Proceedings Paper:
Jia, Yan, McDermid, John orcid.org/0000-0003-4745-4272, Hughes, Nathan et al. (3 more authors) (2023) The need for the human-centred explanation for ML-based clinical decision support systems. In: Proceedings - 2023 IEEE 11th International Conference on Healthcare Informatics, ICHI 2023. 11th IEEE International Conference on Healthcare Informatics, ICHI 2023, 26-29 Jun 2023 Proceedings - 2023 IEEE 11th International Conference on Healthcare Informatics, ICHI 2023 . IEEE , USA , pp. 446-452.

https://doi.org/10.1109/ICHI57859.2023.00064

The need for the human-centred explanation for ML-based clinical decision support systems

1st Yan Jia
Department of computer science
University of York
York, UK
yan.jia@york.ac.uk

2nd John McDermid
Department of computer science
University of York
York, UK
john.mcdermid@york.ac.uk

3rd Nathan Hughes
Department of computer science
University of York
York, UK
nathan.hughes@york.ac.uk

4th Mark Sujan
Human Factors Everywhere Ltd
Woking, UK
mark.sujan@humanfactorseverywhere.com

5th Tom Lawton
Bradford Royal Infirmary and
Bradford Institute for Health Research
Bradford, UK
tom.lawton@bthft.nhs.uk

6th Ibrahim Habli
Department of computer science
University of York
York, UK
Ibrahim.habli@york.ac.uk

***Abstract*—Machine learning has shown great promise in a variety of applications, but the deployment of these systems is hindered by the "opaque" nature of machine learning algorithms. This has led to the development of explainable AI methods, which aim to provide insights into complex algorithms through explanations that are comprehensible to humans. However, many of the explanations currently available are technically focused and reflect what machine learning researchers believe constitutes a good explanation, rather than what users actually want. This paper highlights the need to develop human-centred explanations for machine learning-based clinical decision support systems, as clinicians who typically have limited knowledge of machine learning techniques are the users of these systems. The authors define the requirements for human-centred explanations, then briefly discuss the current state of available explainable AI methods, and finally analyse the gaps between human-centred explanations and current explainable AI methods. A clinical use case is presented to demonstrate the vision for human-centred explanations.**

***Index Terms*—Explainable AI, Human-centred XAI, CDSS**

I. Introduction

Machine learning (ML) models have been proposed as decision support systems (DSS) in healthcare and there is evidence that these systems can perform as well as, or better than, humans in some circumstances [1]. However, comparatively few systems are in widespread use in healthcare and one of the reasons for this is the challenge of demonstrating safety [2], especially of the ML components. Thus, there is a growing interest in how the safety of such systems can be assured, e.g. by assessing the ML models in their system context [3], or by using explainable AI (XAI) methods to gain confidence that they meet safety objectives [4]. To date, many explanations generated for ML systems are technically focused, although we have argued previously that the forms of explanations should vary with their purpose and the stakeholders [5].

In this paper, we focus on ML-based clinical DSS (CDSS), which is also the most common class of ML system that has obtained regulatory approval in healthcare [6]. Such CDSS are often viewed by regulators as low risk compared with more highly autonomous systems as it is still up to human actors to make the final decision, and the CDSS only provides a recommendation. Therefore, they are often evaluated "stand-alone", assuming that as long as the performance of the ML-based CDSS is good enough, then it will improve clinicians' diagnostic performance. However, some studies [7] have pointed out the paucity of evidence that using ML-based CDSS correlates with improved clinician diagnostic performance and suggested that we should consider human decisions as end points, as the ultimate effect or impact of such systems is achieved by human-AI teaming not AI alone. Therefore, we believe that one of the important means to support effective human-AI teaming is explanation.

Despite the significant progress in XAI, some researchers have already shown that technically-focused explanations, e.g. just showing feature importance, are not satisfying and tend to have limited impact on users' responses to system behaviour [8]. Similarly, referring to probabilities or statistical relationships in explanation is not as effective as people believe [9]. Further, the critics of current approaches argue that XAI should build on existing research in other domains, for example philosophy, cognitive psychology/science, and social psychology [9]. Therefore, in this paper, we will build on psychology and decision making theory to derive the requirements for human-centred XAI.

The rest of the paper is structured as follows. Section II introduces our conceptualisation of human-centred XAI. Section III provides a general introduction to the current state of XAI methods. Section IV uses a clinical use case to illustrate our vision for human-centred XAI. Conclusions are presented in section V.

