Artificial Intelligence In Medicine 155 (2024) 102933

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

A Human–AI interaction paradigm and its application to rhinocytology

Giuseppe Desolda [a,*], Giovanni Dimauro [a], Andrea Esposito [a], Rosa Lanzilotti [a], Maristella Matera [b], Massimo Zancanaro [c,d]

[a] Department of Computer Science, University of Bari Aldo Moro, Via E. Orabona 4, Bari, 70125, Italy
[b] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Ponzio 34/5, Milan, 20133, Italy
[c] Department of Psychology and Cognitive Science, University of Trento, Corso Bettini 31, Rovereto, 38068, Italy
[d] Fondazione Bruno Kessler, Povo, Trento, 38123, Italy

ARTICLE INFO

Keywords:
Human-centered artificial intelligence
Explainability
User control
Rhinocytology

ABSTRACT

This article explores Human-Centered Artificial Intelligence (HCAI) in medical cytology, with a focus on enhancing the interaction with AI. It presents a Human–AI interaction paradigm that emphasizes explainability and user control of AI systems. It is an iterative negotiation process based on three interaction strategies aimed to (i) elaborate the system outcomes through iterative steps (*Iterative Exploration*), (ii) explain the AI system's behavior or decisions (*Clarification*), and (iii) allow non-expert users to trigger simple retraining of the AI model (*Reconfiguration*). This interaction paradigm is exploited in the redesign of an existing AI-based tool for microscopic analysis of the nasal mucosa. The resulting tool is tested with rhinocytologists. The article discusses the analysis of the results of the conducted evaluation and outlines lessons learned that are relevant for AI in medicine.

1. Introduction

Although there is an increasing awareness of the potential of Artificial Intelligence (AI), a substantial challenge persists in harnessing its benefits while ensuring reliability, safety, and trustworthiness for humans [1]. AI is often approached with a focus on autonomy and efficiency in decision-making [2]. Yet, while high autonomy can be advantageous, it carries inherent risks [3,4].

The emerging field of Human-Centered Artificial Intelligence (HCAI) suggests employing AI to support and enhance human cognitive capabilities rather than replacing them. In this respect, a paradigm shift from an algorithm-focused view to a human-centered perspective, integrating Human–Computer Interaction (HCI) strategies for designing and testing, is necessary [5]. A crucial aspect of HCAI is the emphasis on *user control*, promoting a new relationship between humans and machines to design systems that are not only efficient and autonomous but also beneficial to humans in various respects [6].

User control can be achieved by building AI models and algorithms that ensure transparency in the system behavior, a concept known as model *explainability* [7]. Although this approach may enable users to comprehend and trust system decisions, a challenge arises as explanations are often tailored for AI specialists, making them less meaningful for end users who are experts in their domain but lack AI expertise. Another avenue for achieving user control is through interactive manipulation of parameters influencing system behavior [8]. In the HCAI vision, user control and system autonomy are not seen as conflicting forces but rather as dimensions that need careful calibration when designing AI-based systems. This implies supporting domain experts in negotiating and reconfiguring algorithm outcomes, and ensuring they play an active role in shaping the system behavior over time.

Calibrating control with efficiency is extremely important when using AI for decision-making in medicine. In this domain, the need to trust AI as a supporting tool has to be counterbalanced, allowing physicians to skeptically inquire about AI assistance to preserve their judgment [9] as well as their professional competence [10].

Despite several examples in the literature, a comprehensive design framework is still lacking. The research proposed in this article aims to fill this gap by presenting a conceptual framework that emphasizes a new notion of explainability embedded in a full-fledged negotiation process for users to understand and modify system behavior iteratively. Specifically, in 2022 we started to explore three strategies to interact with AI systems: *Clarification*, *Negotiation*, and *Reconfiguration* [11]. Clarification involves explaining the AI system's behavior or decision directly, providing alternatives to deep learning systems' "black-box" nature. Negotiation involves reaching system outcomes

* Corresponding author.
E-mail addresses: giuseppe.desolda@uniba.it (G. Desolda), giovanni.dimauro@uniba.it (G. Dimauro), andrea.esposito@uniba.it (A. Esposito), rosa.lanzilotti@uniba.it (R. Lanzilotti), maristella.matera@polimi.it (M. Matera), massimo.zancanaro@unitn.it (M. Zancanaro).

https://doi.org/10.1016/j.artmed.2024.102933
Received 14 December 2023; Received in revised form 17 July 2024; Accepted 19 July 2024
Available online 22 July 2024

through iterative steps driven by user-AI interaction, offering strategies for progressive decision segmentation and recalibration. Reconfiguration allows users to trigger simple retraining of AI models based on new examples or user feedback, providing adaptability within reach of non-expert users. These three broad strategies have been experimented in a main case study in the medical domain, aiming to derive grounded principles of human interaction with AI systems.

The challenging context of the case study is medical cytology, addressing the problem of AI-supported microscopic analysis of cells contained in the nasal mucosa. Our exploration has been based on an existing tool called Rhino-Cyt [12,13]. We designed and deployed a new prototype of Rhino-Cyt, which adopts an interaction paradigm that allows physicians to use the three strategies. The new Rhino-Cyt was evaluated with real users, i.e., physicians specialized in rhinocytology, to investigate if and how the three strategies suggested by the proposed conceptual framework improve the interaction of physicians with the system. The findings reveal the value of the devised strategies and disclose significant aspects that are still underexplored in the literature. These include the importance of offering explanations that can be customized and accessed "on demand", i.e., without presuming their acceptance by every user and at every stage of the interaction with the AI system. We also learned lessons for AI in medicine that can contribute to the broader understanding and application of HCAI principles.

This article is structured as follows. Section 2 discusses the rationale and background of our work. Section 3 outlines the methodology adopted for the human-centered design of the new interaction paradigm. Section 4 introduces the conceptual framework for Human–AI interaction that we applied to redesign Rhino-Cyt and validated through user studies. Section 5 illustrates the interviews conducted with rhinocytology experts to gather initial requirements for designing the new interaction paradigm. Section 6 then details the Rhyno-Cyt redesign, grounded in the conceptual framework for Human–AI interaction introduced in Section 4. Section 7 presents the study that investigated the impact of the three strategies on the physicians' interaction with Rhino-Cyt and Section 8.1 outlines the lessons learned. Finally, Section 9 concludes the article by highlighting the study limitations and suggesting directions for future research.

2. Rationale and background

With the rise of Machine Learning (ML) and AI systems, algorithms that automatically extract information, learn from data and act in the world without human intervention can be developed. In several tasks, this may bring advantages in terms of efficiency and performance [14], but also risks, exacerbating the drawbacks of knowledge bias and lack of trust by the final recipients of algorithmic decisions [3]. To amplify the advantages and avoid the drawbacks, a paradigmatic change is needed to design and develop this new type of system. Recently, "Human-In-The-Loop" approaches in ML processes [15], the HCAI perspective [1,3], and the Interactive Human-Centered AI [16] aim to propose methods to design new interaction paradigms that can amplify, augment, and enhance human performance, in ways that make systems reliable, safe, and trustworthy.

A fundamental element in HCAI is the control by end users. As envisaged by a leading AI researcher [6], a new relationship between humans and machines is needed to design machines that are "not just intelligent but also beneficial to humans". In high-stakes domains, such as healthcare, full automation is often undesirable due to safety, ethical, and legal concerns. End users' control is thus needed. This could be fostered by means of AI models and algorithms that grant transparency of the system behavior, making it easy for the end users (and all the stakeholders, in general) to understand and trust the system decisions, the so-called model Explainability [7] and explainable AI [17].

Reliable user control of the system can also be achieved by granting an interactive manipulation of the relevant parameters determining the system behavior [16]. In the HCAI vision, user control and system autonomy are not considered to be opposing each other, but rather as two dimensions to be adequately calibrated when designing intelligent systems beneficial to people [3]. Users should be enabled to take advantage of the power of AI algorithms, but the importance of the knowledge that users, as domain experts, possess must not be neglected. For example, Cai et al. present an ML-based tool to visually retrieve medical images (tissue from biopsies) from past patients [8]. The tool supports medical decisions with new patients, empowering the physician to cope with the search algorithm on the fly, and communicates what types of similarity are most important in different situations. This interaction between the human and the system determines a step-wise refinement that increases the diagnostic utility of images found, as well as the user's trust in the algorithm.

The opportunity for the users to modify the system behavior and adapt it to their needs (the so-called End-User Development [18,19]), possibly acting on the system's AI models, is crucial in the long term for real empowerment in the use of AI systems. Meta-design principles [19] must be adopted to define a methodological framework in which developers and AI specialists do not design a rigid system; rather, they provide a scaffolding environment where adequate model explanations can empower domain experts to reconfigure algorithm outcomes through negotiation. In [20], the authors discuss Interactive ML techniques enabling model reconfiguration through experts' intervention in medical scenarios. However, they also highlight the need for HCI methods to identify adequate interaction paradigms.

2.1. Explainable AI

The field of eXplainable AI (XAI) addresses the need for transparency and interpretability in AI systems, which is relevant to granting user control. XAI algorithms provide explanations for the decisions made by AI systems, enabling users to understand and trust the system's outputs [17]. However, explainability is generally considered a means to highlight technical features characterizing the performance of AI models. Methods proposed in the literature to open black-box models identify explainability strategies [21]. Important issues remain open. In particular, different scientific communities address explainability from different perspectives. The explanations provided by the AI community are mainly directed to AI specialists. The HCI community considers these approaches inadequate since they are not meaningful to the end users, who are possibly domain experts but not AI experts [5,22]. Furthermore, special care must be given to XAI in the field of medicine. Several studies highlight the complex challenges associated with the use of explanations in medicine, stressing the need for caution due to the potential inaccuracy and irrelevance of explanations [23–25].

An additional aspect is that current research on XAI does not provide clear guidance on generating feature-based explanations starting from Convolutional Neural Networks (CNNs). In a 2018 survey, Guidotti et al. highlighted GradCAM [26] as a way to extract information on pixels relevant to the classification task [21]. The current state-of-the-art computer vision model, the Vision Transformer [27], provides a similar capability through the attention mechanism [28]. However, these techniques fail to generate explanations based on features deemed relevant by the end users (i.e., the physicians). Creating satisfactory explanations requires a novel architecture.

