

Human-AI Teaming in the ICU: A Comparative Analysis of Data Scientists' and Clinicians' Assessments on AI Augmentation and Automation at Work

Nadine Bienefeld, Emanuela Keller, Gudela Grote

Submitted to: Journal of Medical Internet Research
on: June 20, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 25

 Multimedia Appendixes 26

 Multimedia Appendix 1..... 26

 Multimedia Appendix 2..... 26

Related publication(s) - for reviewers eyes onlies 27

 Related publication(s) - for reviewers eyes only 0..... 27

Human-AI Teaming in the ICU: A Comparative Analysis of Data Scientists' and Clinicians' Assessments on AI Augmentation and Automation at Work

Nadine Bienefeld¹ PhD; Emanuela Keller² MD, Prof Dr; Gudela Grote¹ Prof Dr

¹ETH Zurich Zürich CH

²University Hospital Zurich Zürich CH

Corresponding Author:

Nadine Bienefeld PhD

ETH Zurich

Weinbergstrasse 56/58

Zürich

CH

Abstract

Background: Artificial intelligence (AI) and machine learning hold immense potential for enhancing clinical and administrative healthcare tasks. However, slow adoption and implementation challenges highlight the need to consider how humans can effectively collaborate with AI within the broader socio-technical system.

Objective: We aim to explore the optimal utilization of human and AI capabilities by determining suitable levels of human-AI teaming for safely and meaningfully augmenting or automating tasks. We focus on intensive care units (ICUs) as an example and provide recommendations for policymakers and healthcare practitioners regarding AI deployment in healthcare settings.

Methods: We conducted a systematic task analysis in six ICUs in Europe and carried out an international Delphi survey involving 19 health data scientists from academia and industry (response rate = 95%; 21% female; mean age = 38.6 years; mean experience = 12.63). Consensus was reached on the appropriate level of human-AI teaming for each task (Level 1 = no performance benefits from AI; Level 2 = AI augments human performance; Level 3 = Human augments AI performance; Level 4 = AI performs without human input). Ethical and social implications, as well as control and accountability distribution, were also considered by experts.

Results: Levels 2 and 3 human-AI teaming were preferred choices for four out of six core ICU tasks. However, this recommendation relies on AI systems providing transparency, predictability, and user control. If these conditions are not met, reducing to Level 1 or shifting accountability away from users is advised. Additionally, when AI demonstrates near-perfect reliability, Level 4 automation can enhance safety and efficiency, especially when human-AI teaming conditions are not met. Importantly, AI experts agree that certain tasks should not be augmented or automated due to ethical and social concerns related to the physician/nurse-patient relationship and the roles of healthcare professionals in the future.

Conclusions: By considering the socio-technical system and determining appropriate levels of human-AI teaming, our study showcases the potential for improving the safety and effectiveness of AI utilization in ICUs and broader healthcare settings. Regulatory measures should prioritize transparency, predictability, and user control when users bear accountability. Ethical and social implications must be carefully evaluated to ensure effective collaboration between humans and AI, particularly in light of recent advancements in generative AI and large language models.

(JMIR Preprints 20/06/2023:50130)

DOI: <https://doi.org/10.2196/preprints.50130>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/50130>



Original Manuscript

Human-AI Teaming in the ICU: A Comparative Analysis of Data Scientists' and Clinicians' Assessments on AI Augmentation and Automation at Work

Introduction

The rapid development of artificial intelligence and machine learning (AI) offers unprecedented opportunities for supporting physicians and nurses (hereafter clinicians) in a wide range of clinical and administrative tasks [1], [2]. Despite these promises, however, the integration of AI into actual clinical practice remains slow with significant implementation hurdles remaining [3], [4], [5].

One key obstacle is failing to account for the broader socio-technical system (STS) in which humans and AI collaborate. Disregarding the seamless integration of AI with human work practices and the overall systems in which they operate can not only lead to a lack of acceptance [6], [7]; but also introduce unwanted errors, patient safety risks and, in the long run, increase rather than decrease costs.

To fully harness the complementarity between humans and machines [8], advancements in machine capabilities must be accompanied by simultaneously building human competencies for successfully collaborating with machines [9]. Contrary to popular belief, more capable systems do not necessarily reduce human effort or errors the opposite is often true. Paradoxically, complex systems like AI can translate into higher rather than lower demand for skilled personnel and the need for enhanced expertise, ultimately leading to decreased efficiency and potential performance losses [10]. Furthermore, numerous accidents in aviation e.g., [11] serve as tragic demonstrations that increased system complexity and a shift from manual to supervisory control is challenging for humans in terms of maintaining situational awareness, the loss of control, and deskilling [13], [14], [15], [16].

The lack of transparency in today's black-box AI systems [17] raises the additional question of who should be responsible for the system outcomes. If humans are to maintain overall authority over system goals and their attainment, both accountability and control over system functioning must be afforded to humans, requiring transparency, predictability, and means to influence the systems [18]. In most of today's AI applications, these conditions are not met, exposing clinicians to significant legal and professional repercussions and raising questions about reassigning accountability to other entities (e.g., AI development firms or health insurers) [19].

This study integrates different viewpoints on these issues by data scientists developing AI systems for intensive care (ICU) and the clinicians ultimately using these AI systems. We thus tackle the question of how humans can effectively collaborate with AI from a holistic STS perspective [9] and in a context where AI could support overworked clinicians based on AI solutions using high volumes of patient data [20], [21].

Contrary to technology-driven initiatives our human-first approach considers the complementarity between humans and machines to create joint cognitive systems (humans and AI) that achieve better outcomes than humans or AI could achieve on their own [8]. Just because a particular AI application has the potential to augment or automate a given task does not mean that it should do so at all costs. We thus explore the question of social desirability and ethical acceptability of task automation or augmentation both from the perspectives of clinicians currently performing these tasks and data

science experts developing AI solutions.

By defining optimal human-AI teaming in clinical practice and considering the broader socio-technical context in which AI operates, this study contributes to a more effective, safe, socially acceptable, and ethically sound use of AI in healthcare. The resulting decision framework can assist hospital managers, policymakers, and legislators in making well-balanced decisions about AI implementation and use to reduce the workload for clinicians and advance the overall quality of care.

Methods

This multi-method study combines (1) a systematic socio-technical analysis of ICU work tasks with (2) an international Delphi survey among $n = 19$ data science experts to assess optimal levels of augmentation/automation of these tasks and (3) $n = 61$ semi-structured interviews with clinicians exploring their views on AI augmentation/automation. Ethics approval was obtained from the local Ethics Committee (No. EK 2019-N-51 & EK 2019-N-190) and informed consent was obtained from all participants before data collection.