II. Human-centred Explainable AI

The need to understand how users interact with technology is fundamental for any successful deployment of the technology, especially in complex domains such as healthcare [10]. However, it is not just the interaction that is important to consider, but how the user and the technology form a cognitive unit to accomplish tasks together. This view is known as a Joint Cognitive System [11], which originated in the 1980s and explains how humans and technology use knowledge of one another and the environment to plan and modify their actions. This also relates to the theory of distributed cognition [12], which explains how people navigate complex worlds by sharing cognition with other entities (e.g., the work environment, the tools used, and other people).

These views show it is not enough to understand the roles of the human and technology in isolation, as how they work together within a context reveals the true nature of the task. It is therefore essential to understand both how the human and computer interact during tasks, as well as the wider context where the task is performed. To do so, it is important to use established methods such as Applied Cognitive Task Analysis (ACTA) to elicit user requirements based on the task being analysed and its wider system requirements. Alongside this, other human computer interaction design principles (e.g., human-centred design) can be used in refining general requirements. However, the theory of joint cognitive systems has not yet been widely applied to ML-based systems. There is therefore a need to understand how it could be applied to ML-supported tasks, which includes considering how to design human-centred XAI.

To design human-centred explanations in a joint cognitive system such as a CDSS, a first step is understanding the context; specifically, who to explain to and what goal the explanation has. For a CDSS, the users are clinicians, i.e. expert users, and the goal is to *augment* the clinicians' decision making. In other words, when the predictions from ML-based CDSS are right, we want the explanations to help the clinicians to accept the recommendation (right for the right reasons) and when the predictions are wrong, we want the explanations to help clinicians identify and reject the wrong recommendation in order to achieve the best human-AI teaming performance, i.e. better performance than human alone or AI alone. In order to design human-centred XAI that will work effectively with expert users, two concepts must be understood; how experts process information to make a decision, and how decisions are best communicated. Understanding the former highlights what elements human decision-makers are likely to look for in an explanation, and by extension what an explanation will need to contain to be evaluatable. Understanding the latter highlights the need to match a user's expectations of what an explanation "should" look like.

A. How experts process information to make decisions

The main distinction between experts and novices is in their situation assessment abilities, not their general reasoning skills [13]. Further, experts are able to represent available information about a situation in a deeper, more conceptual, and more abstract manner than novices whose representations are more superficial [14].

Abstract representation involves two capabilities that are particularly relevant for the design of human-centred XAI: the ability to select relevant information to make a decision, and the ability to 'chunk' information together for quicker processing. The former ability is the acquisition of skilled knowledge; the ability to separate relevant from irrelevant information, by considering the entire situation rather than only the tasks involved [15]. What information is considered relevant is variable, but interestingly experts do not simply rely on having a larger quantity of information than non-experts. A review of the literature shows that many experts in a variety of fields only use a small number of cues/features to make decisions [16]. Research shows that medical radiologists use 2-6 cues [17], and medical pathologists use only 1-4 cues [18], highlighting that it is not volume but the *type* of information used which varies between experts and novices. Furthermore, although experts naturally single out relevant information, the presence of irrelevant information can negatively effect decisions made. Experts can struggle to ignore irrelevant information leading to less optimal decisions, known as the dilution effect [19]. Therefore, presenting explanations that contain *too much* information irrelevant to the current situation could potentially decrease effective decision-making, as clinicians may struggle to focus on what is important.

The latter ability involves information being considered at different levels of abstraction, by taking raw data and inputs and grouping them together in meaningful ways. Rasmussen [20] argued that complex work systems can be decomposed into five hierarchical levels of information abstraction. This means that experts not only have a larger knowledge base to draw on in the course of decision making; they can also organise their knowledge in a more conceptual and abstract manner [21]. Therefore, it is possible for expert decision-makers within a work context to use different information abstractions when making decisions. For example, [17] shows some cue patterns appear to be of consistent importance among medical radiologists. Using such patterns allows an expert to overcome the limitations of working memory capacities by 'chunking' information together, learning over time how to store information in memory as a collection of patterns [22]. This is known as chunking theory, where experts have both larger and more numerous chunks available to them in their area of expertise, and these also tend to be more abstract than novices' [23]. Increasingly complex chunks, formed by iterative chunking, allow for rapid identification and processing of patterns. Therefore, experts making a decision are likely to rely on abstracted chunks of complex patterns stored in memory, and are less likely to rely solely on raw data. By extension, for an ML-based CDSS to be meaningfully explainable may require the presentation of information to match the structure of the clinicians' chunks, such as by grouping together data known to be related.