Although applicability to CNNs may vary, various techniques allow the post-hoc generation of feature-based explanations (e.g., SHAP [7]). However, the main drawback of using post-hoc explanation techniques lies in a potential lack of fidelity to the original model's computation (as, if that was the case, the explanations would equal the original model, making it white-box) [29]. Furthermore, explanations using post-hoc techniques may provide complex or partial explanations that do not fully allow comprehension of the model's inner workings. Thus, the most effective way to provide feature-based explanations without

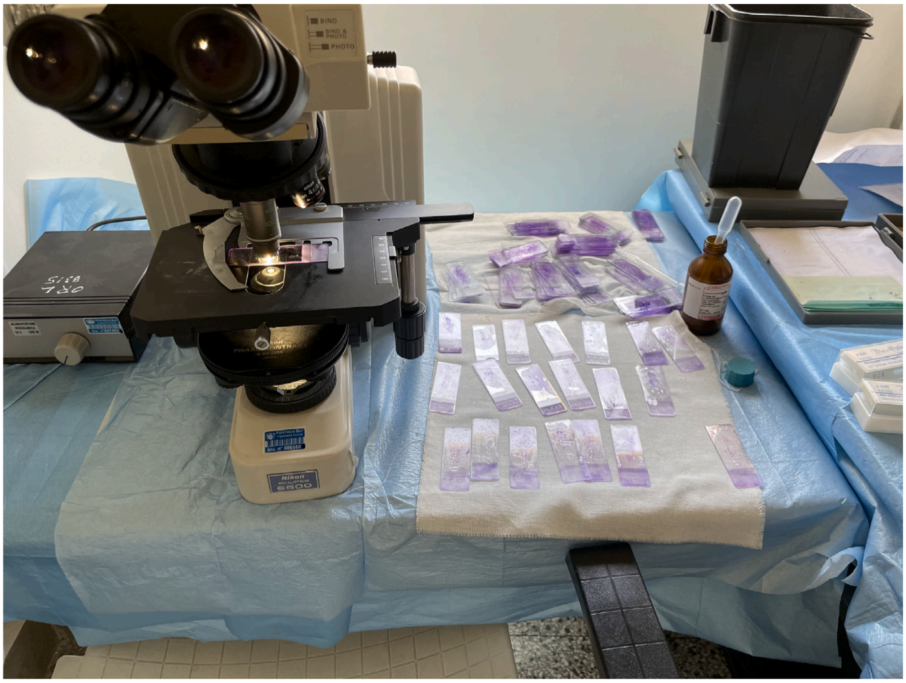


Fig. 1. The typical process of rhinocytology involves direct observation under the microscope, requiring considerable effort by the physician.

using post-hoc techniques involves adopting white-box models (such as decision trees or knowledge-based expert systems).

To overcome these issues while still enabling an adequate level of explainability, in line with recent work proposed in the literature [30], our solution adopts a mixed approach in which black-box models are used for feature extraction while white-box models, e.g., decision trees, generate explanations on top of the extracted features.

2.2. HCAI in medicine

The state of the art in HCI for AI in medicine is rapidly evolving, with researchers exploring various aspects of this intersection. Several studies have investigated the onboarding needs of medical practitioners for Human–AI collaborative decision-making [8]. These studies highlight the importance of providing medical experts with the necessary information when introducing them to diagnostic AI assistants. Additionally, patient apprehensions about using AI in healthcare have been examined, emphasizing the need to address concerns and build trust [31].

One of the key challenges in HCI for AI in medicine is decision-making. While AI systems demonstrate strong predictive performance, full automation is often not desirable [32]. The impact of the COVID-19 pandemic on stroke care has been evaluated using AI, highlighting the need for continuous monitoring and surveillance [33]. Understanding the expectations and requirements of physicians for future AI applications is crucial for successful implementation [34].

Explainability and trustworthiness are critical factors in adopting AI in healthcare. The explainability of AI systems has been recognized as essential, and principles and guidelines have been developed to guide the application and evaluation of AI in medicine [35–37]. The challenges of delivering trustworthy AI in healthcare have been explored, emphasizing the need for transparency and accountability [38].

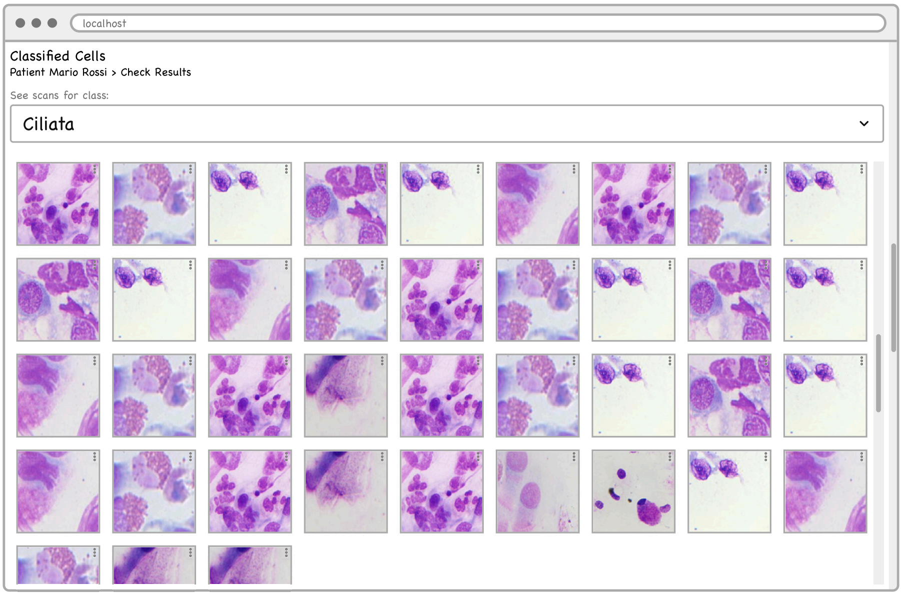
Overall, the state of the art in HCI for AI in medicine is focused on addressing the specific needs and challenges of the healthcare domain. Researchers are working toward developing AI systems that are explainable, trustworthy, and aligned with the requirements of medical professionals and patients. The integration of AI in healthcare has the potential to revolutionize diagnostics, treatment planning, and patient care, but it requires careful consideration of ethical, legal, and social implications. Future research in this area will continue exploring novel applications, improving user experiences, and addressing the challenges associated with adopting AI in healthcare.

2.3. Rhino-Cyt: an AI-enhanced system supporting rhinocytology diagnosis

Our investigation of HCAI paradigms has been organized around a case study in the challenging context of medical cytology, and more specifically, rhinocytology, by tackling the problem of AI-supported microscopic analysis of cells contained in the nasal mucosa [12,13]. Unlike what happens in other medical fields, for example, hematology, nasal cytology does not yet benefit from a network of public or private laboratories that carry out in-depth analyses quickly and at a low cost. Therefore, the diagnostic process is mainly based on direct observation under the microscope, which requires a prolonged effort by rhinocytologists (Fig. 1).

Modern scanning systems for cytological preparations and new affordable digital microscopes enable software systems to be designed to support physicians' activities [39]. By exploiting these capabilities, Rhino-Cyt employs AI models to automate the cytological examination. It encodes a CNN to automatically identify and classify cells in a nasal cytological preparation based on a digital image of the preparation. Compared to standard approaches automating cell counting, Rhino-Cyt aims to move from the current semi-quantitative estimation to a quantitative one, which is more precise and valuable on a scientific level for standardization, to catalog cellular elements and get a more accurate diagnosis in the shortest time. These changes may help in the more widespread use of nasal cytology, a diagnostic investigation that has not yet been widely adopted by the new generation of physicians.

Rhino-Cyt segments histological samples of the nasal mucosa, identifying and classifying individual cells [13] based on nine cytotypes [39]: (i) ciliated, (ii) muciparous, (iii) basal cells, (iv) striated cells, (v) neutrophils, (vi) eosinophils, (vii) mast cells, (viii) lymphocytes, (ix) metaplastic cells [13]. Fig. 2 illustrates how Rhino-Cyt visualizes histological samples classified for a specific cytotype, e.g., *ciliata*. For each cytotype, Rhino-Cyt then supports the cytological examination, producing the cell count.



Classified Cells

Patient Mario Rossi > Check Results

See scans for class:

Ciliata

Fig. 2. The Rhino-Cyt interface.

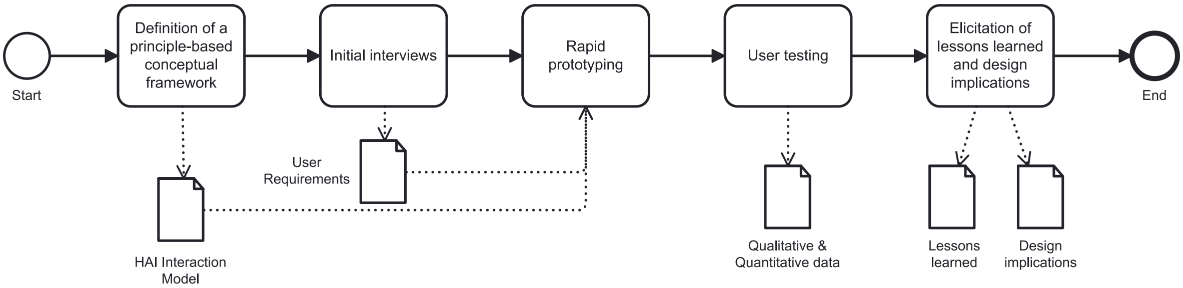

Start
Definition of a principle-based conceptual framework
Initial interviews
Rapid prototyping
User testing
Elicitation of lessons learned and design implications
End
HAI Interaction Model
User Requirements
Qualitative & Quantitative data
Lessons learned
Design implications

Fig. 3. The design process.

Despite its high accuracy [12], Rhino-Cyt might still classify cells incorrectly. Physicians may need to intervene in the model's decisions, assuming control over the cytological examination, to achieve more accurate results. As better illustrated in the following sections, this aspect highlights the importance of designing paradigms that can lead to the final diagnosis through a process of successive steps, with the intervention of the rhinocytologists. It is fundamental to investigate and design interaction mechanisms able to sustain a high level of automation while also simplifying the examination process by providing the domain experts with the capability of understanding, controlling, and reconfiguring the system behavior [8].

3. Design process

Our work follows a *research-through design* approach [40], which emphasizes the early involvement of users, also thanks to the production of prototypes as vehicles for inquiring about foundational aspects of a research challenge. This human-centered design approach aligns with other work on software development in medicine (e.g., [41]). Fig. 3 summarizes the process for the (re)design and evaluation of Rhino-Cyt, which aimed to provide not just a better interface for the tool but also relevant lessons for the design of AI-based tools for decision-making in medicine.

As reported in Fig. 3, we started by grounding our design in already established principles [11], namely (a) the centrality of the control, (b) the need to balance explainability with interactive manipulation of parameters to guide the system's behavior, and (c) the need to enable model reconfiguration. In this first phase, we sketched alternative interfaces for Rhino-Cyt to identify how to accommodate those different principles in an organic view.

In the second phase, we involved rhinocytology experts in semi-structured interviews. The aim was to deepen the diagnostic procedures they employ, also in relation to the automation of cytological examination already coded in Rhino-Cyt, to identify needs and values and explore the potential role of an AI-based system in aiding their work.

The principle-based alternative ideas for the redesign allowed us to prototype a new user interface on top of the Rhino-Cyt system. This phase had an iterative structure and included several occasions of involvement of a sample of rhinocytologists in discussing the opportunity (and the potential risks as well) of adopting new technologies and new working practices. The availability of Rhino-Cyt's already-existing datasets, models, and backend algorithms permitted to immediately focus on new interaction techniques in an effective co-design process.