Socio-Technical Work Task Analysis

The socio-technical work task analysis was conducted using COMPASS, a well-established framework for the assessment of work tasks and systems based on STS theory [22], [23], [24]. Detailed observational notes were analyzed using COMPASS to identify six core tasks performed by ICU clinicians (Textbox 1). Observations were carried out as part of a related study titled “Can AI Help with Healthcare Shortages? A Work System Analysis for Enhancing Job Satisfaction and Well-being [25], during which the first author systematically observed all work-related activities performed by ICU clinicians for 30 morning shifts (8.5 hours each). Additional data were gathered from hospital documents such as descriptions of job profiles, professional competencies, and organizational charts. Two ICU department heads checked and validated the accuracy and adequacy of the analysis concerning the correctness of medical knowledge and adequate use of terminology [26].

Textbox 1. Core tasks performed by ICU clinicians (based on observation and the COMPASS work system analysis [23], in alphabetical order).

1. Monitoring patient data (derived from biosensors such as vital signs, parameters from artificial ventilation, or laboratory values).
2. Documenting clinical information.
3. Analyzing medical data (e.g., from reports, articles, test results, or images).
4. Prescribing medication or treatment.
5. Diagnostic decision-making.
6. Interacting with patients.

Delphi Survey

To assess data science experts' agreement on the levels of human-AI teaming and technological feasibility to augment or automate each of the six core tasks, we conducted an international Delphi survey [27], [28]. The Delphi survey is an iterative process method aiming to forecast future (technological) developments and attain consensus among a group of experts regarding questions where there are no clear right or wrong answers and where there is limited or contradictory information. The Delphi method has been shown to outperform other interactive group approaches regarding accuracy and efficacy [29], making it a commonly utilized technique in health information management and healthcare e.g., [27].

Selection of International Data Science Experts

The quality of Delphi surveys heavily depends on the adequate choice of experts [30]. Therefore, we purposefully selected each expert based on their internationally renowned expertise in the fields of bioinformatics, bioengineering, and health data science. Within these fields, we included academic researchers (professors) from global top-tier Universities and data scientists employed by global healthcare technology manufacturers. Suitable participants were identified through academic publications, participation at conferences, topic-based newsletters [31], and a search of companies and job descriptions on the professional social network LinkedIn, aiming to create a heterogeneous sample regarding geographical regions. Based on the recommended number of participants for Delphi surveys between 10 to 35 [32], we invited 20 data science experts (DS 1-20), out of which 19 agreed to participate (response rate = 95%). Data scientists were 21 percent female¹, between ages 29 and 47, and had between 8 to 21 years of experience as professionals in their domain. 37 percent of experts were employed by global healthcare technology manufacturers and 67 percent were employed as faculty of top-tier Universities worldwide. 42% of experts were from the AMER region (North, Central, and South America), 42% from the EMEA region (Europe, Middle-East, and Africa), and 16% from the APAC region (Asia Pacific excluding China).

Survey Development and Data Collection

As an entry point to the Delphi survey and to provide common ground among experts, we used Russel & Norvig's definition of AI as "machines that mimic cognitive functions that humans associate with the human mind, such as learning and problem-solving" [33] and described each of the six core tasks performed by ICU clinicians in short vignettes to illustrate the ICU work system and workflow (see Appendix A). Experts were then asked to anonymously select which level of automation/augmentation they deemed best suited to enable safe and effective human-AI teaming for each task from a technological feasibility perspective. The levels of human-AI teaming were developed based on Johnson and colleagues' taxonomy that goes beyond the traditional levels of automation [34], in that it "more effectively models the technology, the human, and the work [system] together" [35, p. 82] (see Table 1). In addition, data scientists were asked to indicate which of the core tasks performed in the ICU *should/should not* be augmented or automated by AI based on their subject matter expertise including social and ethical considerations beyond merely considering technological feasibility. For each answer, participants were asked to provide an open-text explanation to justify their choices. The first three experts to respond to our initial invitation were chosen to pilot test the Delphi survey, which resulted in only minor changes (wording). Experts submitted their responses anonymously via email in multiple survey rounds. At each data collection point, two participation reminders were sent via email. There were no dropout cases.

Statistical data about the group's collective choices as well as the qualitative answers given by data scientists explaining their choices were anonymously fed back to all participants in each consecutive round [32]. As recommended by Sumsion[36], agreements above 70% were considered as consensus. This goal was achieved in round three and the Delphi survey was terminated then.

Survey Data Analysis

The Qualtrics survey web-based platform [37] was used to develop the survey and collect data. SPSS Version 24 was used for statistical data analysis.

Interviews with ICU Clinicians

To examine ICU clinicians' perspectives on the social and ethical implications of AI-enabled automation and augmentation technologies, we conducted 61 semi-structured interviews as part of a

¹ The relatively low proportion of female experts represents the gender distribution in the field of data science.

larger research project [38]. Interview questions related to clinicians’ vision of future human-AI teaming solutions and the distribution of control and accountability (see Appendix B). Interviews were conducted in private offices, lasted between 60 to 90 minutes, and were audio-recorded and manually transcribed ad verbatim.

Selection of Clinicians

In line with grounded theory [39], we employed a theoretical sampling approach to include both professional groups of ICU physicians and nurses (25 physicians and 36 nurses). Clinicians were 56 percent female and had between 2 to 30 years of experience after completing their initial education as a registered nurse or board-certified physician. All informants were directly involved in care delivery. We gave each informant a code showing their professional position (attending physician [AP], resident physician [RP], registered nurse [RN]) and personal identifier (1–61).

Interview Data Analysis











The analysis of the interview content was conducted following the grounded theory methodology [39]. The principal investigator engaged in the process of open coding, methodically labeling each discrete conceptual unit within the interview transcripts. This procedure entailed the aggregation of related codes into broader categories and themes. Subsequently, a review of the related textual segments was performed to verify the consistent alignment of the data with the identified themes, a step that is pivotal for maintaining the integrity of qualitative inquiry [40]. The software MAXQDA Version 2024 was employed to facilitate qualitative data analysis [41]. The presentation of the qualitative findings adheres to the protocols endorsed by the Academy of Medicine [42].

Results

Data Science Expert Perspective

Table 1 summarizes the results from the Delphi survey. Data science experts’ assessments about the technological potential to augment or automate each ICU task were quite similar from the start, and increasing consensus between 72.7%-100% was reached across the three Delphi survey rounds. Data science experts also showed high levels of agreement (92%-100%) about which of the six tasks *should/should not* be augmented or automated by AI based on social and ethical considerations, except for one task: For the task “diagnostic decision-making” 53% of data science experts recommended the use of AI whereas 47% did not.

Table 1. Results from the Delphi Survey

Levels of Human-AI Teaming (adapted from Johnson et al. [43])																	
Task	Level 1 Human performs well without AI. <i>Assisted automation</i> (AI is unable to significantly augment human performance)			Level 2 AI augments human performance. <i>Partial automation</i> (Human benefits from AI augmentation but is always needed)			Level 3 Human augments AI performance. <i>Conditional automation</i> (AI could perform alone but human input increases reliability)			Level 4 AI automates task without human input. <i>High automation</i> (AI performs reliably. Humans serve as trouble-shooters)			Should AI be used?				
																	