Specifically, [14] provides a more granular description of

the mechanisms that experts use to excel at decision making. In contrast to a novice decision-maker, experts (1) are tightly coupled to cues and contextual features of the environment; (2) have a larger knowledge base that is organised differently from non-experts; (3) engage in pattern recognition; (4) have better situation assessment and representations of problems; (5) have specialised memory skills; (6) self-regulate and monitor their processes; (7) automate the small steps; (8) seek diagnostic feedback; and (9) engage in deliberate and guided practice. The first six features are particularly relevant for designing human-centred XAI.

The abstract representation abilities described above can be seen in many of these features — the selection of relevant information is present in features 1, 3, and 5, whereas the ability to chunk information together is present in features 2, 3, 4, and 5. Therefore, how experts process information to make decisions can be condensed into the following five aspects: the use of **contextual cues**, the use of **selective information**, **chunking information** together, **pattern recognition**, and **monitoring/reflecting on their decisions**. Viewing expert decision-making in this way indicates that human-centred XAI will need to present explanations of how a decision has been reached in ways that are understandable to an expert. This will facilitate user understanding and comparing the explanation against their own decision-making process.

B. How decisions are communicated

It is important to understand how decisions are communicated as well as how they are made. For a user to work effectively with an ML-based system, the system must be able to communicate its decisions and explanations to the user. Many rules for effective communication between people have been shown to apply to human-machine communication (such as between users and chatbots [24]), and have been considered for XAI design in the past, e.g. [25]. Overall, the goal of communication is to exchange not only information, but also the meaning behind information. Both "speakers" work together to allow the sharing of meaning, known as the cooperative principle, that was introduced by the philosopher of language Paul Grice, [26], which is composed of four maxims, as follows: **quantity** (what is said provides sufficient information), **quality** (what is said is genuine and relevant), **relation** (what is said is contextually relevant to the conversation and the current topic), and **manner** (what is said is understandable). These maxims imply that effective communication is an ongoing state, rather than something achieved at one specific time – both parties work together to select the correct information at the correct level of abstraction to be understandable to each other. If one party is unclear, the other can ask for more information. Conversely, a party can indicate they understand enough about the smaller details and wish to talk at higher levels (such as strategies and concepts). Therefore, human-centred XAI should communicate its decisions effectively to a user, which involves the ability to adjust provided explanations both to the needs of the user and the context in which they are given.

C. The requirements for human-centred XAI

As discussed at the beginning of this section, designing human-centred XAI for CDSS may benefit from a foundation based on the theory of joint cognitive systems. Specifically, it is important to understand the task requirements and the wider context when designing a CDSS. For example, what kind of support does a clinician need to enable them to work effectively in its clinical context? In this section, our focus is to derive requirements for the design of human-centred XAI for a CDSS. The purpose is not to replace the well-established joint cognitive systems design principles, instead it is to build on them and add specific requirement for explanation.

Therefore, we derived four requirements for designing human-centred XAI, see Figure 1. The first requirement is to *provide salient and timely information*, which reflects the element "selective information" in expert decision making and the elements "correct", "sufficient" and "relevant" information from effective communication. The second requirement is to *provide different types of explanations*, which reflects the element "being understandable" from effective communication. The third requirement is to *provide different levels of abstraction*, which reflects the element "information chunks" and "pattern recognition" in expert decision making. The final requirement is to *provide interactive explanation*, which reflects that different experts might use different information to make decisions and that effective communication is an ongoing process.

Figure 1 condenses these insights into a framework for designing human-centred XAI, showing how the requirements arise from reconciling how experts make decisions with key elements of effective communication. The requirements above are intended to be general. That means that they need to be refined when designing a specific ML-based CDSS. For example, when designing a ML-based CDSS using structured data, counterfactual explanations might be preferred by experts in order to understand what features to change to achieve a desired outcome. However, when designing a ML-based CDSS using an imaging dataset, a saliency map might be preferred.

III. The current state of XAI

A. Types of explanation

ML encompasses a diverse set of methods, some of which are considered to be inherently interpretable, such as linear models, decision rules or trees, and general additive models. It is worth noting that even with such models, interpreting the results can be challenging for humans, particularly when the model input features exhibit heterogeneity or the dimensionality of the input features surpasses human cognitive capacity, which is a frequent occurrence in healthcare settings. In contrast, other ML models, such as neural networks (NNs), are more complex or opaque, but *post-hoc* explanations can be generated to provide insights into how and why the models make their decisions, even if they are too complex for direct human interpretation. This paper briefly discusses four main *post-hoc* explanation types, but a more comprehensive overview of XAI has been presented in [4].