Finally, through a usability test in conjunction with a thinking-aloud protocol, we compared two versions of the Rhino-Cyt system and distilled the lessons learned through the entire design process.

4. Conceptual framework for Human-AI interaction

Despite the novel contributions in the field of HCAI illustrated in Section 2, the literature still lacks a comprehensive framework able to guide designers in creating HCAI systems. Our work aims to fill this gap: its ultimate goal is to propose design models and methods promoting *explainability techniques* that can be adequate for non-technical people besides being useful to their understanding of the AI system, and new paradigms for the interaction between the human and the system that can enable a *progressive exploration* of the available data. These elements should shape up a *negotiation process* that can empower the human not only to understand the reasons that determine a specific system behavior, but also to intervene and modify it through iterative reconfigurations. Negotiation is an important aspect of Human–AI interaction, as it promotes an iterative process that allows users and AI systems to reach mutually satisfactory outcomes [11]. Unlike a one-sided interaction where the AI simply outputs results based on predefined algorithms, negotiation involves a dynamic and continuous exchange where both the user and the AI adapt and respond to each other's inputs. This synergy can be achieved through adequate interaction paradigms, which are the focus of our research.

Recently, industry players, i.e., Google and Microsoft, have proposed guidelines, toolkits, and design patterns to build the interaction with AI tools. They address explainable AI features but, to the best of our knowledge, they specify what to explain while they do not provide guidelines on the specific strategies to be adopted to provide explanations. The interaction process, in which explainability can allow the users to control the system and trigger iterative reconfigurations, is not addressed. In particular, our analyses of current research highlighted three lacking design dimensions that, if investigated, can provide building blocks for the creation of HCAI systems:

1. *Clarification*: involves providing clear, usable explanations about the inner workings of an AI system, enhancing user comprehension of its behavior and decisions. It implies strategies for explanations that can be meaningful and accessible to the target users. It aims to move away from the "black-box" model typical of deep learning systems (e.g., [42]), where many existing methods reveal technical aspects of the AI model but fail to be helpful to those without a technical background [43]. It aims to propose new models of explanations, which should consider the needs and characteristics of users and the specific requirements of the application domain, as identified through human-centered design methods.
2. *Reconfiguration*: refers to enabling non-technical users to initiate the retraining of AI models easily. This becomes relevant when users identify incorrect predictions or when there is a need to integrate new data into the model (e.g., [20]). Retraining an AI model, to make it adapt and evolve, is inherently technical and requires continual learning techniques to facilitate the incremental absorption, updating, and application of new knowledge [44]. However, equally crucial is the establishment of effective interaction paradigms that empower users to contribute their knowledge to the retraining process.
3. *Iterative Exploration*: relates to reaching the system outcome through a sequence of iterative steps driven by the interaction between the user and the AI system. It involves meta-strategies designed to break down the decision-making process into progressive, manageable steps that can favor human-in-the-loop paradigms [16] and provide means to assess and recalibrate the request to the system (e.g., [8]).

These three dimensions contribute to building a scaffolding layer for negotiation that fosters Human–AI interaction as a *dialog* for mutual comprehension. Negotiation is not just a sequence of isolated actions but a process characterized by continuous feedback and adaptation: users progressively engage with clarifications and AI model reconfiguration to understand and control the system. Concurrently, the system adapts and evolves, leveraging the knowledge users contribute through their interactions. Negotiation thus lies in the overall Human–AI interaction, characterized by an alternating sequence of iterative exploration, clarification, and reconfiguration actions (see Fig. 4).

While AI represents a novel design material [46,47], the negotiation process outlined above can still be understood through foundational HCI principles. Referring to Norman's execution gulf, where the users try to understand how to operate with a system [45], adequate intervention mechanisms [16] can constitute the channel for the user to control the AI model. These mechanisms are meant to improve the user's perception of the automated process outcomes, suggest options for intervention, and allow the user to adapt the behavior of the currently running processes with immediate effect [4]. As for Norman's concept of the evaluation gulf, where the users try to figure out what happened after their actions [45], it is crucial to support users in *perceiving*, *interpreting*, and *evaluating* the AI system's status. In particular, the key to interpretability is the provision of explanations for specific outcomes in a manner that users can easily comprehend. Clarifications, when highlighting incorrect system behavior, may prompt requests for reconfiguration. Thus, for a successful negotiation cycle, it is essential to offer explanations that are not only accessible when needed but also easily interpretable. As Section 4 will illustrate, in our conceptual framework intervention mechanisms introduce actions for the user to give two types of input to the AI model: (i) *requests for clarifications*, when the user wants to understand the reasons behind a given outcome and (ii) *indications for reconfigurations*, when the user identifies incorrect outcomes and tells the system how to change its behavior.

For these mechanisms to be effective, they must emphasize simplicity and minimality, mirroring the language of the users to let them form their intentions and identify and carry out the actions to control and reconfigure the AI model. The users should also be in control, being enabled to explicitly ask for both clarification and reconfiguration. Consequently, adopting human-centered design methods, and involving the target users in the design process becomes crucial. This ensures the creation of intervention mechanisms that are not only functional but also usable by the intended end users.

The following sections will illustrate how this conceptual framework has informed and guided the following design activities, suggesting relevant dimensions that are critical for identifying constituent elements promoting understandability and control in a paradigm for Human–AI interaction.

5. Interviewing experts in rhinocytology

In November 2022, we started our investigation by involving rhinocytology experts in semi-structured interviews aimed to confirm the diagnostic procedure, already coded by Rhino-Cyt, and explore the potential role of an AI-based system in aiding their work. Thus, three experts in rhinocytology were interviewed as a preparatory phase more akin to co-design than to requirement analysis. Specifically, the goal of this initial phase was to set the objectives of the redesign of Rhino-Cyt by instantiating an approach that, while mainly principled-based, aimed at being informed by experts in the field, with a specific emphasis not just on their procedures and practices, but rather on their attitude toward AI as a new design material [47].

5.1. Participants, data gathering, and data analysis

The three rhinocytologists work in hospitals and collaborate with their local universities. They signed a digital consent form to permit the collection of audio recordings and the management of sensitive data. Participants were not given any remuneration or reward for participating in this study.

The semi-structured interview was composed of 5 sections (see Appendix A for details). After welcoming the interviewee, the first section asked about the interviewees' medical specialization, where they work, and their experience in rhinocytology. The second section is related to information about their patients and the process of arriving at a diagnosis. The third section concerns the cell population process, how it is performed, and if and what tools the physicians use. The fourth section investigated the physicians' expectations of using an AI-based system to support their work, to what extent they trust the system, and their interest in understanding how the system works. The fifth

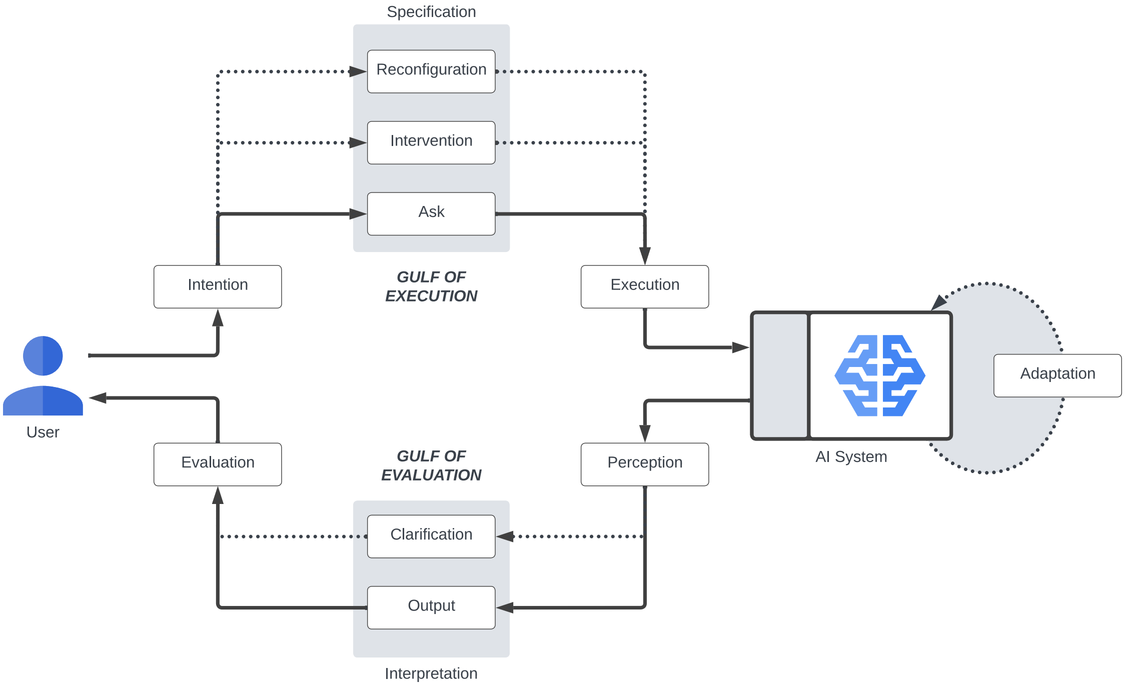


Fig. 4. Structuring the interaction with AI systems (based on Norman's *evaluation* and *execution gulfs* [45]).

section pertained to the physicians' opinion about the AI technological revolution and how they envision the future usage of an AI-based system in their work.

Two researchers were involved in each interview: one researcher served as the interviewer, and the other one assisted by taking notes. The interviews were audio-recorded. Each interview was transcribed before analysis. An inductive thematic analysis of the collected data was performed. Two researchers independently examined the interview transcripts and analyzed them in terms of themes. The interrater reliability was 70%. The remaining 30% of the results were discussed until a consensus was reached.

5.2. Results

Table 1 presents the two primary themes and their corresponding codes derived from the thematic analysis of the answers provided by the rhinocytologists. Specifically, the theme of "*Traditional Rhinocytological Analysis*" delineates the core activities undertaken by physicians in establishing a diagnosis, while "*AI-enhanced Rhinocytological Analysis*" elucidates the principal implications physicians anticipate upon integrating AI-based systems into their practices. The first theme encompasses various aspects, including the patients the physicians receive, the clinical data they collect, and the main important diagnostic investigation the rhinocytologist performs to arrive at an accurate diagnosis, i.e., the cell population process. The theme of "*AI-Enhanced Rhinocytological Analysis*" highlights the factors physicians perceived as pivotal in seamlessly integrating an AI-based system into their practice. This encompasses the desired level of support from AI, the trust physicians wish to have in the system, as deriving from their control over the system decisions, and the assurance that all AI systems should be easily comprehensible.

5.2.1. Themes for the Traditional Rhinocytological Analysis

Patients. Their patients are people of all ages with symptoms such as rhinorrhea and nasal obstruction for some months. Generally, the patients have already consulted other specialists, mainly allergological ones, without resolving their situation. The patients may also be patients of their colleagues who are not rhinocytologists.

Table 1
Themes and codes identified in the thematic analysis.