	1	2	3	1	2	3	1	2	3	1	2	3	Yes	No			
Monitoring													63.6	63.6	72.7	92	8

patient data

Documenting clinical information				90.1	90.1	90.1		100	-
Analyzing medical data				81.8	90.1	100		92	8
Prescribing medication or treatment		81.8	90.1	90.1				92	8
Diagnostic Decision-Making		81.8	90.1	90.1				53	47
Interacting w. patients	90.1	90.1	90.1					-	100

In what follows we provide a detailed account of how the consensus-building process among data science experts unfolded as part of the Delphi survey process. For each task and level of human-AI teaming, we provide illustrative statements by data science experts justifying their choices.

Monitoring patient data. In the first round, 63.6% of data science experts agreed that AI could eventually fully automate this task and that human assistance would bring no further benefits (Level 4). This result remained the same also in the second round but increased to a 72.7% consensus in the third round. One data science expert justified this choice by stating:

Large longitudinal data like heart rate, BP, etc., are much harder to interpret for humans. It's like weather forecasting where you have so many data points and supercomputers can more accurately model weather than humans. Even the most experienced clinicians are overwhelmed with today's data overload and since human monitoring is not feasible at all times, full automation without human interference is really the best option. (DS-17)

27.3% of experts selected the design choice “AI could do it, but humans increase reliability” in the first and second rounds:

Monitoring of the signs is possible to learn by ML; however, having human help when an obvious mistake happens increases efficiency and reliability and seems unavoidable. AI could not do it as reliably and reproducibly as doing it together with the human. (DS-05)

Regarding the question as to whether “monitoring patient data” *should* be automated by AI (Level 4), all but one expert (92%) agreed that there were no social or ethical concerns:

Automated monitoring is bound to be better than what there is now. Even if the system fails at times, it's not that patients will die from this. (DS-11)

Documenting clinical information. In the first round, 90.1% of experts agreed that AI could eventually automate this task, but human assistance would increase reliability (Level 3). Expert agreements remained stable throughout all three rounds. The following argument was given by a data science expert further explicating their choice:

There are a lot of approaches that involve humans in the loop such as active learning/weak supervision that can greatly augment the reliability of AI techniques used for these types of tasks. With greater data standardization and curated datasets, the goal of greater automation becomes more and more feasible but for now, physicians will have to be in the loop. (DS-07)

Also, 100% of experts agreed, that this task *should* be augmented by AI, for instance, as one data science expert argued:

That's [the task of documenting clinical information] what [clinicians] don't like doing and where AI can really help so they can focus on more interesting things.
(DS-15)

Analyzing medical data. Already in round one, most data science experts (81.8%) agreed that AI could eventually do this task autonomously, but that human assistance would increase reliability (Level 3), for instance, stating that:

So many fields are already doing automated reporting and with humans in the loop, this becomes easier and easier for AI systems to automate and for medical professionals to interpret. Again, not fully autonomous but moving closer on the continuum towards it. (DS-09)

Two Data science experts adopted the consensus opinion in the second and third rounds of the Delphi process resulting in a 100% agreement. One of them provided the following explanation for changing from Level 4 to Level 3:

I agree. The process of automatically producing "draft analysis reports" can be fully automated but if we take human sense-making as a kind of "assistance" then I would also say that this [Level 3] is the right choice. (DS-18)

Furthermore, 92% of experts agreed that there were no social or ethical concerns and that this task *should* be augmented by AI (Level 3).

Prescribing medication or treatment. In the first round, 81.8% of data science experts agreed that this task could be augmented by AI but would always require humans in the lead (Level 2). This opinion increased to 90.1% in rounds two and three. The following arguments were given by data science experts explaining their choices:

Although automated diagnostics and algorithmic medication prescriptions have become extremely accurate because mistakes are very costly, a medical professional should always review and approve the suggestions produced by the software. (DS-10)

One expert (9.1%) argued that AI could perform this task with human assistance (Level 3) and did not change his or her opinion throughout the Delphi survey process based on the following explanation:

I don't think human oversight is always required. It will be like how self-driving cars are gradually phasing into our society as well. There are certain use cases where automated diagnostics can yield greater results in a well-designed system that triages events earlier in the healthcare system and prevents issues later downstream that have bigger healthcare impacts in terms of patient health and costs. (DS-19)

92% of data science experts agreed that this task *should* be augmented by AI as long humans are accountable and in the lead (Level 2):

AI tools will never be ready to prescribe meds without a human in the loop simply

because they don't take legal liability for the decisions which ultimately will have to be made by the physician. (DS-06)

Diagnostic Decision-Making. In the first round, 81.8% of data science experts indicated that AI could augment this task but humans would always be required (Level 2) and in the second and third rounds 90.1% selected this choice. As one Data science expert stated:

I think our [AI] solutions are far too narrow for this type of decision-making. They [AI solutions] can assist but need human input always. (DS-05)

Two experts (18.2%) chose Level 3 for this task in the first round stating that:

Diagnostics and prognosis can be automated really well with AI/ML but depends on how it is accepted. But in my opinion, having 100% human supervision defeats the purpose of AI/ML. (DS-16)

Concerning social and/or ethical considerations 47% of data science experts believed, that diagnostic decision-making *should not* be augmented. by AI, even if it was technologically feasible to do so:

I strongly believe that for as long as we [data scientists] are still struggling with issues of bias, transparency, and equity etc., we should shy away from such consequential tasks [diagnostic decision-making]. ML can be utilized in the form of a "recommendation engine", but doctors should never rely on it fully. (DS-19)

Besides the ethical issues of bias, transparency, and equity, some data science experts also felt that using AI in diagnostic decision-making would undermine physicians' role identity which could be problematic.

This is what being a doctor is all about. Applying one's knowledge and experience to diagnose patients is the very core of medicine. It would be like messing with the Hippocratic oath and taking away the reason why they chose to become a doctor in the first place. (DS-08)

Interacting with patients. With a strong 90.1% agreement in all three rounds and no changes throughout, experts believed that the task of patient interaction cannot be augmented or automated by AI.

AI/ML technologies simply cannot properly mimic human soft skills which are essential in these sorts of interactions. (DS-13)

Only one expert argued that AI could augment humans in their patient interactions (Level 2) by stating:

In most cases, it is challenging for AI/ML to interact empathically with patients, but there are some new use cases with promising results that could potentially augment this task. (DS-07)

However, regardless of the technological possibilities of AI mimicking empathic patient interactions, all data science experts (100%) agreed, that AI *should not* be used to interact with patients, mainly based on social concerns and what it means to be human:

Medicine is all about empathy, one human caring for another [...]. Therefore, [interactions with patients] should never be replaced by an AI [ro]bot or the like. (DS-11)

The Role of AI is to provide more time for physicians to do what this empathy-driven field requires most: human interaction and connection. No AI should interfere with that. (DS-17)

Clinician perspective

In the following, we report clinicians' views on how to best team up with AI when augmenting or automating ICU tasks and conclude with a brief comparison between the two stakeholder groups.