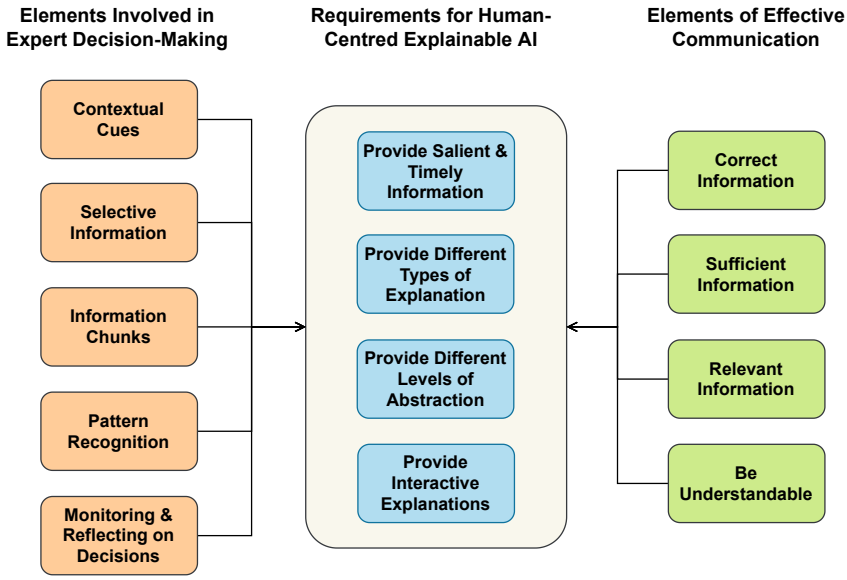


Fig. 1. A framework to inform design of human-centred XAI.

Feature importance explanation involves ranking or scoring all of the input features for a given model. A higher score means the specific feature has a greater effect on the model's prediction. Perturbation or gradient-based methods can be used to obtain these scores. One example of a popular perturbation method is SHAP [27], which is based on Shapley values. Meanwhile, Integrated Gradients is a gradient-based method that calculates the average gradient of the output in relation to each input feature [28].

Explanation by example selects specific instances from the dataset or generates new ones, which can include prototypical examples, counterfactual examples and influential instances. Prototypical examples are able to capture and encapsulate a complex underlying data distribution, making them a useful tool for explaining specific predictions. By identifying the training instance that is most similar to the current instance of interest, they can provide a clear and concise representation of the reasoning behind a given prediction. Counterfactual explanations are hypothetical scenarios that explore "what if" situations. Influential instances are the training instances which have a significant impact on the model or a specific prediction. Note that influential instances are not necessarily representative for the current instance of interest although they are influential in the training process.

Explanation by approximation refers to the use of surrogate models, which are often inherently interpretable, such as linear models, to approximate the complex model, e.g. a NN. Depending on whether the surrogate model is approximating a single prediction or the entire model, it can be further classified as either local or global explanation. For instance, Anchors [29] generate if-then rules for predictions in a particular region by employing a perturbation based approach. LIME [30] trains a local surrogate, e.g. a linear model, to explain an individual prediction by mimicking the complex model behaviour.

Other **visual explanation** techniques can illustrate how input features of a ML model interact with the model prediction or other input features in order to improve model comprehension. Note that the categories presented here are not orthogonal. For example, although LIME is categorised as providing explanation by approximation, it can also provide feature importance explanations.

B. Different levels of abstraction

Turning to deep learning, three levels of abstraction have been explored, which are particularly relevant to producing explanations to expert users. The **first level** is to provide explanations at the input feature level, i.e. explaining what features are important for the ML model to make a specific prediction. For example, a "saliency map", which highlights the pixels that were relevant for a certain image classification by a NN, would be considered to belong to feature level explanation. Many XAI methods produce explanation at the input feature level, e.g. DeepLift [31], and LIME [30].