Theme	Code
1. *Traditional Rhinocytological Analysis*	1. Patients 2. Diagnosis process 3. Cell population process
2. *AI-enhanced Rhinocytological Analysis*	1. Support 2. Trust and Control 3. Understandability

Diagnosis process. When the rhinocytologist receives the patient, they collect the patient's data (e.g., age, symptoms, possible allergy, previous visits to the ENT) in a record. The most critical data is whether the patient is allergic; if yes, the physician notes what they are allergic to. Then, the physician proceeds with the rhinocytological analysis. Along with this diagnostic investigation, three other activities are carried out: patient anamnesis, endoscopy, and rhinometry. An interviewee claimed, "*Each activity, taken individually, says nothing but all together defines the diagnosis*".

Cell population process. The cell population process is an important step of rhinocytological analysis, starting with nasal cytology, which consists of scraping cells from the nasal mucosa that are placed on a slide. The physicians collect 7/8 slides. These slides are then analyzed with a microscope for the cell population process. Cells taken by nasal scraping are gently swiped onto a slide and left to dry in a dedicated box. After the slide is dry, the physician stains it using the May-Grunwald Giemsa method [39]. This step takes about 30 min. At this point, the slide is ready to be analyzed under a microscope at 1000x magnification for a total of 50 microscopic fields. The cells on the slide are classified using the standardization described in the Atlas of Rhinocytology [39]. It is a semi-quantitative classification. The physician looks at the four types of inflammatory cells: neutrophils, eosinophils, mast cells, and lymphocytes. The analysis of a slide takes, on average, 15 min; even if the patient situation is clear from reading the first fields (e.g., the physician finds a large number of eosinophils in the first fields, that means inflammation is ongoing), the physician still reads all 50 fields. The physician uses a table with the cell classes on the rows. Each time the physician recognizes a cell type, they mark a tick in the class cell to which the cell belongs. Ultimately, they count the number of cells in each class. An interviewee said: "*The result obtained is similar to that of a blood count*".

5.2.2. Themes for the AI-Enhanced Rhinocytological Analysis

Support. All interviewees would appreciate using an AI-based system that performs the cell population process in their place. An interviewee said, *"If a system comes along that reads the slide well, it would be nice"*. Another interviewee added, *"Beyond the passion of reading a slide, having the possibility of the system reading it and giving me an automated answer would be interesting. Then, the physician can still look at the slide. The two things are not incompatible!"* The last one added: *"After all, we do not risk losing our art"*. The interviewees reported three main reasons why they consider using an AI-based system useful for the cell population stage, which refers to *automation*, *increased accuracy*, and *easy comparison between different diagnoses*. First, as an interviewee pointed out, *"Automating an operator-dependent technique, such as the cell population process, is very valuable"*. Time is one of the physicians' more critical resources, so this automation would save them time. Regarding classification time, an interviewee said, *"The classification time is irrelevant if this can affect the cell classification. Once the physician has taken the sample from the patient, he will meet him again after 10–15 days to make the diagnosis. Thus, if the system needs more time to be more accurate, there is no problem. There is no hurry!"* Another interviewee added, *"It's better to have more precise and standardized data rather than one that is quick but maybe leaves something behind"*. Another interesting aspect that all three physicians highlighted is that an AI-based classification system would allow them to look at the cells of the same class altogether; one interviewee claimed, *"This experience would be very nice... because seeing all these cells together... is something we don't see"*.

Trust and control. All interviewees agreed they could eventually trust an AI-based classification system, provided they could exert their own expertise for the final decision. The main reason is that physicians would somehow control the results provided by the system to ensure they are correct because they must correlate them with the patient's medical history. If they find that the classification made by the system connects well with the other evidence, the diagnosis can be confidently made. If, on the other hand, the physician has doubts, they may end up reading the slide; an interviewee claimed: *"Nobody prevents you from going to see the slide anyway"*. Lastly, the fact that a system visualizes a classification that the physician can assess is regarded as a possibility for a double check; as an interviewee highlighted, *"The safer you are, the better"*. The same interviewee pointed out, *"We are not talking about a cancer slide"*, so even if the system misclassifies a few cells, it would not be severe, *"We are not endangering people's lives"*. Even if the physician trusts the system classification, they will not rely entirely on it. An interviewee said, "*Artificial Intelligence could change our work, but we can hardly be replaced. The clinical aspect should always remain in our hands. Otherwise, it becomes a bare laboratory examination*". The physicians' clinical reasoning, based on their experience and knowledge, is fundamental. Thus, an interviewee concluded, *"The figure of the physician remains fundamental. The rhinocytologist does not want to lose his dignity. It must be clear that the physician must confirm the classification. We trust the system, but our relationship with the cells must remain"*. The same interviewee, however, emphasizes that *"The system is mechanical, so I don't know to what extent it could match my past. It could only help me with the cell count, but I must do the correlation! I would trust the cell count because I think the system must be taught by me!"*

Understandability. All physicians wish to understand the reasons behind a system classification. They believed the explanations could be the basis for a dialogue with the system; an interviewee claimed, *"I would like to dialogue with the system that would become a real interlocutor with which to carry out the cell classification process"*. And continued, *"I thought the system should have classified a cell as eosinophilic: since that did not happen, I discarded its decision. In this case, the system should tell me why it classified it that way and then ask why I discarded it"*. The system would become an interlocutor, a mechanical collaborator, and must remain so. One interviewee added that system explanations could be helpful when the physician is a neophyte and is approaching rhinology for the first time. In fact, thanks to the dialogue with the system, the neophyte can learn how to classify the cells of the nasal mucosa.

6. Human-centered redesign of Rhino-Cyt

By combining the conceptual framework illustrated in Section 4 with the insights gained through the initial interviews with the rhinocytology experts, we identified a set of requirements to redesign the user interface for the Rhino-Cyt system. As highlighted in Fig. 5, the new user interface comprises five distinct areas that serve different tasks in the negotiation process, allowing the users to scrutinize the AI model outcomes, also based on explanations, and enact model reconfigurations when they identify wrong model behaviors.

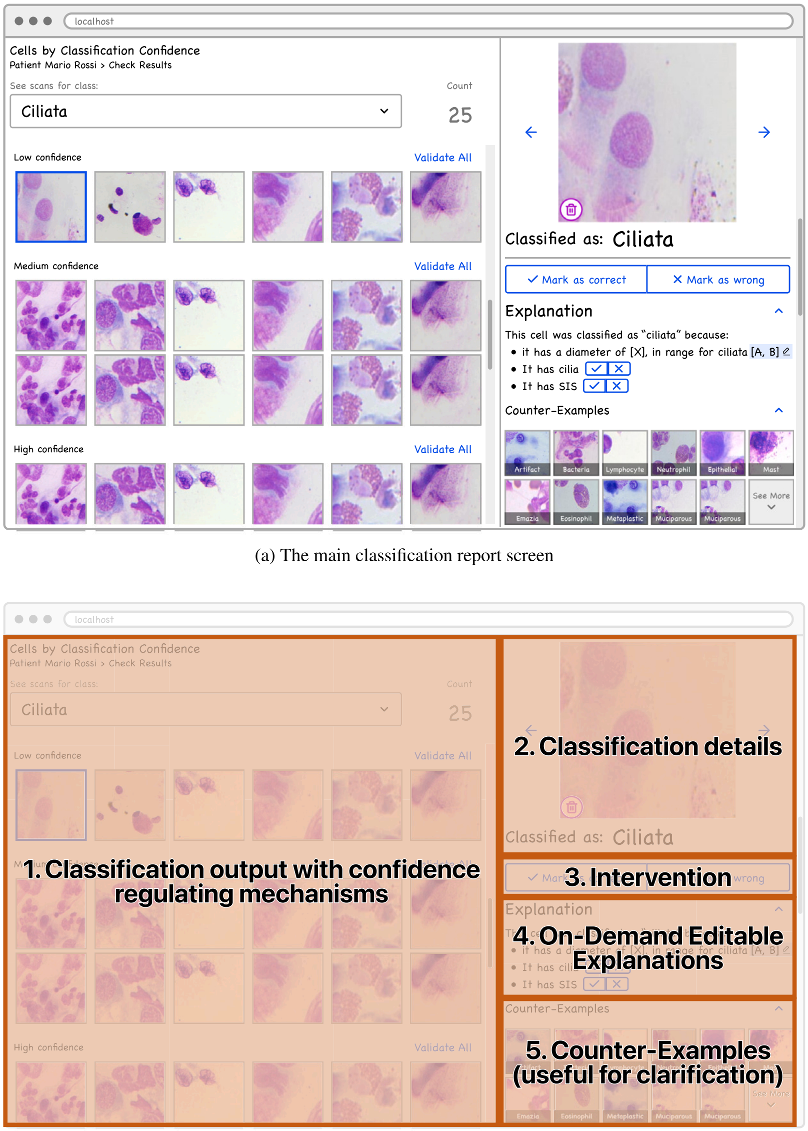
The new Rhino-Cyt prototype has been implemented as a web-based application integrating all the required functionalities. Nevertheless, for the sake of experimentation, the outcome of the analysis has been purposely crafted to test the different cases needed for the evaluation.

6.1. Displaying the classification output

The classification result is the main output of the Rhino-Cyt system; according to the insights gathered from the interviews (Theme 2.1), this is the result in which the users are most interested. As discussed with the rhinocytologists, it is shown on a class-by-class basis in an overview screen that shows the high-level results (Fig. 6). Physicians can then select a class to examine the results more in detail. Rhino-Cyt shows all instances classified in the selected cytotype, and the users can use a dropdown menu to change the visualized class (Fig. 5(a)). The instances are grouped by three levels of classification confidence, computed by the AI model as *low*, *medium,* and *high*. In accordance with the Theme "Trust and control" (Theme 2.2 in Section 5), an additional "verified" group is created if the user manually confirms a classification: physicians clearly highlighted their need to confirm the system classification. In addition, the confidence-based grouping shows the system's capabilities, adhering to guidelines for Human–AI interaction that suggest helping users understand how often the system could commit errors [48,49]. This also aligns with the need for understandability the physicians highlighted during the interviews (Theme 2.3). A sidebar allows the user to inspect the properties of a single instance by clicking on its image. The topmost area of the sidebar then shows an enlarged image of the selected instance and details on the classification results.

6.1.1. Allowing interventions and reconfiguration

To accommodate the need for control and trust highlighted by the interviewees (Theme 2.2), a specific area of the user interface empowers the users to intervene in the AI model to take control of its decisions: dedicated UI controls available in this area allow the users to mark cell classification as correct or wrong and these actions constitute feedback for the system to learn how to modify its behavior. This feature holds particular significance when addressing exceptional cases, outliers, or instances that the AI model may not have accurately identified [4]. By incorporating this feature, the Rhino-Cyt interface also aligns with best practices in the HCAI field, addressing the efficient auditing and editing by the users of an AI model outcome [48,49].



localhost

Cells by Classification Confidence

Patient Mario Rossi > Check Results

See scans for class: Ciliata

Count: 25

Classified as: Ciliata

Explanation

This cell was classified as "ciliata" because:

- it has a diameter of [X], in range for ciliata [A, B]
- It has cilia ✓ ✗
- It has SIS ✓ ✗

(a) The main classification report screen

1. Classification output with confidence regulating mechanisms
2. Classification details
3. Intervention
4. On-Demand Editable Explanations
5. Counter-Examples (useful for clarification)

(b) The five areas of the classification report screen

Fig. 5. The Rhino-Cyt redesigned interface.