Monitoring patient data. Like data scientists, clinicians acknowledged the benefits of automation of the monitoring task. Many clinicians, however, pointed out the need for highly reliable systems so they would no longer have to constantly supervise AI, which defeats the purpose of gaining time and increasing efficiency:

Taking over the monitoring [task] would really save us time and reduce our workload. But of course only if the AI is so good that we don't have to constantly monitor it. Because until now, with AI [systems], we still have to do the monitoring so it is more like a double effort and I actually have less time for the patient (RN-43)

Documenting clinical information. Clinicians perceived the documentation task as "the most time-consuming [task] of all" imagining a future where AI could "take away that burden" (RP-12) and welcoming high levels of automation for this task. As one attending physician explained, their future roles would ideally consist of merely checking for potential errors in AI-produced documents:

This [automation of clinical documentation] would be a huge relief, especially for residents because they are the ones who spend the most time on administrative tasks. I see them sit here [in the ICU] for hours on end way past they clock out [after the end of their shift] and I think to myself, "All those years of medical school to do what - office work!" That can't be right. (AP-39)

Analyzing medical data. Clinicians saw immense potential in using AI to aid with the analysis of medical data, but only if they could "stay in the driver's seat". Most clinicians said they would always want "to decide whether to follow the advice [given by AI] or not" (AP-11). One nurse gave the example of using AI to analyze medical data so they could foresee the risk of delirium in ICU patients so they could initiate prophylactic actions before problems occurred:

All the vast amounts of data that we can use to predict the onset of conditions such as delirium. That would be brilliant because delirious patients have a considerably higher mortality rate and are very resource-intensive. In some ICUs, [the use of AI to analyze medical data] is already available, for instance, to predict sepsis and so on. A lot is going on at the moment and I see a huge opportunity [for AI] to recognize all these risks in patients so that we can simply say, "Oh yes, that makes sense" and take prophylactic action. (RN-36)

Prescribing medication or treatment. Similar to data scientists, clinicians realized AI's potential to augment the task of prescribing medication and treatment, but they stressed the importance of keeping the ultimate decision power in their own hands. As one resident physician explained this important precondition of using AI for this task:

Yes, of course, an AI system that can provide us with information on adverse drug interactions or patient-specific intolerance, for example, would be great because we humans are much more prone to error for these kinds of tasks than machines.

This also goes a bit in the direction of personalized medicine and that is the future. But one thing that must not happen is that the system then directly initiates a therapy, that would be dangerous. (RP-58)

One use case that was discussed intensively by clinicians was the use of AI to assist decision-making around which treatment patients should receive based on ethical considerations and long-term prognosis:

Where AI would help us a lot would be in deciding whether and if so which therapy really makes sense for a patient in terms of prognosis and quality of life long-term. Now we have interdisciplinary ethics boards for this [decision], but it is often unclear what the outcome will be for each patient. If AI could show all the facts and the prognosis and perhaps also visualize them, our decisions would ultimately be more ethical, objective, comprehensible, and better in terms of health economics for sure. (RN-30)

Diagnostic Decision-Making. Unlike data scientists, not a single clinician worried about themselves. Rather than focusing on the changes AI would incur in their professional lives, clinicians always focused on the benefits AI could bring to their patients. As one attending physician explained, putting patient benefits front and center is what matters most in the decision on when and how to use AI:

What's important here is not whether we like [the changes] or not but how the patient can benefit. That AI helps as many patients as possible. I mean it's well known that it is more difficult for us [human physicians] to diagnose rare diseases for instance and that we have all kinds of cognitive biases. That's where I see the big potential of AI, that it can help [improve diagnostic accuracy]. (AP-01)

Unlike data science experts, issues relating to biased or unfair AI algorithms did not feature in clinicians' answers. Instead, clinicians highlighted how important it was to "see every patient as a unique human being" and to assess each patient "holistically including individual, social, and environmental factors" (RN-45):

Diagnostic support from AI is great. But where I would have problems is if everything was treated in the same way and we could no longer have a say. Because every patient is unique, one patient feels fine with a blood pressure of 60 and another one is already lying flat. Medicine cannot be generalized, it cannot be standardized according to 08/15 standards and norms, that would be a setback. (RN-43)

Interacting with patients. Clinicians agreed with data scientists on the importance of social interaction, empathy, and human connection in patient interactions. They too were unable to imagine a future where AI would interact socially and demonstrate human-level empathy with patients:

It has been proven that people also need human connection and closeness to get well. Our social skills, caring, and empathy are enormously important for a patient to overcome even very difficult crisis situations. I can't imagine that an AI system, that is a robot, will ever be able to do this. Especially not here in the ICU where patients are on the border between life and death. (RN-36)

Furthermore, interactions with patients were seen as one of the core tasks why clinicians had chosen their profession and one that they hoped they could invest more time in thanks to AI augmentation and automation:

The human aspect in nursing is actually the reason why I chose this profession. I think it would be incredibly hard for me if it was just a case of operating the AI, maybe turning the patient once a shift and that would be it. So that's what I'm hoping for with AI, that we'll find time again to care for our patients more closely, to talk to them or their relatives in peace, to do all the actual caring again. Unfortunately, there's not enough time for that these days. (RN-17)

To summarize, clinicians' and data scientists' perspectives regarding the optimal levels of human-AI teaming, and ethical/social concerns were very similar across all tasks. However, clinicians provided a more nuanced picture regarding the effects different levels of augmentation/automation would have on their ability to validate AI's performance. Specifically, they argued for higher rather than lower levels of automation for the tasks of monitoring and documentation, provided that AI performance was highly reliable. Clinicians agreed with data scientists, that they would need to retain the ultimate control and decision-making power for the tasks of analyzing medical data, prescribing medication/treatment, and clinical decision-making but not because they feared losing professional status or purpose but because were worried about patient safety. Clinicians worried less about bias or inequity in AI but more about standardization and failing to see each patient as a unique human being., i.e., that it made sense given a specific patient situation in the clinical context. Finally, for the task of interacting with patients, both stakeholder groups were clear about not wanting AI to interfere because they both valued "real" human empathy, care, and connection.

Discussion

In this study, we identified six core tasks characterizing the work in hospital ICUs and assessed optimal levels of human-AI teaming for each task, based on the potential for automation/augmentation from data science experts' and ICU clinicians' perspectives. Such a human-centered assessment based on STS theory contributes to the ongoing debates about the future use, and social and ethical implications of AI in healthcare in the following important ways.

First, contrary to the predominant focus on what AI can and will be able to achieve, our approach considers the strengths and weaknesses of both humans and AI in a complementary fashion. Such an approach reduces the risks of misalignment and aims at achieving better overall system outcomes than humans or AI could achieve alone [8], [22], [35]. The ICU context served as an ideal context due to the availability of large amounts of patient data and need for technological support but our methodology itself can be used in any work system be it in healthcare or beyond.