The **second level** is to provide explanations at the concept level, i.e. explaining the concept that the ML uses to make prediction, e.g. a small "crater" present in the X-ray image. When the number of input features exceeds a certain amount, the ranking of all the input features would not be sufficient to provide human intelligible explanation and as we mentioned above, experts use abstract representation of the information, not only the information itself. Concept-based explanation could address this limitation and provide human-friendly explanation. There are two ways to generate concept level explanations. One is to integrate concepts used by humans into the design of the NN itself. For example, concept bottleneck models map the input features to concepts,

then map the concepts to predict model outputs [32]. Another is to use *post-hoc* interpretations to explain already-trained NNs in terms of high-level concepts, e.g. Kim et al use a linear probe to predict concepts from hidden layers [33]. However, these post-hoc concept-based XAI techniques rely on models automatically learning those concepts despite not having explicit knowledge of them. That means if the models doesn't learn the concepts used by humans, then the approach can fail.

The **third level** is to provide explanation at a strategic level, i.e. explaining the decision strategy that the ML model uses to make predictions. This can reveal whether the ML model's problem-solving behaviours are naive, short-sighted, or well-informed and strategic. For example, Lapuschkin et al uses spectral relevance analysis to characterise the decision behaviour of ML models [34].

Current XAI methods are mostly focused on input feature level explanations. Concept-based XAI methods are better developed for image-based datasets, where it is much easier to show that a concept present in an image is used for a certain prediction task. In comparison, it is much harder to develop concept-level explanations for structured datasets, where it is less obvious how to group raw features as the raw features are already meaningful to humans, e.g. patient blood pressure. We will illustrate this further in our clinical use case in section IV. Further, strategy level explanations are less well explored. Also, although there are methods to provide explanations at different level of abstraction as discussed above, it still might not be possible to provide explanations for all of the levels of abstraction for a given ML-based CDSS.

IV. A CLINICAL USE CASE

In this section, we present a clinical use case to illustrate our vision for human-centred XAI. It focuses on an ML-based CDSS using convolutional neural networks (CNN) to predict extubation readiness for patients on mechanical ventilation in an Intensive Care Unit (ICU). Extubation is the final step in liberating a patient from mechanical ventilation and is a potentially dangerous time for ICU patients as the range of breathing support options from the ventilator is markedly reduced when the endotracheal tube has been removed, and the lack of the tube also means that there is limited access for suction catheters to be passed to remove respiratory secretions. Both early and delayed extubation can cause patient harm. Determining the right time to extubate a patient is a complex clinical task. Reported rates of reintubation in the literature range from 3% to greater than 30% [35]. Although a lot of effort has been made to come up with criteria for extubating patients, e.g. the rapid shallow breathing index (RSBI), there is no consensus on a standardised protocol and such indices are still not very accurate and are unlikely to be sufficient in themselves [36]. This shows the difficulty of predicting extubation readiness, therefore ML can potentially be helpful to support clincians to make such decision.

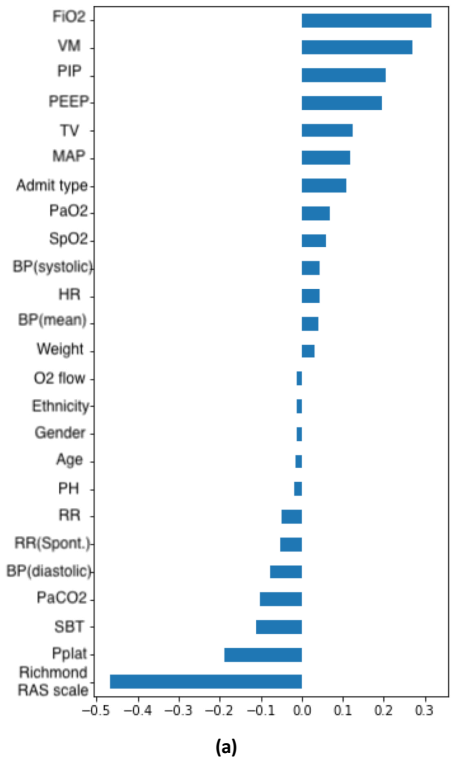
Thus, we developed the CNN model using the MIMIC-III clinical database incorporating 25 patient features, such as demographics, vital signs and laboratory values (see our previous paper [4] for more detail). Here, we used DeepLift [31] to generate feature importance for the CNN model. Figure 2 (a) illustrates what features are important for a particular prediction. The length of the bars shows importance, with those to the right contributing to extubation and those to the left indicating otherwise. Showing features that influenced a prediction can allow clinicians to recognise contextual cues.