6.1.2. *Providing adaptable clarifications*

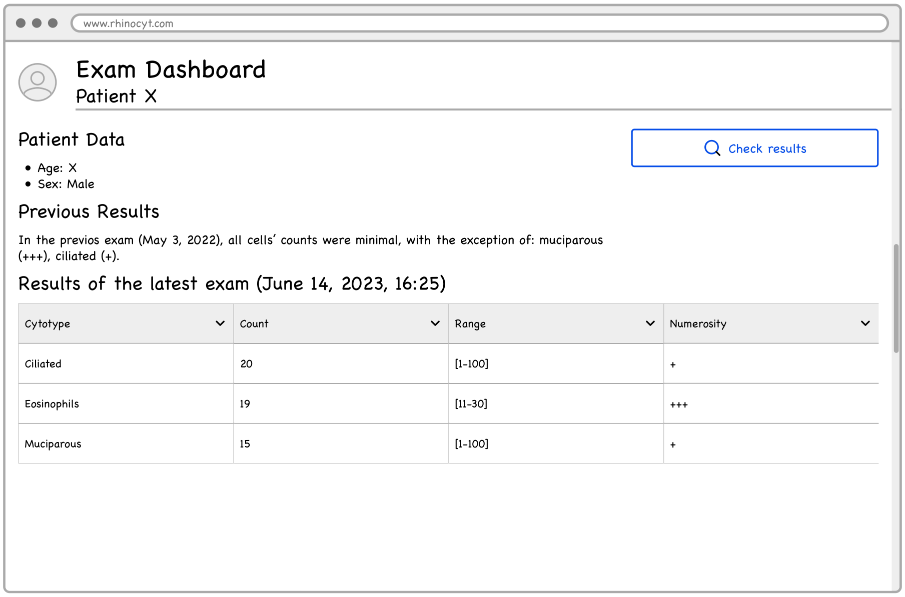
The final two areas in the lower-right corner of the interface are dedicated to explanations. Among different types of explanation (e.g., textual, visual, etc. [50]), a textual explanation is provided, designed in accordance with social science guidelines that emphasize the importance of providing the most distinctive characteristics for each class [51]. The textual explanations can be edited on-demand to let the users indicate wrong behaviors and possible corrections. Depending on the nature of the features highlighted in the explanation, users can take actions such as marking the explanation as accurate or wrong or adjusting feature values used in the explanation (e.g., if the system recognizes a cell as red, while the user perceives it as purple), thus allowing the user to align the model with their perspective.

To enforce the understandability of the model decision-making process, which in turn can favor users' trust [52,53], the last area of the interface shows additional counter-examples that enable the users to compare the selected instance with instances from other classes [50].

To provide on-demand access to information while maintaining the interface minimal [54], both clarification areas are collapsible, ensuring that users can request additional information when needed without being overwhelmed by data that may be deemed irrelevant or excessive for the task at hand. This responds to the aspects highlighted by the code "Support" of Theme 2, identified in the initial interviews.

6.2. *AI-model architecture*

Besides adopting an adequate interaction paradigm, the strategies discussed in Section 4 require adequate AI model architectures. To avoid typical drawbacks of post-hoc feature-based explanations [29], we propose the "gray-box" architecture depicted in Fig. 7 to create an AI model that can provide explanations while still utilizing images as input data set. The architecture comprises two distinct blocks: the first focuses on feature extraction from images, employing feature-specific black-box models (e.g., CNNs in the current Rhino-Cyt prototype). The rationale is that explanations do not need to detail why a particular qualitative feature was obtained, as this is likely to be of minimal interest to end users [51]. The second block handles the final classification by employing a decision tree that generates explanations starting from the features extracted by the previous block, thus enabling *clarifications*. A pivotal role in the classification is played by the model calibration, which is carried out to output a classification with three levels of confidence: low, medium, and high. This calibration activity



Exam Dashboard
Patient X

Patient Data

- Age: X
- Sex: Male

Previous Results

In the previos exam (May 3, 2022), all cells' counts were minimal, with the exception of: muciparous (+++), ciliated (+).

Results of the latest exam (June 14, 2023, 16:25)

Cytotype	Count	Range	Numerosity
Ciliated	20	[1-100]	+
Eosinophils	19	[11-30]	+++
Muciparous	15	[1-100]	+

Fig. 6. Rhino-Cyt's redesigned overview screen.

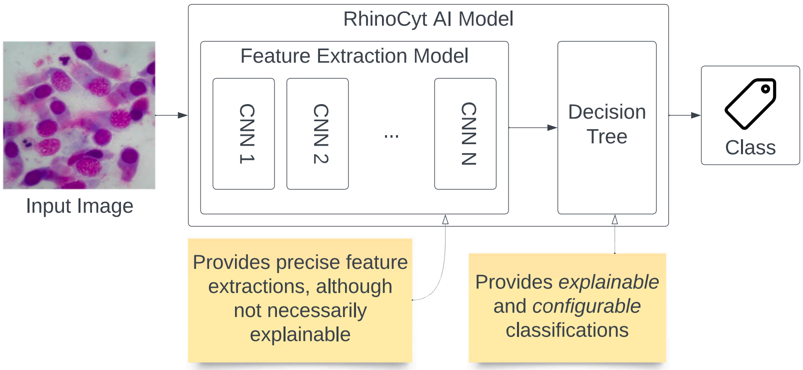

RhinoCyt AI Model
Feature Extraction Model
CNN 1
CNN 2
...
CNN N
Decision Tree
Class
Input Image
Provides precise feature extractions, although not necessarily explainable
Provides *explainable* and *configurable* classifications

Fig. 7. The proposed AI-model architecture.

could be enhanced by the adoption of specific metrics, such as the *ECI* (Estimated Calibration Index) provided in the evaluation framework described in [55].

Given its white-box nature, this model is configurable, allowing end users to edit the features and the parameters used in classification, thus enabling *reconfiguration*. In particular, the reconfiguration can be operated in different ways: (i) *automatically at the feature-level*: when the user marks an explanation as right or wrong, the instance is added to the dataset to train the feature-specific model that is responsible for the explanation; (ii) *automatically at instance level*: when the user marks a classification as right or changes the class, the instance is added to the dataset to train the final classification model; (iii) *manually at feature-level*: the system may expose parameters of the final classification model, allowing the user to disable the usage of certain features that may have been wrongly learned in the training process.

7. Evaluation study

An exploratory study was conducted to assess to which extent the redesigned interaction paradigm, based on the three strategies illustrated in Section 4, improved the users' understanding and control in the interaction with the AI system. The study was structured as a within-subject usability test with a thinking-aloud protocol. The details of the study are reported in the following.

7.1. Participants

We employed convenience sampling to recruit a diverse group of 9 physicians (4 F, 5 M), from different Italian hospitals. Their demographic details are reported in Table B.6. The participants have been recruited through personal contacts. These participants varied in seniority and expertise within the field of rhinocytology. Participants' ages ranged from 30 to 70 years, with a mean age of 53.44 years (SD = 13.74). Their professional experience varied from 10 to 20 years, with a mean of 14.89 years (SD = 5.78), except for one participant with only 2 years of experience. Before their participation, all physicians provided explicit informed consent for their involvement in the study. None of the participants had previous experience with the Rhino-Cyt system. None of the interviewees recruited for the initial investigation (Section 5) took part in this second study.

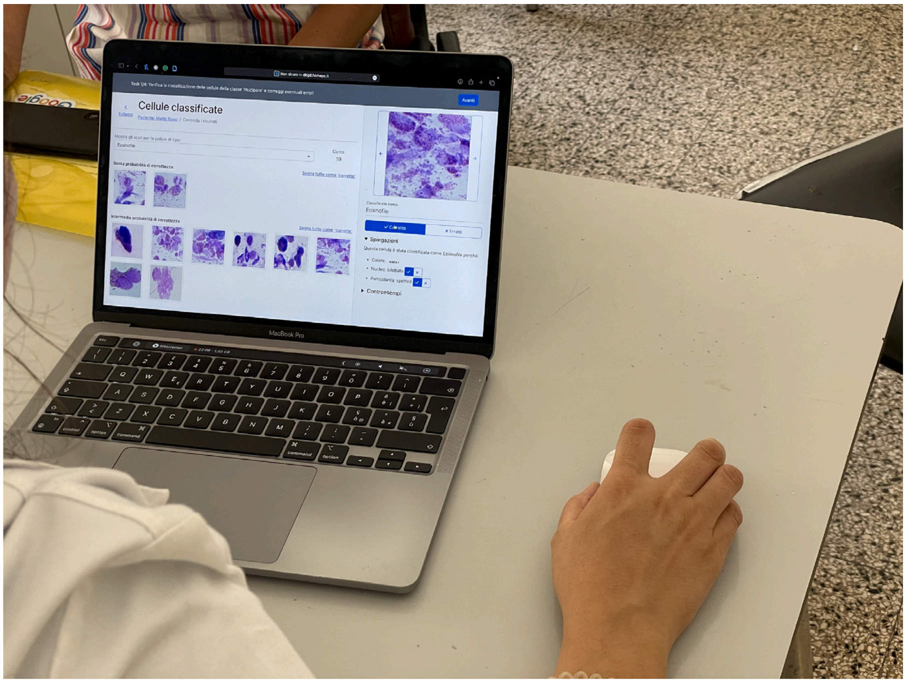


Fig. 8. A photo of a user during the user test, showing Rhino-Cyt's redesigned layout.

7.2. Methods

Rhino-Cyt prototypes. We utilized the two versions of the Rhino-Cyt system: the initial Rhino-Cyt (referred to as "original") and the redesigned prototype introduced in Section 6 (referred to as "redesigned"). Both prototypes were implemented as Web applications. By default, in the "redesigned" interface, explanations were visible while counter-examples were collapsed. To maintain control over the dataset shown in the prototype, we introduced incorrect instances of cells to evaluate the usefulness of features such as explanation and reconfiguration in case of both correct and wrong classified cells. Efforts were made to optimize UI component reuse, to ensure a consistent visual experience between the two prototypes, minimizing potential biases from purely aesthetic differences. Fig. 8 shows a participant during the user test, showing the default layout of the prototype.

Platform for remote test. A custom platform developed using Node.js managed the study. Once opened, participants were presented with a consent form, and upon acceptance, one of the prototypes was opened. Participants completed the tasks, and the platform administered the questionnaires. The platform then displayed the second prototype, with the order of administration counterbalanced across participants.

Outcome measures. The study focused on assessing participants' workload, acceptability, trustworthiness, User eXperience (UX), as well as the impact of the system on human and professional values.

To assess the workload, we used the *NASA Task Load Index (NASA-TLX)* questionnaire, a well-established tool for assessing subjective workload [56]. It includes questions and rating scales that assess mental, physical, and temporal demands, performance, effort, and frustration and provides insight into user-perceived workload.