Second, whereas the current consensus on the future of AI in healthcare suggests that augmentation is always better than automation, the results from our Delphi study and the responses from clinicians themselves paint a more nuanced picture. Augmentation (Levels 2 and 3) was considered the ideal form of human-AI teaming for four out of six tasks but only if certain requirements are met. From data scientists' perspectives, these requirements include full transparency, predictability, and sufficient means to influence the system [18], [22]. Currently, the goals of higher transparency and predictability are not always attained due to black-boxed AI systems and the unresolved explainable AI conundrum. Clinicians said they would not need to understand the AI algorithms as long as it provided interpretable results (e.g., people can drive a car without knowing how the engine works). Involving clinicians in the co-design of interpretable rather than fully transparent systems could thus be a solution to solving the explainable AI conundrum as a recent study has shown [44]. Given the high safety risks for patients, these considerations are particularly important for the tasks of "diagnostic decision-making", "prescribing medication or treatment", and "analyzing medical data". Also, both stakeholder groups agreed that at Levels 2 and 3, both the control over and responsibility for system outcomes must reside with the user, i.e., with clinicians.

If full human control is not possible (due to a lack of interpretability or because the necessary skills or knowledge are absent or lost over time), legal responsibility would need to shift to another entity (e.g., AI providers). In cases where AI applications advance to a level of robustness and near-perfect reliability yet human control is not guaranteed, higher levels of automation can increase the overall system safety and efficiency [15], [18]. For the task of "monitoring patient data", data science experts and clinicians chose level 4 automation because it is impossible for humans to continuously monitor an AI system.

Looking into the future and considering the fast-moving developments in generative AI (GenAI) applied to healthcare [45], tasks such as "documenting clinical information" or "analyzing medical data" could benefit from higher levels of automation but also introduce new ethical and security risks[46]. Because these models are capable of combining multiple modalities (text, image, code, or video), AI applications may soon no longer be restricted to performing single tasks. Contemplating the state-of-the-art performance of such future AI applications on a wide variety of problems e.g., [47], [48], they have the potential to automate tasks such as "documenting clinical information" by integrating and conjointly analyzing laboratory, visual, and patient history data. Also, while such systems currently do make mistakes, they are also able to self-correct when prompted to do so [2].

Third, for the task of interacting with patients both data science experts and the clinicians in this study believed AI should not assist humans even if it was technologically feasible. They unanimously agreed that compassion and empathy lie at the core of the physician-patient or nurse-patient relationship and that these innately human qualities should remain in the sole hands of humans. Intriguingly, the chatbot GPT-4 [49] was recently rated as more "empathetic" than human physicians by a panel of expert physicians who assessed physician-patient interactions on social media [50]. Findings like these raise ethical questions about how AI's increasingly powerful capabilities should be used in the future. For instance, GenAI could be used to assist clinicians in brainstorming ideas on how to best structure difficult patient conversations or to simulate which questions patients are likely to have. This is not to say that clinicians lack these skills altogether but that they often lack the time to prepare for challenging patient interactions [2]. After all, if AI can help free up time for tasks spent away from patients, clinicians' number one desire—(re-)gaining time spent with their patients—could finally be realized.

Finally, the question of whether "diagnostic decision-making" should or should not be augmented or automated by AI revealed the largest disagreement both within our sample of data science experts (53% yes vs. 47% no) and between the sample of data science experts and clinicians. Opposing the 47% of data science experts who believed AI should not be used for this task, all clinicians in our study believed that AI should be used to support them in diagnostic decision-making, as long as they could retain the final decision-making power. Unlike data science experts, clinicians were not concerned about biased or unfair AI algorithms. Instead, they stressed the importance of making holistic patient assessments and that every patient is unique with specific personal or environmental circumstances that must be considered. Moreover, some data science experts were concerned that AI augmentation or automation of the diagnostic decision-making task would threaten clinicians' role identity and professional sense of meaning and that AI should therefore not be used. Interestingly, clinicians put these concerns behind the benefits that AI would bring to their patients. One important question remains in terms of the joint optimization of this task from a socio-technical perspective: If AI applications are capable of increasing diagnostic accuracy beyond human levels, how can clinicians retain the ability to build medical expertise and double-check AI results? Also, with highly capable systems and increasing accuracy of results, the risk of overreliance on AI is likely to increase with potential long-term consequences such as loss of expertise and deskilling [51].