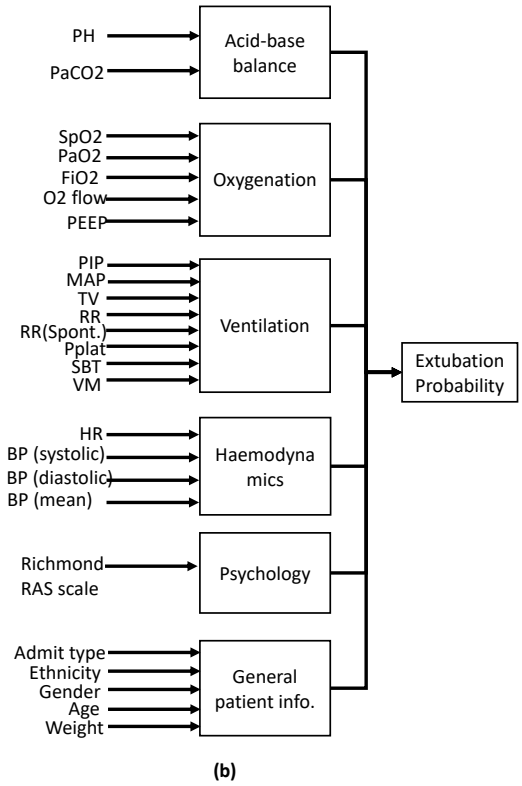
In contrast, Figure 2 (c) illustrates counterfactual explanations for a particular patient. The leftmost column lists the features (the same as in Figure 2 (a)), and the central column shows the current values that lead to the prediction. In this case the patient is not ready for extubation (predicted probability >0.5). The rightmost column shows one counterfactual example – changes in input features that can reverse the prediction (predicted probability <0.5). Some features cannot be changed, e.g. age and ethnicity, and the methods for producing counterfactuals seek to identify the smallest change that brings about the desired change in prediction. Some methods can generate multiple alternative counterfactuals [4].

It is worth noting that feature importance is a more process-based than outcome-based way to explain an ML prediction. In other words, with feature importance clinicians can see that the ML is using the "right" set of features to make its prediction, comparing with their own knowledge, but it is hard to know exactly how they affected that prediction. In contrast, one of the benefits of counterfactuals is that they make clear the effects of feature change on the outcome. Thus, there is benefit in generating different types of explanation so clinicians can select the most effective type for the task at hand.

Further, when clinicians make decisions about extubation, they not only look at each individual feature, they also organise the information based on their clinical knowledge. They are more likely to group or "chunk" the features into a more abstract level to assess extubation readiness. For example, they are more likely to ask "whether the acid-base balance for the patient is appropriate", which indicates that CO2 in the patient blood is normal. This can involve comparing multiple variables such as PaCO2, pH, and SpO2. Also, they may ask whether the patient has adequate oxygenation and adequate pulmonary function to extubate, whether the patient has haemodynamic stability, and whether the patient's psychological status is suitable for extubation. They might also consider other patient information, e.g. patient weight in this case, as being "relevant information". These are the questions that clinicians are more likely to consider to make the extubation decision, rather than just considering the 25 raw features. Thus, just showing the input feature importance might not be sufficient to support clinicians to make decisions as feature importance alone does not fully support "information chunks" or "pattern recognition".

Based on these considerations, we have grouped the raw patient features used in our CNN model based on how clinicians will look at them, as shown in Figure 2 (b). In principle, we can just sum up the raw feature importance based on the grouping and derive more abstract measures of





Input features	original input	Counterfactual example
Admit type	Emergency	--
Ethnicity	White	--
Gender	Female	--
Age	78.2	--
Weight	86.5	--
HR	119	110
RR	24	--
SpO_2	98%	--
FIO_2	100%	40%
PEEP	10	5
MAP	14	10
TV	541	--
PH	7.46	--
RR(Spont.)	0	24
Richmond RAS Scale	−1	0
PIP	21	--
O_2 Flow	5	--
Pplat	19	--
PaO_2	124	118
$PaCO_2$	33	--
BP (systolic)	101	--
BP (diastolic)	65	--
BP(mean)	76	--
SBT	No result	Completed
VM	CMV/ASSIST/ AutoFlow	SIMV/PSV
Predicted outcome	0.93	0.17

(c)

Fig. 2. An example of representing different types and levels of explanation for the CNN model. (a) Feature importance for a specific prediction which ranks the 25 input features. (b) More abstract level of explanation for the model by grouping raw features together in a meaningful way. (c) Counterfactual explanation for a specific patient instance showing what features to change in order to extubate the patient. "–" shows that the features don't change. Legend: PaCO2, arterial carbon dioxide pressure; SpO2, oxygen saturation pulseoxymetry; PaO2, arterial oxygen pressure; FiO2, inspired oxygen fraction; PEEP, positive end expiratory pressure; PIP, peak inspiratory pressure; MAP, mean airway pressure; TV, tidal volume; Pplat, plateau pressure; SBT, spontaneous breathing trial; VM, ventilatory mode; RR, total respiratory rate; RR (spont.), spontaneous breaths; HR, heart rate; BP, blood pressure; Richmond RAS scale, Richmond Agitation and Sedation scale.

importance, e.g. showing how acid-base balance contributed to the decision. For example, this has been done in [37] where they simply summed up the feature importance of the raw features based on which clinical domain they belong to.