To evaluate the UX, we performed a content analysis of qualitative data obtained from answers to the open questions and from the recorded video footage of participants interacting with the prototypes. Content analysis was used in this study because it provides valuable information about user engagement, task performance, and user satisfaction [57] and can help identify patterns and trends in users' behaviors.

Regarding acceptability, the *Unified Theory of Acceptance and Use of Technology (UTAUT)* was used to measure performance expectancy, effort, expectancy, attitude toward using technology, behavioral intention to use the system, self-efficacy, and social influence [58].

Trust in the Rhino-Cyt system was measured using the *Trust in Automated Systems Test (TOAST)* questionnaire [59]. This tool assesses trust in different aspects of automated systems, such as reliability, transparency, and overall trustworthiness.

Finally, we also engaged physicians in evaluating how design choices might impact human and professional values within the context of cytological analysis. This assessment followed the principles of *Value-Sensitive Design* [60], which defines ethical acceptability in terms of how much the technology supports their personal (or, in this case, professional) values. Four values were selected as relevant for physicians:

- *Autonomy*: the degree to which physicians using the tool can plan and act in a way that supports their goals.
- *Accountability*: the possibility of uniquely tracing physicians' actions and the tool's responses.
- *Freedom from Bias*: concerns about potential systematic unfairness due to technical or socio-technical biases.
- *Identity*: understanding who physicians are in terms of professional skills and competencies when using the tool.

7.3. Procedure

This section details the procedure adopted for the user study. The overall procedure is shown in Fig. 9. Twenty potential participants from a group of rhinocytologists were invited to join the study through personal emails, SMSs, and WhatsApp messages. Nine physicians expressed their availability to participate in the study. Participants were scheduled for 90-min appointments over the following two weeks. Due to their locations across Italy, a remote user test was organized using Microsoft Teams (with the exception of one participant, that participated in presence). Participants received a link to the MS Teams room and the study platform via email.

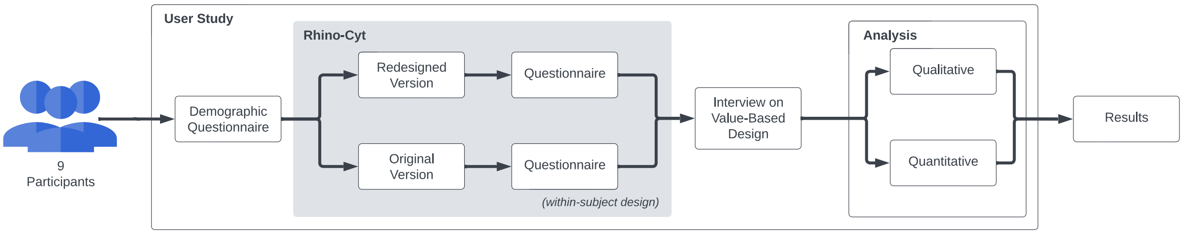


Fig. 9. Overall process of the user study.

On clicking the MS Teams room link, the participant found two researchers (a conductor and an observer) waiting. After welcoming and thanking the participants, the conductor instructed them to open the link to the study platform. A consent form was presented, and participants were asked to sign it digitally. All participants gave their consent. Participants then provided their gender, age, and email address. Being a within-subject study, the participants experience both conditions but the order was randomized.

In each of the two conditions, participants were presented with one of the prototypes and asked to complete four tasks. These tasks were designed to observe user behavior and preferences concerning the AI model's clarification and intervention features. The tasks were as follows (presented in random order):

1. Check the classification of class "muciparous" cells and correct errors, if any.
2. Check the classification of class "eosinophils" cells and correct errors, if any.
3. Identify misclassified cells in class "neutrophils" and correct errors, if any.
4. Identify cells misclassified in class "ciliated" and correct misclassifications and explanations, if any.

After completing the tasks, participants filled out the NASA-TLX, UTAUT and TOAST questionnaires. They were then given a 5-min break. After the break, participants repeated the entire procedure with the second experimental condition.

A semi-structured individual interview was conducted after participants completed tasks with both prototypes. Specifically, participants were asked to share general comments on the two prototypes and to discuss the value-based design dimensions. Then, they were introduced to various values related to their role as professional physicians and were asked to rank the impact of the system (Rhino-Cyt) on these values. Additionally, they discussed any differences they observed between the two interfaces regarding these values and were invited to suggest other values for consideration.

7.4. Quantitative data analysis and results

In the following subsections, the study results along the outcome measures, i.e., workload, acceptability, and trustworthiness, are reported. Wilcoxon Signed Rank test was computed to analyze the results of the NASA-TLX, UTAUT, and TOAST because of the violation of normal distribution (assessed with the Shapiro–Wilk test). An alpha level of 0.05 was used for all statistical tests. The rank-biserial correlation (r) was calculated as a measure of the effect size of the difference between experimental conditions; this test is suited for non-normally distributed interval data, as the one of our study (normality assessed with Shapiro–Wilk test). A value of $r = 0$ implies the absence of a relationship. Values of r below ±0.29 are considered indicative of a weak correlation, between ±0.30 and ±0.49 indicate a moderate correlation, and finally, values between ±0.50 and ±1 suggest a strong correlation [61].

7.4.1. Workload

The analysis of the workload perceived by participants while using the two prototypes resulted in very similar low, thus, positive values (Original M = 25, SD = 12.13; Redesigned M = 23.7, SD = 9.12). The result of the Wilcoxon Signed-Rank test was not statistically significant. Also the observed effect size r is very small, 0.06.

To gain more insights from this analysis, the six subscales of the NASA-TLX, i.e., *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration* (each scale ranges from 0 = low to 100 = high), were analyzed separately. The details for each subscale, as well as the results of the Wilcoxon Signed Rank tests, are reported in Table 2. Also, in these cases, no significant differences emerged and all the observed effect sizes are r small.

7.4.2. Acceptability

The analysis of the acceptability perceived by participants while using the two prototypes resulted in very similar high scores (Original M = 3.87, SD = 0.7; Redesigned M = 3.98, SD = 0.89). The result of the Wilcoxon Signed Rank test was not statistically significant. Also the observed effect size r is small in all cases, except for physical demand.

To gain more insights from this analysis, the six subdimensions of the UTAUT, i.e., *Performance Expectancy*, *Effort Expectancy*, *Attitude toward Using Technology*, *Behavioral Intention to Use the System*, *Self-Efficacy*, and *Social Influence*, were analyzed separately. The details for each subscale, as well as the results of the Wilcoxon Signed Rank tests, are reported in Table 3. Also, in these cases, no significant differences emerged and all the observed effect sizes are r small in all cases.

7.4.3. Trust

The analysis of the trust perceived by participants while using the two prototypes resulted in very similar high scores of the TOAST index (M = 5.7, SD = 1.5; Redesigned M = 5.7, SD = 1.5). The result of the Wilcoxon Signed Rank test was not statistically significant. Also, the observed effect size r is very small in all cases.

To gain more insights from this analysis, we analyzed the TOAST subdimensions, i.e., understanding and performance. The details for each subscale, as well as the results of the Wilcoxon Signed Rank tests, are reported in Table 4. Also in these cases, no significant differences emerged. Also, the observed effect sizes r are very small.

7.5. Qualitative data analysis and results

In this subsection, we report the results of the analysis of the qualitative data, i.e., the observers' notes and the participants' oral and written comments, and finally the results of the value-based evaluation.

7.5.1. Content analysis

Various insights were extracted from the video analysis of the participants during the interaction with both versions of Rhino-Cyt. Generally, participants approached the system addressing it as a tool for visualizing the collected and classified cells, rather than as a tool that could automatize the diagnostic process: this highlights that Rhino-Cyt was able to correctly set the expectations and trust level.

By examining the participants' interactions with both system versions, we discovered interesting patterns of how physicians approach an AI-enabled diagnostic tool, which we discuss in the following.

Table 2
Results of the NASA-TLX questionnaire. For the means, 95% confidence intervals are provided.

	NASA-TLX		Mental Demand		Physical Demand		Temporal Demand		Performance		Effort		Frustration	
	x̄	Mdn	x̄	Mdn	x̄	Mdn	x̄	Mdn	x̄	Mdn	x̄	Mdn	x̄	Mdn
Classic	25 ± 9.33	21.67	36.67 ± 13.31	30	24.44 ± 14.44	20	28.89 ± 13.56	20	15.56 ± 19.26	10	26.67 ± 13.86	20	17.78 ± 10.01	10
Redesigned	23.7 ± 7.01	23.33	33.33 ± 14.89	30	18.89 ± 7.13	20	25.56 ± 11.6	20	14.44 ± 19.64	10	34.44 ± 19.26	30	15.56 ± 6.78	10
w-Test	$Z = -0.2$ $p = .859$ $r = -0.06$		$Z = -0.5$ $p = .608$ $r = -0.2$		$Z = -0.9$ $p = .395$ $r = -0.3$		$Z = -0.09$ $p = .931$ $r = -0.03$		$Z = -0.3$ $p = .766$ $r = -0.1$		$Z = 0.3$ $p = .792$ $r = 0.1$		$Z = 0$ $p = 1.000$ $r = 0$	

Table 3
Results of the UTAUT questionnaire. For the means, 95% confidence intervals are provided.

	UTAUT		Performance Expectancy		Effort Expectancy		Attitude Toward Using Technology		Behavioral Intention to Use the System		Self-Efficacy		Social Influence	
	x̄	Mdn	x̄	Mdn	x̄	Mdn	x̄	Mdn	x̄	Mdn	x̄	Mdn	x̄	Mdn
Classic	3.87 ± 0.54	4	3.72 ± 0.82	4	4.53 ± 0.33	4.5	4.17 ± 0.71	4.5	3.75 ± 0.96	4.25	4.39 ± 0.52	4.75	2.5 ± 0.86	2.5
Redesigned	3.98 ± 0.69	4.17	3.97 ± 0.98	4.5	4.64 ± 0.32	4.75	4.11 ± 1	4.75	3.86 ± 1.11	4.25	4.42 ± 0.43	4.5	2.89 ± 0.79	3
w-Test	$Z = 0.4$ $p = .676$ $r = 0.1$		$Z = 0.7$ $p = .514$ $r = 0.2$		$Z = 0.3$ $p = .752$ $r = 0.1$		$Z = 0.7$ $p = .512$ $r = 0.2$		$Z = 0.2$ $p = .833$ $r = 0.07$		$Z = 0.2$ $p = .833$ $r = 0.09$		$Z = 0.2$ $p = .833$ $r = 0.09$	

Table 4
Results of the TOAST questionnaire. For the means, 95% confidence intervals are provided.

