Users as Allies: Why Clinicians Are Key to Explainable Clinical AI *

Hendrik Knoche and Hamzah Ziadeh

Aalborg University, Aalborg 9000, DK

Abstract. Building on Angela Sasse's human-centred vision of users as knowledgeable partners in system performance, this paper examines how human-centred concerns from usable security research can inform explainable artificial intelligence (xAI) in clinical practice. Both domains face the challenge of aligning high-stakes technologies with human cognition, effort, and trust. Drawing on a qualitative case study of xAI applied to stroke care quality registry data, we analyse clinicians' interactions with prototype explanations and identify recurring breakdowns in actionability, comprehension, and trust. We synthesise these observations into a design-oriented typology of clinical variable roles in xAI, distinguishing predictors by actionability, clinical relevance, and data visibility, and relating these distinctions to user cost in descriptive and counterfactual use. Interpreted through a human-centred lens emphasising comprehension, effort awareness, and warranted trust, the typology shows how misaligned explanations shift cognitive burden onto clinicians and invite miscalibrated reliance. We argue that xAI succeeds not when explanations are merely available, but when they reduce the interpretive and vigilance burden placed on users—realising Sasse's vision of users as informing and informed partners.

Keywords: User-centered design · explainable AI · user cost.

1 Introduction

Angela Sasse's work in usable security consistently challenged a deeply rooted assumption in system design: that users are the problem. Through decades of

* This paper is dedicated, with deep gratitude, to Professor M. Angela Sasse, whose work has profoundly shaped how technology engages with its users. Through her scholarship and advocacy—most visibly in her collaborative work on usable security—she helped establish the principle that users are not the enemy, transforming thinking in security and human–computer interaction by foregrounding understanding, trust, and cooperation as foundations of reliable systems. By extending this human-centred ethos into the emerging field of explainable AI, this contribution honours not only her research legacy, but also her enduring influence on how we design for, think with, and learn from the people who use our systems.

research, she demonstrated that many security failures arise not from careless or malicious users, but from systems that demand excessive cognitive effort, misalign with users' goals, or invite misplaced trust.

Explainable artificial intelligence (xAI), particularly in clinical settings, now faces a strikingly similar challenge. Although predictive models achieve impressive performance, their safe and effective use depends on whether clinicians can understand, assess, and appropriately act on model outputs. Explanations that are statistically faithful but poorly aligned with clinical reasoning, actionability, or available data can increase cognitive burden, foster over- or under-trust, and ultimately compromise care. As in usable security, model transparency alone is insufficient if it does not support meaningful human understanding.

In this paper, we draw on human-centred concerns articulated in Sasse's work to examine explainable clinical AI through the lens of user cost and trust calibration. Our contribution is a design-oriented typology of clinical variable roles in xAI, distinguishing predictors by actionability, clinical relevance, and data visibility. By making explicit where model representations diverge from clinicians' mental models, the typology highlights how explanation design choices shape effort, trust, and responsibility—reinforcing the enduring insight from usable security research that users are not the enemy, but essential allies in creating usable and safe systems.

2 Background

Research in explainable Artificial Intelligence (xAI) has increasingly turned toward human-centred approaches that prioritise understanding as a design goal rather than an afterthought. Researchers have argued that explanations should be modelled on human explanatory practices drawn from social and cognitive psychology [9, 10]. Ehsan and Riedl similarly advanced a human-centred xAI framework that treats explanation as a communicative and reflective process [4], while Kulesza et al. emphasised iterative, user-participatory methods to improve explanation usability [8]. Within Responsible AI scholarship, authors such as Dignum [3] and Floridi and Cowls [6] define transparency and explicability as core ethical principles, positioning human understanding as a prerequisite for accountability. The EU's Ethics Guidelines for Trustworthy AI (HLEG, 2019) and the EU AI Act (2024) have since codified interpretability and effective human oversight as design obligations, further institutionalising this shift from post hoc transparency to understanding by design.

While these works collectively frame understanding as essential to responsible AI, few trace its intellectual lineage to earlier human-centred computing traditions. This paper situates contemporary xAI within the trajectory of usable security and media systems research led by Angela Sasse, which consistently foregrounded comprehension, effort, and trust as preconditions for effective system use. These concerns are not unique to usable security, nor were they invented by Sasse; rather, her distinctive contribution was to make them visible, measurable, and unavoidable in domains where system failure is often blamed on users. By

connecting the framing of "users are not the enemy" with current debates on explainability and human oversight, this paper offers a historical and conceptual synthesis: human understanding is not merely an ethical ideal, but a design resource and infrastructure that sustains both security and trustworthy AI.