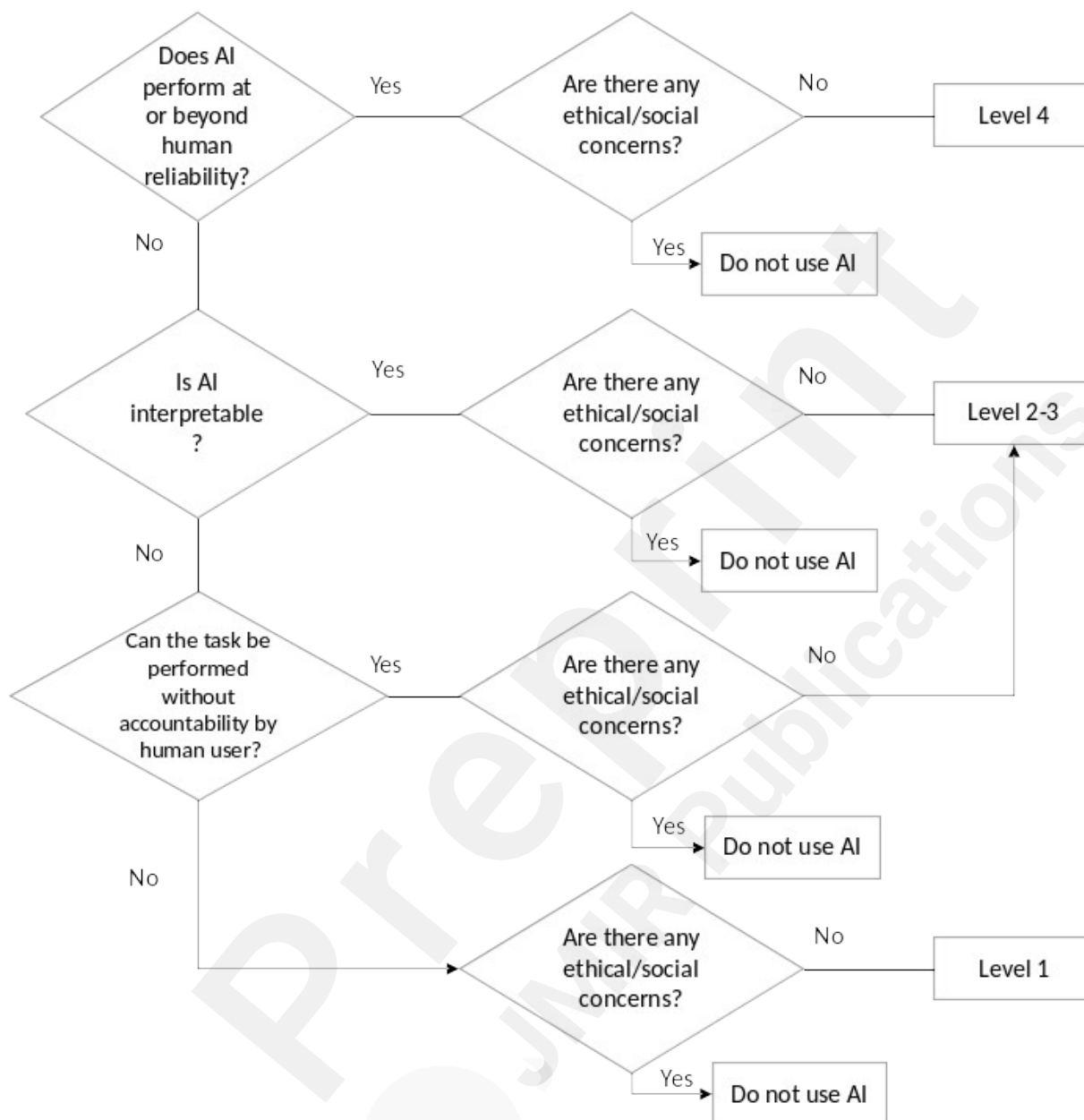
Implications for Policy and Practice

Based on these results, Figure 1 shows a decision diagram aimed to assist hospital managers, IT managers, or policymakers in assessing optimal levels of human-AI teaming to enable AI's safe and effective use. The decision process is based on the assumption that reliability, interpretability, and accountability are key factors in defining optimal levels of human-AI teaming by augmenting/automating a task. Furthermore, regardless of technological capabilities, the question of whether there are any ethical or social concerns to be considered should always be asked. As depicted in Figure 1, if AI reliability is near perfect and there are no ethical/social concerns, Level 4 automation is advised. If AI does not (yet) perform highly reliably but system interpretability is granted—or if accountability is shifted away from human users—Levels 2 or 3 are recommended. Finally, if AI systems are neither reliable nor interpretable and accountability remains with human users, Level 1 is advised or AI should not be used. Textbox 2 summarizes this study's key takeaways and implications for policymaking and practice.

Textbox 2. Implications for Policy and Practice.

1. Adopting a STS perspective can prevent costly misalignment between AI, humans, and their work and facilitate implementation into clinical practice.
2. Optimal human-AI teaming is based on considering the strengths and weaknesses of humans and AI in a complementary fashion thus creating better outcomes than either one could achieve on their own.
3. Future users (i.e., clinicians) should be involved in co-designing future AI applications (e.g., [44]) to leverage the complementarity of humans and AI and to continuously monitor human-AI teaming effectiveness also post-implementation.
4. Human users need specific knowledge, skills, and attitudes to effectively work in human-AI teams; human-centered interaction design is not enough.
5. At the level of augmentation (Levels 2-3), users must be enabled to understand which data are used to make predictions, interpret results within the clinical context, and have adequate means of controlling the system.
6. If AI demonstrates near perfect (human-level or above) reliability yet the conditions 1-5 are not met, Level 4 automation can create higher safety and efficiency.
7. All relevant stakeholders (e.g., clinicians, patients, regulators, management, patient safety experts) must decide whether AI should/should not be used based on social/ethical considerations beyond technological feasibility.
8. Regulatory foresight must address the question of control and accountability and reconsider liability assignments and/or insurance arrangements.
9. AI solutions are continuously evolving with the need for reevaluation of the STS.
10. Policies around AI-enabled in-context learning are needed, especially regarding new generations of generative AI and LLMs [1].

Figure 1. Decision flow-chart summarizing recommended levels for human-AI teaming based on study results and STS literature.



Limitations and Future Research

Several limitations must be considered when interpreting the results of this study. First, as we used the example of ICUs to forecast the future use of AI to augment or automate various tasks, the peculiarities of other medical domains in terms of tasks, level of risk, and ethical considerations should be specified in future research.

Second, although the Delphi technique has proven to be the most adequate method for developing probable future scenarios based on experts' present knowledge and outperforms comparable interactive group forecasting techniques [29], the realization of the forecast depends on multiple other factors such as the availability of resources, legal frameworks, or the acceptance by clinicians. Although the clinicians in our study agreed with data science experts about the optimal levels of human-AI teaming for most tasks, there were some differences such as the importance of considering a patient situation holistically (e.g., social context or environmental factors), beyond measurable data. Also, in the rapidly evolving landscape of AI, expert assessments may change and will need to be adapted over time. Future research, and ideally also policymaking, should adopt our proactive approach to regularly ask a panel of data science experts, clinicians, and potentially additional stakeholders such as health insurers, or hospital management, about their opinions to always be one step ahead of upcoming technological developments. This presents an alternative route to the currently predominant technology-driven approaches to AI implementation in healthcare.

Third, cultural aspects are known to influence people's perceptions regarding the social and ethical considerations of AI use [52]. For this reason, we paid special attention to including a broad sample of data science experts from AMER, EMEA, and APAC regions. In this study, we found no notable differences between these geographical areas but we have to acknowledge that we were unable to recruit data science experts from China and Africa. Future research should involve China and African countries especially. Also, the clinicians in our study, although representing multiple nationalities, are not a fully representative sample of the global ICU clinician population. Even though most ICUs adhere to similar standards and work tasks, there may be other factors such as the availability of resources to consider.

Lastly, despite the widespread use of the Delphi technique, particularly in medical informatics research [27], [28], several concerns have been raised. Anonymity, a central characteristic of the Delphi technique, is intended to promote unbiased assessments from experts by eliminating social desirability bias. Nevertheless, it has been argued that anonymity may lead to hasty judgments, as experts feel relieved from the responsibility of defending their responses [53]. Additionally, the Delphi technique necessitates the disclosure of interim results in each round to facilitate the generation of a group consensus, which some scholars contend compromises independent judgment because this disclosure may exert social pressure on outliers to revise their assessments, thereby promoting group conformity instead of genuine changes in opinion [54]. Hence, as noted by McKenna [55], group consensus should not be regarded as the only "correct answer," but seen as a way to structure the discussion among experts in domains where no right or wrong answers exist. In this study, we believe we achieved this goal by additionally asking experts to provide open-text answers explaining their motivation to change previous ratings [56].

Conclusions

With the overall aim of enabling safe, socially accepted, and ethically sound use of AI to reduce workload and improve quality of care, this study offers valuable insights into the potential of human-AI teaming in ICUs and wider healthcare settings. By adopting an STS perspective, we emphasize

the importance of considering human and AI strengths and weaknesses in a complementary fashion to optimize outcomes and minimize misalignment. The findings challenge the prevailing notion that augmentation is always preferable to automation by demonstrating a more nuanced picture of the ideal interaction levels for each task.

We propose several key principles for implementing AI in healthcare for policy and practice, e.g., incorporating user-centered co-design, promoting transparency, and predictability, and ensuring control over AI systems in cases where accountability resides with users. Moreover, considering both ethical and social implications when deciding on AI applications is important, especially considering the promising new developments in GenAI. Regulatory foresight and knowledge about optimal fit between work tasks and levels of human-AI teaming will be crucial in shaping the future of AI in healthcare and beyond.