This is an evolving use case and we also plan to develop other XAI methods to represent the abstract level explanations, e.g. using fuzzy logic to map the raw patient features to abstract levels, e.g. acid-base balance, then using NN to map the result from fuzzy logic to extubation probability.

We also plan to evaluate the impact of human-centred XAI using two main criteria. We will test task performance in this context and assess experts' mental models. This will help to determine whether or not the explanation provided is "salient and timely", providing the right amount of support and not diluting the experts' decision-making, thus achieving the goal of explanation.

V. Conclusions

This paper has set out the need for human-centred explanations for ML-based CDSS. We have systematically derived four requirements for human-centred XAI from an analysis of how experts make decisions and considering elements of effective communication. These requirements are intended to be general and they will need refinement for a specific CDSS. Then we discussed the current state of XAI methods and the gap between these and human-centred XAI, especially for development of different levels of abstraction. The concepts we have developed have also been illustrated using a clinical use case of a CDSS supporting clinicians about extubation readiness decisions for patients on mechanical ventilation.

We intend to explore further the utility of our concepts, on several fronts. Most importantly, we are engaged in a programme of experimentation with clinicians which will enable us to evaluate and refine the requirements we have developed for human-centred XAI. For example, we will be able to explore what types of explanations clinicians prefer and how interactive explanation can best be supported. Whilst the ideas set out here have been informed by current research on XAI methods, we think it likely that new methods will be needed, and we expect the experimental evaluation of human-centred XAI for extubation readiness will provide insights into the need for new development of XAI method.

We are also interested in understanding how human-centred explanation could improve overall transparency (of the design

and deployment process as well as the machine output) that is needed for assuring the ethical acceptability of AI in safety-critical applications [38], particularly in healthcare. Additionally, we are exploring the concept of accountability for AI in terms of two core features: giving an explanation and facing the consequences [39]. In particular, we are mapping how these two features relate to each other and the conditions under which the former may or may not be necessary for establishing the latter from a moral and legal perspective.

ACKNOWLEDGEMENT

This project is funded by Lloyd's Register Foundation and University of York through Assuring Autonomy International Programme (Project ref 06/22/04) and by the Engineering and Physical Sciences Research Council through the Assuring Responsibility for Trustworthy Autonomous Systems project (EP/W011239/1).

REFERENCES

[1] E. Topol, *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.

[2] D. Higgins and V. I. Madai, "From bit to bedside: a practical framework for artificial intelligence product development in healthcare," *Advanced intelligent systems*, vol. 2, no. 10, p. 2000052, 2020.

[3] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli, "Guidance on the assurance of machine learning in autonomous systems (AMLAS)," *arXiv preprint arXiv:2102.01564*, 2021.

[4] Y. Jia, J. McDermid, T. Lawton, and I. Habli, "The role of explainability in assuring safety of machine learning in healthcare," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 4, pp. 1746–1760, 2022.

[5] J. A. McDermid, Y. Jia, Z. Porter, and I. Habli, "Ai explainability: The technical and ethical dimensions," *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 2021.

[6] D. Lyell, E. Coiera, J. Chen, P. Shah, and F. Magrabi, "How machine learning is embedded to support clinician decision making: an analysis of fda-approved medical devices," *BMJ Health & Care Informatics*, vol. 28, no. 1, 2021.

[7] B. Vasey, S. Ursprung, B. Beddoe, E. H. Taylor, N. Marlow, N. Bilbro, P. Watkinson, and P. McCulloch, "Association of clinician diagnostic performance with machine learning–based decision support systems: a systematic review," *JAMA network open*, vol. 4, no. 3, pp. e211 276–e211 276, 2021.