3 The case: xAI in the context of stroke registry data

To ground our conceptual considerations, we draw on a case study involving xAI applied to reviewing or making predictions about stroke survivors based on RES-Q registry data capturing care quality indicators and patient outcomes. This tabular data reflected routinely collected clinical and process variables used for quality monitoring such as demographics, risk factors, clinical presentation, diagnostic tests, treatments, and outcomes, making it a representative setting for current clinical AI deployments. We developed an interactive xAI prototypes that generated patient-level predictions accompanied by feature-based explanations (SHAP values and their visualizations) and explored their use with practicing clinicians.

We conducted task-based tests in which clinicians were invited to interpret predictions, assess their plausibility, and reason about potential actions or counterfactuals. These sessions revealed recurring points of friction between model explanations and clinical reasoning, particularly around actionability, missing context, and trust. Follow-up discussions with clinical project partners were used to deepen and validate these observations. The qualitative insights derived from this process informed the variable typology presented in this paper, linking concrete interaction breakdowns in xAI use to broader principles of comprehension, effort-aware design, and trust calibration.

The qualitative data revealed that difficulties with explainable AI rarely stemmed from a lack of information, but from misalignments between what the model highlighted and how clinicians reason about decisions in practice. In particular, participants repeatedly distinguished between variables they could act upon, variables that were informative but irrelevant for intervention, and clinically important factors that were absent or only implicitly represented. We synthesize these patterns into a design-oriented typology of variable roles in explainable clinical AI, shown in Table 1. The typology distinguishes predictors by their clinical relevance, actionability, and data availability, and relates these distinctions to the cognitive effort and trust judgments required of (stroke) clinicians when engaging with xAI systems.

While the typology is grounded in observations from the stroke xAI case, its reach likely extends beyond this specific setting. The distinctions captured in the table reflect recurrent tensions identified in Angela Sasse's usable security research: between understanding and compliance, between system demands and user effort, and between appropriate and misplaced trust. Interpreting the typology through these three human-centred design concerns – comprehension over compliance, effort-aware design, and trust calibration – reveals how explainable AI systems can inadvertently shift cognitive burden onto users or invite miscal-

ibrated reliance. In the following, we use this framing, consistently articulated in Sasse's work, as an interpretive lens to unpack how different variable roles systematically shape clinician effort, understanding, and trust when engaging with xAI—both when reviewing explanations of a classification or prediction and when manipulating parameters in what-if analyses.

3.1 Comprehension over compliance

Sasse's work in usable security, including her collaboration with Adams, demonstrated that procedural compliance does not guarantee secure behaviour [1]. Users may follow prescribed steps correctly while failing to internalise their purpose, resulting in behaviour that is brittle, context-dependent, and prone to breakdown when conditions change. An analogous risk arises in clinical explainable AI: users may learn when to accept or reject system outputs based on surface cues—such as prominent features or visual emphasis—without understanding the reasoning, assumptions, or limitations that underlie a prediction. In high-stakes domains such as healthcare, such shallow engagement is particularly problematic, as it can mask misunderstanding behind apparently appropriate behaviour.

The variable typology makes this risk explicit by distinguishing predictors that support clinical reasoning from those that merely appear influential. Variables that are non-actionable or clinically misleading may still be highlighted by explanations, encouraging blind compliance with model output. Designing for comprehension therefore requires treating explanation as an act of sense-making rather than model information delivery. Prior work shows that people seek explanations that are selective, contrastive, and purpose-driven [9, 8]. Applied to clinical xAI, this implies explanations that are aligned with the user's task, indicate why certain factors matter or do not matter, and clarify the boundaries of the model's competence. Evaluation should likewise focus on whether users form accurate mental models—anticipating model behaviour and limitations—rather than on recall, satisfaction, or apparent agreement with system recommendations. In this view, explainability supports comprehension as a prerequisite for responsible action, not as a mechanism for enforcing compliance.