Acknowledgments

We thank the esteemed data science experts and clinicians who devoted their time to participate in this study. N.B. and G.G. contributed to the initial study design and preliminary analysis plan. E.K. provided her medical expertise to set up the study, verify findings from the work system analysis, and formulate questions. N.B. conducted all parts of the study and implemented the study protocol. N.B. and G.G. contributed to the review and analysis of the results. N.B. and G.G. contributed to the preparation of the manuscript. All authors approved the final version of this manuscript.

Conflicts of Interest

None declared.

References

- [1] M. Moor *et al.*, “Foundation models for generalist medical artificial intelligence,” *Nature*, vol. 616, no. 7956, Art. no. 7956, Apr. 2023, doi: 10.1038/s41586-023-05881-4.
- [2] P. Lee, S. Bubeck, and J. Petro, “Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine,” *N. Engl. J. Med.*, vol. 388, no. 13, pp. 1233–1239, Mar. 2023, doi: 10.1056/NEJMs2214184.
- [3] F. Gama, D. Tyskbo, J. Nygren, J. Barlow, J. Reed, and P. Svedberg, “Implementation Frameworks for Artificial Intelligence Translation Into Health Care Practice: Scoping Review,” *J. Med. Internet Res.*, vol. 24, no. 1, p. e32215, Jan. 2022, doi: 10.2196/32215.
- [4] M. Sharma, C. Savage, M. Nair, I. Larsson, P. Svedberg, and J. M. Nygren, “Artificial Intelligence Applications in Health Care Practice: Scoping Review,” *J. Med. Internet Res.*, vol. 24, no. 10, p. e40238, Oct. 2022, doi: 10.2196/40238.
- [5] J. Shaw, F. Rudzicz, T. Jamieson, and A. Goldfarb, “Artificial Intelligence and the Implementation Challenge,” *J. Med. Internet Res.*, vol. 21, no. 7, p. e13659, Jul. 2019, doi: 10.2196/13659.
- [6] R. Abdullah and B. Fakieh, “Health Care Employees’ Perceptions of the Use of Artificial Intelligence Applications: Survey Study,” *J. Med. Internet Res.*, vol. 22, no. 5, p. e17620, May 2020, doi: 10.2196/17620.
- [7] I. M. Olaye and A. A. Seixas, “The Gap Between AI and Bedside: Participatory Workshop on the Barriers to the Integration, Translation, and Adoption of Digital Health Care and AI Startup Technology Into Clinical Practice,” *J. Med. Internet Res.*, vol. 25, no. 1, p. e32962, May 2023, doi: 10.2196/32962.
- [8] E. Hollnagel and D. D. Woods, *Joint Cognitive Systems: Foundation of Cognitive Systems Engineering*. Taylor and Francis, New York, 2005.
- [9] M. Johnson and A. Vera, “No AI Is an Island: The Case for Teaming Intelligence,” *AI Mag.*, vol.

- 40, no. 1, pp. 16–28, Mar. 2019, doi: 10.1609/aimag.v40i1.2842.
- [10] J. L. Blackhurst, J. S. Gresham, and M. O. Stone, “The autonomy paradox,” Oct. 2011. Accessed: Sep. 02, 2019. [Online]. Available: <http://armedforcesjournal.com/the-autonomy-paradox/>
- [11] BEA, “Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro – Paris.” French Civil Aviation Safety Investigation Authority, 2012.
- [12] NTSB, “NTSB Preliminary Report: Highway HWY18FH013 Fatal Tesla Crash.” Accessed: Jan. 17, 2019. [Online]. Available: <https://www.nts.gov/investigations/AccidentReports/Pages/HWY18FH013-prelim.aspx>
- [13] C. E. Billings, *Aviation automation: the search for a human-centered approach*. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers, 1997. Accessed: Sep. 05, 2019. [Online]. Available: <https://trove.nla.gov.au/version/16507648>
- [14] R. I. Cook and D. D. Woods, “Operating at the Sharp End: The Complexity of Human Error,” in *Human Error in Medicine*, 1st ed., M. S. Bogner, Ed., CRC Press, 2018, pp. 255–310. doi: 10.1201/9780203751725-13.
- [15] M. R. Endsley, “From Here to Autonomy: Lessons Learned From Human–Automation Research,” *Hum. Factors*, vol. 59, no. 1, pp. 5–27, 2017.
- [16] J. D. Lee, “Review of a Pivotal Human Factors Article: ‘Humans and Automation: Use, Misuse, Disuse, Abuse,’” *Hum. Factors*, vol. 50, no. 3, pp. 404–410, Jun. 2008, doi: 10.1518/001872008X288547.
- [17] C. McLellan, “Inside the black box: Understanding AI decision-making,” ZDNet. Accessed: Apr. 27, 2018. [Online]. Available: <https://www.zdnet.com/article/inside-the-black-box-understanding-ai-decision-making/>
- [18] G. Boy and G. Grote, “The Authority Issue in Organizational Automation,” in *Handbook of Human-Machine Interaction: A Human-Centered Design Approach*, U.K.: Ashgate, 2011, pp. 131–150.
- [19] N. Helberger and N. Diakopoulos, “ChatGPT and the AI Act,” *Internet Policy Rev.*, vol. 12, no. 1, Feb. 2023, Accessed: Jun. 05, 2023. [Online]. Available: <https://policyreview.info/essay/chatgpt-and-ai-act>
- [20] D. van de Sande, M. E. van Genderen, J. Huiskens, D. Gommers, and J. van Bommel, “Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit,” *Intensive Care Med.*, vol. 47, no. 7, pp. 750–760, Jul. 2021, doi: 10.1007/s00134-021-06446-7.
- [21] D. van de Sande *et al.*, “Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter,” *BMJ Health Care Inform.*, vol. 29, no. 1, p. e100495, Feb. 2022, doi: 10.1136/bmjhci-2021-100495.
- [22] G. Grote, C. Ryser, T. Waefler, A. Windischer, and S. Weik, “KOMPASS: a method for complementary function allocation in automated work systems,” *Int. J. Hum.-Comput. Stud.*, vol. 52, no. 2, pp. 267–287, Feb. 2000, doi: 10.1006/ijhc.1999.0289.
- [23] D. Boos, G. Grote, and H. Guenter, “A toolbox for managing organisational issues in the early stage of the development of a ubiquitous computing application,” *Pers. Ubiquitous Comput.*, vol. 17, no. 6, pp. 1261–1279, Aug. 2013, doi: 10.1007/s00779-012-0634-y.
- [24] T. Waefler, G. Grote, A. Windischer, and C. Ryser, “KOMPASS: A method for complementary system design,” in *Handbook of cognitive task design*, in Human factors and ergonomics. , Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2003, pp. 477–502. doi: 10.1201/9781410607775.ch20.
- [25] N. Bienefeld, E. Keller, and G. Grote, “Can AI Help with Healthcare Shortages? A Work System Analysis for Enhancing Job Satisfaction and Wellbeing,” JMIR Preprints. Accessed:

- Nov. 12, 2023. [Online]. Available: <https://preprints.jmir.org/preprint/50852>
- [26] C. Seale, "Quality in Qualitative Research," *Qual. Inq.*, vol. 5, no. 4, pp. 465–478, Dec. 1999, doi: 10.1177/107780049900500402.
- [27] A. Ermolina and V. Tiberius, "Voice-Controlled Intelligent Personal Assistants in Health Care: International Delphi Study," *J. Med. Internet Res.*, vol. 23, no. 4, p. e25312, Apr. 2021, doi: 10.2196/25312.
- [28] F. Hasson, S. Keeney, and H. McKenna, "Research guidelines for the Delphi survey technique," *J. Adv. Nurs.*, vol. 32, no. 4, pp. 1008–1015, Oct. 2000, doi: 10.1046/j.1365-2648.2000.t01-1-01567.x.
- [29] A. Graefe and J. S. Armstrong, "Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task," *Int. J. Forecast.*, vol. 27, no. 1, pp. 183–195, Jan. 2011, doi: 10.1016/j.ijforecast.2010.05.004.
- [30] G. Rowe, G. Wright, and A. McColl, "Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence," *Technol. Forecast. Soc. Change*, vol. 72, no. 4, pp. 377–399, May 2005, doi: 10.1016/j.techfore.2004.03.004.
- [31] E. Chen, S. Johri, P. Rajpurkar, and E. Topol, "Doctor Penguin," Jul. 2020. [Online]. Available: <https://doctorpenguin.com/newsletters>
- [32] G. Wright, M. J. Lawrence, and F. Collopy, "The role and validity of judgment in forecasting," *Int. J. Forecast.*, vol. 12, no. 1, pp. 1–8, Mar. 1996, doi: 10.1016/0169-2070(96)00674-7.
- [33] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Third. Malaysia: Pearson Education Limited, 2010.
- [34] L. Onnasch, C. D. Wickens, H. Li, and D. Manzey, "Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 56, no. 3, pp. 476–488, May 2014, doi: 10.1177/0018720813501549.
- [35] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. van Riemsdijk, and M. Sierhuis, "Coactive design: designing support for interdependence in joint activity," *J. Hum.-Robot Interact.*, vol. 3, no. 1, pp. 43–69, Feb. 2014, doi: 10.5898/JHRI.3.1.Johnson.
- [36] T. Sumsion, "The Delphi technique: an adaptive research tool.," *Br. J. Occup. Ther.*, vol. 61, no. 4, pp. 153–156, 1998.
- [37] Qualtrics, Provo, Utah, USA, 2020. [Online]. Available: <https://www.qualtrics.com>
- [38] N. Bienefeld, "Human-AI Teaming in High-Risk Work: The Role of Cognitive-Affective Frames, Sociomaterial Practices, and Socio-Technical System Integration," ETH Zurich, Zurich, 2024.
- [39] B. G. Glaser and A. L. Strauss, *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge, 2017.
- [40] D. A. Gioia, K. G. Corley, and A. L. Hamilton, "Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology," *Organ. Res. Methods*, vol. 16, no. 1, pp. 15–31, Jan. 2013, doi: 10.1177/1094428112452151.
- [41] U. Kuckartz and S. Rädiker, *Analyzing Qualitative Data with MAXQDA: Text, Audio, and Video*. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-15671-8.
- [42] B. C. O'Brien, I. B. Harris, T. J. Beckman, D. A. Reed, and D. A. Cook, "Standards for Reporting Qualitative Research: A Synthesis of Recommendations," *Acad. Med.*, vol. 89, no. 9, pp. 1245–1251, Sep. 2014, doi: 10.1097/ACM.0000000000000388.
- [43] M. Johnson, J. M. Bradshaw, and P. J. Feltovich, "Tomorrow's Human–Machine Design Tools: From Levels of Automation to Interdependencies," *J. Cogn. Eng. Decis. Mak.*, vol. 12, no. 1, pp. 77–82, Mar. 2018, doi: 10.1177/1555343417736462.
- [44] N. Bienefeld *et al.*, "Solving the Explainable AI Conundrum: How to Bridge the Gap Between Clinicians Needs and Developers Goal." 2023. doi: 10.21203/rs.3.rs-2326665/v1.

- [45] D. Yim, J. Khuntia, V. Parameswaran, and A. Meyers, "Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review," *JMIR Med. Inform.*, vol. 12, no. 1, p. e52073, Mar. 2024, doi: 10.2196/52073.
- [46] T. Minssen, E. Vayena, and I. G. Cohen, "The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models," *JAMA*, vol. 330, no. 4, pp. 315–316, Jul. 2023, doi: 10.1001/jama.2023.9651.
- [47] S. Reed *et al.*, "A Generalist Agent." arXiv, Nov. 11, 2022. doi: 10.48550/arXiv.2205.06175.
- [48] K. Singhal *et al.*, "Large Language Models Encode Clinical Knowledge." arXiv, Dec. 26, 2022. doi: 10.48550/arXiv.2212.13138.
- [49] OpenAI, "GPT-4." 2023.
- [50] J. W. Ayers *et al.*, "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," *JAMA Intern. Med.*, Apr. 2023, doi: 10.1001/jamainternmed.2023.1838.
- [51] M. Beane, "Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail," *Adm. Sci. Q.*, vol. 64, no. 1, pp. 87–123, Mar. 2019, doi: 10.1177/0001839217751692.
- [52] P.-H. Wong, "Cultural Differences as Excuses? Human Rights and Cultural Values in Global Ethics and Governance of AI," *Philos. Technol.*, vol. 33, no. 4, pp. 705–715, Dec. 2020, doi: 10.1007/s13347-020-00413-8.
- [53] H. Sackman, "Summary Evaluation of Delphi," *Policy Anal.*, vol. 1, no. 4, pp. 693–718, 1975.
- [54] C. M. Goodman, "The Delphi technique: a critique," *J. Adv. Nurs.*, vol. 12, no. 6, pp. 729–734, Nov. 1987, doi: 10.1111/j.1365-2648.1987.tb01376.x.
- [55] H. P. McKenna, "The Delphi technique: a worthwhile research approach for nursing?," *J. Adv. Nurs.*, vol. 19, no. 6, pp. 1221–1225, Jun. 1994, doi: 10.1111/j.1365-2648.1994.tb01207.x.
- [56] M. J. Bardecki, "Participants' response to the Delphi method: An attitudinal perspective," *Technol. Forecast. Soc. Change*, vol. 25, no. 3, pp. 281–292, May 1984, doi: 10.1016/0040-1625(84)90006-4.

Supplementary Files

Multimedia Appendixes

Delphi Survey Vignettes to describe ICU tasks and workflow.

URL: <http://asset.jmir.pub/assets/4d52a6f7de1dcfca6d0b1210e5b8fa17.pdf>

Interview Guidelines.

URL: <http://asset.jmir.pub/assets/228e9cc5df5f16da346c802bb2bc9ae5.pdf>

Related publication(s) - for reviewers eyes onlies

50130_Response Letter_Human-AI Teaming in ICU.

URL: <http://asset.jmir.pub/assets/20e75085d0b18d71d13c044dbf3b9702.pdf>