	TOAST		Reliability		Transparency	
	x̄	Mdn	x̄	Mdn	x̄	Mdn
Classic	5.7 ± 0.66	5.88	5.97 ± 0.35	6	5.42 ± 1.14	6
Redesigned	5.75 ± 0.82	6	5.94 ± 0.63	6	5.56 ± 1.11	6
W-test	$Z = 13$ $p = .547$ $r = 0.2$		$Z = -0.07$ $p = .944$ $r = -0.02$		$Z = 0.2$ $p = .866$ $r = 0.06$	

Pattern 1: Trade-off between experience and use of explanations. A striking behavior observed in most participants concerns the use of explanations. In fact, explanations were almost always used only by the less-experienced physicians, while the experienced ones felt they did not need to be helped by an AI-based system. For example, participant P3, an experienced physician, said: *"I don't need the system to tell me why a cell is classified in a certain class. I can understand whether a cell is correctly classified by looking at the picture rather than reading the explanation. Therefore, I would only use the system because it speeds up the classification process and the number of cells in each class, which helps me to make a diagnosis"*. This comment could be useful to explain the high workload of this participant (NASA-TLX score = 32.5/100), mainly caused by the Effort sub-dimension; indeed, this sub-dimension concerns how difficult it was to achieve the goal and the presence of classification and explanations could be an obstacle to diagnosis if the physician simply wants a system that counts the cells in each category. On the contrary, less-experienced physicians found the explanations very useful to support their decisions, especially in cases where they were undecided. For example, the participant, who has only 2 years of experience in this field, said: *"I find the explanations very useful because, in some cases, I could not remember the details of a class of cells and the explanations helped me to remember and decide on the correct classification."*. This facilitation offered by the explanations to less-experienced participants also emerges clearly as a positive aspect in the participant's questionnaire results. Indeed, she achieved one of the lowest workloads (NASA-TLX score = 20/100), the highest acceptability value among the participants (UTAUT score = 4.9/100), and the highest trust value (TOAST score = 4.9).

Pattern 2: Lack of visual relationship between the explanation and its target. A total of 3 users expressed difficulty linking the explanation to the cells reported in each picture, especially when the pictures presented several cells inside. Indeed, it was unclear to them which cell the explanation referred to, which was more complicated in the presence of similar cells. Participant P6, for example, said: *"I can't understand if this explanation refers to one or both cells in the image"*. This comment may explain the low acceptability of the system on his part (UTAUT score = 3.2/10), since an ineffective explanation may be perceived as causing a low quality of the whole system, which in turn results in a low users' inclination to adopt the system in real contexts. Another participant, P7, also said: *"It should be visually evident which cell the explanation refers to"*. For this participant, we also observed a particularly high workload (NASA-TLX score = 35.8/100) with the redesigned system, caused mainly by the dimensions of Mental demand, Temporal demand and Effort. The high values of these 3 dimensions may explain the user's effort to relate the explanations to the cells shown in the images. An improvement in this aspect could probably positively impact the workload of the entire system. This brings to mind the possibility of using heat maps on images to highlight the connection between the explanation and the cell involved.

Pattern 3: Experts are not omniscient — they cannot always correct the system. Both systems made it possible to reclassify the cells deemed to be wrong, albeit with different interaction mechanisms. Regardless of the underlying interaction mechanism provided by the system, one comment from participant P7 concerns the lack of complete knowledge of an expert, who may know that a cell does not belong to a class but may not know which class it belongs to. This implies the need to include a new class to collect cells that need to be reclassified but for which the physician cannot specify the correct class. In addition to having an implication on the interaction and the UI, this comment also implies a re-training of the model that differs from what would be done by explicitly tagging an instance with another class: in this case, the probability for a certain instance to belong a certain class must only be lowered, and nothing can be deduced about the remaining classes.

Pattern 4: Cells classification features must be chosen also by experts, not only by AI models. Both systems used by the participants started from the actual training of a CNN-based model for the recognition and classification of cells within slide images. However, the comments of 2 participants (P3 and P6) revealed that the visual features selected by the model to classify cells do not consider the real needs of physicians, who instead need explanations of features not included in the classification process. For example, participant P3 said: "*Other features that are much more indicative for human classification would be useful in the explanation, e.g. nuclear reinforcement useful for identifying muciparous calciforms*". Similarly, participant P6 said that "*For eosinophils there would be Degranulation, a typical feature of an allergic phase*". These comments suggest an a priori selection of a set of cell characteristics that takes into account the needs of the end user in the diagnosis process, otherwise explanations may be useless.

Pattern 5: The correction of explanations is often neglected. The redesigned version of the system offered the possibility of editing the explanation on demand so that the model could be fine-tuned. However, in very few cases the participants used this function of the system. It is unclear whether this is due to a secondary role of this functionality in the UI or whether the functionality itself was considered to be of little use. In some cases, it appeared that users had difficulty using this function because they were confused by the explanation, i.e., not understanding which cell it referred to made it impossible for them to correct the explanation. In other cases, on the other hand, it seemed unnecessary for some users to correct the explanations because they were considered to be of little relevance to them, particularly the more experienced ones.

7.5.2. Value-sensitive assessment

As part of the debriefing interview, we ask participants to rank the importance of four values that may pertain to this type of technologies. Of the nine participants, one did not agree to participate in this part of the evaluation. Table 5 reports the relative ranking of the other physicians.

Table 5
Relative ranking of values as reported by the physicians. Each cell reports the number of participants that ranked each value as of a particular relevance.

	Very relevant	Relevant	Relatively relevant	Not much relevant
Autonomy	3	4	0	1
Freedom from bias	3	3	2	0
Accountability	2	0	4	2
Identity	0	1	2	5

The values of autonomy and freedom from bias are those that seem to be more at risk of being impacted by the system: the impact of these values is regarded as either "very relevant" or "relevant" by two participants. The physicians reporting those values acknowledged potential risks that the system's suggestions are taken for granted and induce errors (that is, an impact on "freedom from bias"). Two of them explicitly mentioned a possible risk to autonomy because the tool may prevent or somehow discourage a deep analysis of the cells not explicitly selected. On the other hand, one physician claimed that autonomy is not a relevant value because collaboration is fundamental in the medical domain, and that requires reducing autonomy (it is worth noting that the participant was not referring specifically to the use of Rhino-Cyt but to the medical professional in general).

The risk of accountability raised more disagreement among the participants: while four physicians dismissed it as relatively relevant, other two found it quite a concern. For one of the two, the emphasis is on the negative side: this kind of system can induce errors, impacting the value of accountability (that is, whose fault is it if the physicians accept the wrong suggestion?). The other participant stressed the other aspect: the physician should always be held responsible, and human decisions should easily override this system. This stance confirms the adequateness of including reconfiguration mechanisms in a paradigm for Human–AI interaction in the medical domain.

Finally, the risk posed to the physicians' professional identity seems overall not perceived as relevant. The only physician who ranked this risk as "relevant" did not comment on this decision but claimed that the other values are even less impacted. We can assume that all these professionals are used to employing digital interfaces in their work, and both the Rhino-Cyt interfaces might look like standard tools.

All the physicians are positive about using this type of tool overall. As emerged during the initial interviews, they commented that cytological analysis is just a step in the diagnostic process, and possible errors cannot have dramatic consequences. One of them actually recognized that using such tools might be beneficial not only for the sake of efficiency but also as a way of fostering more objectivity in this kind of analysis.

Only one of the physicians had an explicit preference for the interface without explanation because it was regarded as more efficient: they claimed that they do not have time to delve into long digressions about the system's motivation for a suggestion, yet they also acknowledged that, in case the only task of the specialist is cytological analysis, such explanations might be helpful.

Two of the eight participants strongly preferred the new interface because, in case of misalignment between the user's and the system's classification, the explanation may help understand the system's interpretation and clarify the issue. On the other hand, the other two participants preferred the old interface because they claimed the physician should not be bothered by the reasons for the system's different interpretation (that is, the physician is always correct). It might be interesting to note that the two former participants are the younger (and less experienced), while the latter are among the most experienced. This aspect may highlight a potential educational or scaffolding role for AI-based tools, as already emerged from the initial interviews.

8. Discussions

In HCI and AI, the integration of computer-aided diagnosis tools for physicians has been a subject of significant research and experimentation (see, for example, [62–64]). In this work, we tried to enlarge the perspective from the usability and the acceptance of a specific tool to the wider proposal of a conceptual framework, together with its application and evaluation to a specific tool. This section discusses the lessons learned toward future adaptations and application of this framework.

8.1. Lessons learned

Lesson 1—Expertise-driven insights: Tailoring AI clarifications for optimal user engagement. Our conceptual framework emphasizes the importance of clarification in promoting user understanding and control over AI systems. The study reaffirms this, revealing a pattern in user engagement based on expertise levels, and this divergence in preferences aligns with the framework's call for meaningful and accessible explanations. As reported in Pattern 1 (*Trade-off between experience and use of explanations*), a significant portion of users opted not to utilize clarification. Specifically, experienced physicians emerged as a group showing a high confidence level in the system's decisions, thanks to their knowledge, without additional explanations. In contrast, the physician with less experience (only 2 years) exhibited a perceived benefit from clarification, finding it supportive in their decision-making process. This divergence aligns with established principles in user interface design, which promote flexibility and adaptation to give control and freedom to the users so that they can feel that they are in control of the system themselves — see, for example, the notion of *Internal Locus of Control* promoted by the Shneiderman's golden rules for UI design [65]. Existing research, such as [66], substantiates the notion that expert users, like senior physicians, often rely on robust mental models and pre-existing knowledge. Thus, they require less explicit information. On the other hand, novice users, akin to younger physicians, tend to derive substantial benefits from detailed clarifications [67]. This phenomenon also aligns with the principles of cognitive load [68] and the expertise reversal effect [66]. Therefore, future investigations in this domain could delve into identifying specific expertise levels at which explanations become more or less beneficial. This exploration may also lead to adaptive interfaces [69] capable of tailoring the provision of explanations to align precisely with the user's expertise, thereby optimizing user experience and system trust. This lesson further confirms the importance of the negotiation process outlined in the conceptual framework (Section 4), with an emphasis on the system's capability to understand when and how users seek clarification and contribute to the iterative cycle of clarification and reconfiguration.

Lesson 2—Customizing explanations: User-centered design for enhanced cognitive resonance. Our investigation into the utilization of explanations uncovered a pivotal user behavior: a preference for explanations tailored to features that are perceived as distinctive for the user and, thus, useful. This aspect clearly emerged from the observation of the user interaction, as described in Pattern 4 (*Cells classification features must also be chosen by experts, not only by AI models*). The emphasis on "distinctive and odd" features in users' preferences for explanations also reflects the well-established concept of saliency in visual perception, i.e., users tend to prioritize information that stands out [70]. In addition to the system-driven personalization discussed above, this lesson, therefore, emphasizes the necessity of enabling user-driven customization, which can be highly favored by an active negotiation process during runtime interaction with the AI system. By gathering and incorporating user feedback and preferences dynamically, explanations can be tuned in real time to align with individual cognitive styles and actual needs, optimizing both user engagement and understanding [71,72]. The process of negotiation is essential in this context and requires mechanisms enabling a continuous and interactive exchange between the user and the AI system.