3.2 Effort-Aware Design

Across decades of human-factors research, Sasse highlighted cognitive and attentional effort as a scarce resource. Be it in media consumption[1], security mechanisms that demand excessively complex password rules, or repeated authentications that produce avoidance, fatigue, and workarounds. In clinical xAI, explanations that center on model explanations instead of contextual understanding and sensemaking impose similar burdens. This is especially worrying for junior

[1] With colleagues, Sasse emphasized the notion of user cost in studies of degraded multimedia Quality of Service (QoS), showing that users could maintain task performance by investing additional effort – at personal, including physiological, cost [15]

Table 1. Types of variables in clinical xAI decision-making, model/data status, and mode-dependent user cost

Variable type	**Examples**	**Model status**	**Role in decision-making**	**explanatory mode (user cost)**	**what-if mode (user cost)**
Fixed, clinically relevant	Age, sex, prior stroke, chronic comorbidities	Modeled	Baseline risk; non-modifiable	Low — mentally discount	Medium — cannot adjust; reason about constraints
Changeable, clinically relevant (temporally actionable)	BP, 2 hr BP trend, identified critical time window / state change	Modelled	Actionable at bedside; guideline-referenced	Low — aligns with reasoning	Medium — verify suitability before manipulation
Changeable, not actionable per patient	Time/day of admission, ward occupancy, available experts, arrival prenotification	Modelled	Predictive population-level only	Medium — discard implied actionability	High — manipulating meaningless; wasted effort risk
Changeable, clinically misleading	Aggressive BP lowering, intensified anticoagulation in high-risk patients	Modelled	Safe only in specific patients	Medium — override misleading signals	Very high — counterfactuals could suggest unsafe actions
Granularity-misaligned variable	Any e.g. frailty, functional baseline, adherence, social support	Present but not operationalised	Relevant but ignored by model, clinician might want inclusion in model	High — integrate mentally	Very high — cannot simulate; mental compensation required
Clinically meaningful but not modelled	Any e.g. weight, medical history, depression	Present but not chosen	Relevant but ignored by model, clinician might want inclusion in model	High — integrate mentally	Very high — cannot simulate; mental compensation required
Outcome-relevant but care-process invisible	planned surgery, ongoing procedure, Nursing intensity, rehab quality, timing of supportive care	Not present in data	Influences outcomes but not captured	Very high — reconstruct context mentally	Extreme — cannot simulate or manipulate

clinicians who might be more prone to use and benefit from AI support in clinical decision support systems.

Table 1 highlights where explanations impose additional user cost – such as when clinicians must filter non-actionable predictors, compensate for unmodelled variables, or reconstruct missing context. Effort-aware design responds to these patterns by directing attention toward clinically relevant and temporally actionable variables, while de-emphasising or flagging those that invite unnecessary cognitive work. Established HCI strategies such as progressive disclosure and interactive "why" and "why-not" queries allow users to control explanatory depth and allocate effort where it is most valuable. Reducing avoidable effort preserves respect for users' cognitive resources that lies at the heart of Sasse's work.

3.3 Trust Calibration

Perhaps the most direct bridge between usable security and explainable AI lies in the challenge of trust calibration. Sasse's research repeatedly showed that both under-trust and over-trust lead to system failure: users who distrust controls circumvent them, while those who trust blindly ignore risks. In their seminal paper "Shiny Happy People Building Trust?", Sasse and colleagues proposed the concept of warranted trust—trust that is earned through evidence of competence and reliability [13].

Clinical xAI faces an analogous problem and some steps, e.g., model cards [11] have been taken. Model cards address trust calibration by documenting a model's intended use, performance characteristics, and known limitations, helping users judge when reliance on a system is warranted. While model cards help bound trust at the level of the system, they leave unresolved the variable-level distinctions captured in our typology, effectively shifting the work of interpreting actionability, relevance, and missing context onto the user. Conceptual modelling [14] suggests that junior clinicians may exhibit more fragile trust in AI recommendations—particularly as workload and diagnostic complexity increase—compared with senior clinicians who maintain more calibrated trust by leveraging experience to identify potential model errors. So, when xAI highlight predictors that are statistically influential but not clinically actionable or relevant for a specific patient—so-called "untrustworthy predictors" that can mislead users if interpreted without context [12] because they summarise statistical associations rather than underlying biological mechanisms or cause-and-effect relationships [2]. The goal, as in Sasse's security work, is not to increase trust indiscriminately but to ensure it is warranted—aligned with the evidence at hand. Explanations should communicate why predictors can be trusted (e.g. by providing published evidence about predictors), when they may fail or require checking, and which uncertainties remain, while accounting for both modelled, unmodelled, and data-absent variables.