[8] M. Nagendran, A. Gordon, and A. Uk), "Impact of xai dose suggestions on the prescriptions of icu doctors aldo faisal (a.faisal@imperial.ac.uk)." [Online]. Available: https://2022.ccneuro.org/proceedings/0000328.pdf

[9] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[10] R. J. Holden, P. Carayon, A. P. Gurses, P. Hoonakker, A. S. Hundt, A. A. Ozok, and A. J. Rivera-Rodriguez, "Seips 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients," *Ergonomics*, vol. 56, no. 11, pp. 1669–1686, 2013.

[11] E. Hollnagel and D. D. Woods, "Cognitive systems engineering: New wine in new bottles," *International journal of man-machine studies*, vol. 18, no. 6, pp. 583–600, 1983.

[12] E. Hutchins, *Cognition in the Wild*. MIT press, 1995.

[13] J. Orasanu and T. Connolly, "The reinvention of decision making," *Decision making in action: Models and methods*, vol. 1, pp. 3–20, 1993.

[14] M. A. Rosen, E. Salas, R. Lyons, and S. M. Fiore, "Expertise and naturalistic decision making in organizations: Mechanisms of effective decision making," in *The Oxford Handbook of Organizational Decision Making*. Oxford University Press, 2008.

[15] P. Benner and J. Wrubel, "Skilled clinical knowledge: the value of perceptual awareness, part 2," *Journal of Nursing Administration*, pp. 28–33, 1982.

[16] J. Shanteau, "How much information does an expert use? is it relevant?" *Acta psychologica*, vol. 81, no. 1, pp. 75–86, 1992.

[17] P. J. Hoffman, P. Slovic, and L. G. Rorer, "An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment." *Psychological bulletin*, vol. 69, no. 5, p. 338, 1968.

[18] H. J. Einhorn, "Expert judgment: Some necessary conditions and an example." *Journal of applied psychology*, vol. 59, no. 5, p. 562, 1974.

[19] J. Shanteau, "Averaging versus multiplying combination rules of inference judgment," *Acta Psychologica*, vol. 39, no. 1, pp. 83–89, 1975.

[20] J. Rasmussen, "Human information processing and human machine interaction," *Amserdam: North Holland*, 1986.

[21] P. J. Feltovich, M. J. Prietula, and K. A. Ericsson, "Studies of expertise from psychological perspectives." in *The Cambridge handbook of expertise and expert performance*. Cambridge University Press, 2006, pp. 41—-67.

[22] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956.

[23] W. G. Chase and H. A. Simon, "Perception in chess," *Cognitive psychology*, vol. 4, no. 1, pp. 55–81, 1973.

[24] B. Jacquet, A. Hullin, J. Baratgin, and F. Jamet, "The impact of the gricean maxims of quality, quantity and manner in chatbots," in *2019 international conference on information and digital technologies (idt)*. IEEE, 2019, pp. 180–189.

[25] M. Ribera and A. Lapedriza, "Can we do better explanations? a proposal of user-centered explainable ai." in *IUI Workshops*, vol. 2327, 2019, p. 38.

[26] H. P. Grice, "Logic and conversation," in *Speech acts*. Brill, 1975, pp. 41–58.

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[28] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104*, 2017.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[30] M. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 97–101. [Online]. Available: https://aclanthology.org/N16-3020

[31] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[32] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5338–5348.

[33] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[34] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, p. 1096, 2019.

[35] A. N. Miltiades, H. B. Gershengorn, M. Hua, A. A. Kramer, G. Li, and H. Wunsch, "Cumulative probability and time to reintubation in united states intensive care units," *Critical care medicine*, vol. 45, no. 5, p. 835, 2017.

[36] M. Karthika, F. A. Al Enezi, L. V. Pillai, and Y. M. Arabi, "Rapid shallow breathing index," *Annals of thoracic medicine*, vol. 11, no. 3, p. 167, 2016.

[37] K.-C. Pai, S.-A. Su, M.-C. Chan, C.-L. Wu, and W.-C. Chao, "Explainable machine learning approach to predict extubation in critically ill ventilated patients: a retrospective study in central Taiwan," *BMC anesthesiology*, vol. 22, no. 1, p. 351, 2022.

[38] Z. Porter, I. Habli, and J. McDermid, "A principle-based ethical assurance argument for ai and autonomous systems," *arXiv preprint arXiv:2203.15370*, 2022.

[39] Z. Porter, A. Zimmermann, P. Morgan, J. McDermid, T. Lawton, and I. Habli, "Distinguishing two features of accountability for ai technologies," *Nature Machine Intelligence*, vol. 4, no. 9, pp. 734–736, 2022.