4 Discussion

Angela Sasse's research transformed security from a technical control problem into a human-centred design discipline. The same transformation is now needed for explainable AI. By extending human-centred concerns articulated in her work into the AI domain, this paper argues that understanding is not a secondary design goal but the structural foundation of ethical and trustworthy systems—one that recognises users as co-constructors of meaning and essential participants in design. Explanations that merely expose system logic without supporting human sense-making risk reproducing the same failures that plagued early security design: cognitive overload, blind compliance, and misplaced trust.

The topology proposed here also complements contemporary governance frameworks such as the EU AI Act [5] and Ethics-by-Design initiatives [7]. These instruments enshrine transparency and oversight as legal duties, yet they rarely specify how understanding can be achieved in practice. Sasse's empirical tradition—grounded in usability testing, cognitive workload assessment, and behavioural evaluation—offers precisely the methodological bridge that policy lacks. In this way, Sasse's legacy of usable security provides an empirical and methodological backing for a genuinely human-centred implementation of ethical AI principles.

5 Summary

Many failures of expert systems and explainable AI in clinical practice can be understood as a transfer of cognitive cost from the system to the clinician. Following Sasse's notion of user cost, clinicians compensate for misaligned explanations by investing additional mental effort, sustaining performance at the expense of fatigue and miscalibrated trust. This paper introduced a design-oriented typology of clinical variable roles that makes these costs visible and links them to breakdowns in comprehension, effort, and trust. The typology can be used by data scientists collaborating with clinical partners to unpack and create the necessary epistemic metadata. Because explainable AI succeeds not when explanations are merely available, but when they reduce the interpretive and vigilance burden placed on users—realising Sasse's vision of clinicians as informing and informed partners rather than passive recipients of system claims. We hope that this will help reframe xAI tools as clinical communication interfaces, not mere transparency artifacts.

Acknowledgments. Funding was provided the European Union as a part of the Horizon Europe research initiative RES-Q+ (grant number 101057603). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Adams, A., Sasse, M.A.: Users are not the enemy. Communications of the ACM **42**(12), 40–46 (1999). https://doi.org/10.1145/322796.322806

[2] Carriero, A., de Hond, A., Cappers, B., Paulovich, F., Abeln, S., Moons, K.G.M., van Smeden, M.: Explainable AI in healthcare: to explain, to predict, or to describe? Diagnostic and Prognostic Research **9**, 29 (2025). https://doi.org/10.1186/s41512-025-00213-8

[3] Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30371-6

[4] Ehsan, U., Riedl, M.O.: Human-centered explainable ai: Towards a reflective sociotechnical approach. In: Stephanidis, C., Kurosu, M., Degen, H., Reinerman-Jones, L. (eds.) HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence. pp. 449–466. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-60117-1_33

[5] European Parliament and Council of the European Union: Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, L 2024/1689, 12 July 2024 (2024), https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[6] Floridi, L., Cowls, J.: A unified framework of five principles for ai in society. Harvard Data Science Review **1**(1) (2019). https://doi.org/10.1162/99608f92.8cd550d1

[7] High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI. European Commission (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, brussels

[8] Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th International Conference on Intelligent User Interfaces. pp. 126–137. IUI '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2678025.2701399

[9] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007

[10] Miller, T., Howe, P., Sonenberg, L.: Explainable ai: Beware of inmates running the asylum. In: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI). Melbourne, Australia (2017). https://doi.org/10.48550/arXiv.1712.00547, https://doi.org/10.48550/arXiv.1712.00547

[11] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 220–229. FAT* '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287596

[12] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778

[13] Riegelsberger, J., Sasse, M.A., McCarthy, J.D.: Shiny happy people building trust? photos on e-commerce websites and consumer trust. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 121–128. CHI '03, Association for Computing Machinery, New York, NY, USA (2003). https://doi.org/10.1145/642611.642634

[14] Shamszare, H., Chaudhry, Z., Berenji, M., Choudhury, A.: Conceptualizing clinicians' trust in artificial intelligence as a function of their expertise, workload, patient outcome, diagnosis difficulty, and ai accuracy: A systems thinking approach. IEEE Access **13**, 119601–119618 (2025). https://doi.org/10.1109/ACCESS.2025.3586555

[15] Wilson, G.M., Sasse, M.A.: Listen to your heart rate: counting the cost of media quality. In: International Workshop on Affective Interactions. pp. 9–20. Springer (1999)