Users' preference for explanations tailored to distinctive features of the classified entity (the medical image in our study) aligns with the framework's call for strategies that move away from the "black-box" model. The innovative aspect is the opportunity given to the users to negotiate with the system what the explanations should focus on. This lesson also reinforces the importance of involving users in the design process, as suggested by the conceptual framework, to identify, already at design time, possible dimensions along which to tailor explanations, coupled with interaction mechanisms to let the users express their preferences.

Lesson 3—Unveiling user needs: The unspoken desire for intervention in AI-assisted systems. The study brought forth a noteworthy aspect: although participants did not explicitly express the need for intervention functionality in the system, the usage of the two systems showed that this is a desirable requirement. This was driven by users utilizing functions in both system versions to correct erroneously classified cells. Notably, the simplicity of the reconfiguration process played a pivotal role: while the original version allowed for reconfiguration with a single click, the redesigned version required two steps, leading to a notable drop in the user completion rates. The implicit user inclination toward intervention and reclassification highlights the importance of adequate intervention mechanisms; this aligns with the concept of user agency and control in interactive systems to give the users the capability to intervene and modify system outputs as needed [73]. This result also highlights the importance of devising intuitive and high-efficiency interaction paradigms that are able to compete with the immediateness of the AI systems' autonomous behaviors. Concerning our conceptual framework, the user inclination toward intervention functionality confirms the importance of mechanisms allowing users to initiate clarification and reconfiguration sessions. This, in turn, highlights the need for interactive negotiation strategies, enabled by a dialog for mutual comprehension between the user and the AI system.

Lesson 4—Focusing attention: Enhancing medical imaging interpretation with localized explanations. A notable user behavior observed in interactions with the redesigned system highlights the importance of localized contextual explanations (see Pattern 2, *Lack of visual relationship between the explanation and its target*). Participants expressed a preference for explanations to be specific to the part of the image under model classification. Some participants articulated the need for a visual cue, such as highlighted cells, to accompany explanations. This would assist in associating the explanation with the addressed image features, preventing ambiguity, especially in scenarios where the presence of multiple cells, makes it challenging to discern which cell the explanation pertains to. In the domain of medical imaging and diagnosis, localizing explanations to specific regions of interest aligns with the principles of attention and visual perception [74]. This approach ensures that users can direct their focus to relevant information within complex visual stimuli, a situation occurring for medical images with multiple cells or structures.

The emphasis on localized contextual explanations also supports the negotiation process outlined in our conceptual framework: providing explanations specific to the part of the image under classification can contribute to improving user comprehension within the negotiation cycle, as it improves the system's ability to present outcomes in perceivable and interpretable ways. It is also in line with recent results in XAI that highlight the significance of providing contextual explanations that are pertinent to the user's ongoing task and environment [26,27].

Lesson 5—Empowering trust: The role of transparent learning in human-centered AI systems. The study highlighted how not all users utilized the functionality for reconfiguring the model. On the contrary, users expressed appreciation for the model's transparent learning from its mistakes instead of explicitly correcting the model, as reported for Pattern 5 (*The correction of explanations is often neglected*). The model's capacity to learn from its errors reinforces the feedback-driven adaptation concept, and favors active user participation in the learning process. This lesson, therefore, emphasizes the role of the negotiation process, where users systematically assess the AI model's outputs and take intervention steps. It also highlights the importance of transparency, user control, and robust feedback mechanisms, especially in sensitive domains like medical diagnosis. By gathering and accommodating user preferences for transparent learning over explicit reconfiguration, AI systems can improve trust and offer a sense of user agency, thus contributing to the evolution and refinement of these systems within critical domains.

Lesson 6—Considering experts' knowledge limitations in the design process. In the overall design process, as the one proposed in this research, it is important to recognize the limitations of expert knowledge within the context of system correction and reclassification. This was highlighted by the content analysis, and in particular emerged from Pattern 3 (*Experts are not omniscient — they cannot always correct the system*), which revealed that experts may not always possess the complete knowledge required to correct the decision of an AI-based system [75,76]. This lesson has multifaceted implications for the design of these kinds of systems. Firstly, it highlights the need for system designers to consider the limitations of expert knowledge when developing interaction mechanisms and UI elements. The comment from participant P4 highlights the potential impact on user experience and the design of system interfaces, emphasizing the importance of accommodating the expert knowledge uncertainty within the system [76]. Furthermore, it also extends to model re-training within Human-centered systems. Unlike the conventional reclassification of an instance into another class, the scenario described by participant P4 necessitates a different approach. For example, the probability of a specific instance belonging to a certain class can only be lowered, without providing information about the remaining classes. Different aspects, from interaction mechanisms to model reconfiguration techniques, need to consider the fallibility of expert knowledge [75].

Lesson 7—Preserving values: A human-centered approach to mitigating risks in AI interactions. Physicians identified autonomy and freedom from bias as values most at risk of being impacted by the AI system. The perceived risk to accountability showed mixed opinions, while their professional identity was considered less relevant. A fear of deskilling in case of long-run use clearly appears (Section 5.2.2). This aspect needs to be further investigated. More in general, this lesson reinforces the importance of understanding and addressing user values in designing ethical and value-sensitive AI interactions. Identifying and prioritizing values at potential risk is essential. Addressing concerns related to autonomy and bias should be a focal point in system design, emphasizing features that empower users and mitigate biases to ensure ethical and value-sensitive AI interactions. This lesson acknowledges the importance of the negotiation process where users progressively engage with clarifications and AI model reconfiguration, preserving their capability to intervene and give feedback on the outcome reliability to mitigate biases. Beyond allowing for interactive improvement of the model, the negotiation process might also be helpful in preventing deskilling [77–79]. Therefore, AI applications should adopt clear mechanisms to suggest and enable intervention.

Lesson 8—Efficiency and objectivity as drivers for adoption. Physicians expressed positive attitudes toward using AI tools for cytological analysis, recognizing potential benefits in terms of efficiency and objectivity of the diagnosis process. However, opinions varied on the necessity of detailed explanations, with some favoring efficiency over elaborate justifications. Highlighting the efficiency gains and objectivity enhancement brought by AI tools can be key in promoting adoption. Offering customization options for explanation depth caters to diverse preferences, ensuring that both efficiency-focused and detail-oriented users find value in the system. This lesson aligns with the framework's focus on new paradigms for interaction and clarifications that cater to diverse user preferences. Favoring efficiency gains and objectivity enhancement is in line with the negotiation process, where users are in control, assess the AI model's outputs, and seek clarifications when necessary, as well as being able to customize the explanation depth.

8.2. Discussion on the quantitative results

Although the quantitative analysis did not find statistical differences, possibly due to the small sample size, the results can still be helpful for reproducibility. In addition, the results highlight situations that are worth to be investigated in the future. However, we want to emphasize that, given the absence of significant differences, the following discussion is mostly speculative and should be considered as a guide for future research and additional longitudinal studies.

The Rhino-Cyt redesigned version generally seems to improve the original for all evaluated properties, although the effect sizes are mediocre for all dimensions, except for the physical demand in the NASA-TLX, where a moderate effect is observed.

The workload is generally lower (25.00 for the original version and 23.70 for the redesigned version). More specifically, the participants felt that fewer mental, physical, and temporal demands were requested by the redesigned version. The effort required to fulfill the task increases. However, this increased effort may aid in decreasing the risk of deskilling, promoting the redesigned version of the system as an example of *frictional AI*: friction is added deliberately to reduce the risk of deskilling [80].

Similarly, the system acceptability slightly improves (3.87 for the original version and 3.98 for the redesigned version). This result suggests that the negotiation between users and the AI in the redesigned system reduces the negative feelings toward AI. The negotiation process allows users to edit the AI output interactively, eliminating the feel of replacement that users may experience when dealing with AI-enabled systems (Section 5).

Finally, trust in the system also increases (5.69 in the original version and 5.75 in the redesigned version). More precisely, the perceived reliability slightly decreases while the perceived transparency increases. Although it may seem trivial, this suggests that introducing explanations may introduce an additional point of failure in the system. However, overall, the added transparency increases the trust in the system.

9. Conclusions

Within the field of HCAI, the work illustrated in this article aims to investigate new classes of AI-assisted systems, where the drawbacks of Black-Box approaches, and especially model biases due to the lack of domain knowledge, are overcome by means of tools that can empower domain experts to control and customize the outcome of AI algorithms. The work promotes the importance of human-centered methods as the key to better understanding the final users' domain. The applied research methodology, emphasizing users' involvement in inquiring about foundational aspects of a research challenge, is commonly used in HCI but rarely used in AI. This article emphasizes the benefits of applying human-centered methods. It also proposes and validates a Human–AI interaction paradigm based on three strategy that open new perspectives for designing AI-based systems. Overall, the goal is to overcome Black-Box approaches, so that the end users can understand the algorithmic decisions and possibly influence them with their perception and knowledge; these benefits can also contribute to building increased trust in the AI model. These are indeed among the main criticalities observed for AI tools; they need to be addressed, especially in the field of medicine where physicians' judgment still plays a central role and experts must be in control and trust the technology.

The insights gathered through the human-centered redesign of the Rhino-Cyt tool are encouraging. However, additional user studies are needed to identify and characterize the salient features for explanations that can effectively sustain rhinocytologists' tasks within the proposed conceptual framework. One interesting result would be the definition of a human-centered model for explainability, capable of capturing the specificity of the addressed domains and the specific needs of the target users for providing customized and meaningful support. Technical experiments will also compare the performance of the proposed AI-model architecture to that of Black-Box models, to investigate how the reconfigurations operated by the end users can contribute to improving the model performance. Further and extensive investigations will also assess the validity of the proposed methodological framework beyond the specific domain Rhino-Cyt refers to. In particular, we acknowledge that, for the sake of simplicity, we did not fully investigate the important case in which the user does not agree with the system when the system is actually right. This scenario, which is of course plausible, may trigger further problems if not recognized during the negotiation phase.

In rhinocytology, the application of AI to cell image classification represents a solution where the consequences of misclassification are nuanced and generally less critical than in other medical domains. Although the AI system will demonstrate high levels of accuracy in the future (a preliminary evaluation of classification models for our system has been reported in [12]), it is important to note that AI tools will never be fully accurate. In rhinocytology, misclassification of a cell type does not usually result in serious patient harm. This is mainly because rhinocytology is often used as a preliminary assessment rather than a definitive diagnostic tool. In other medical fields, such as oncology or cardiology, the margin for error is much smaller: AI misclassifications can lead to incorrect treatment decisions with potentially serious consequences, including delayed diagnosis of life-threatening conditions, administration of inappropriate treatments, or even patient death. This difference in error tolerance highlights the need for domain-specific considerations when integrating AI into medical diagnostic processes. It emphasizes the need for strong fail-safes, human oversight and the use of AI to complement human expertise rather than replace it. It also highlights the importance of AI systems being transparent and able to defend their decisions so that medical professionals can make informed decisions based on the insights generated by AI.

It is worth noting how the quantitative assessment of the redesign did not find statistically significant differences between the two interfaces or significant effect sizes, while the verbal protocol and the observations allowed us to gain interesting insights. Of course, there might be multiple reasons, from the small sample to the lack of real challenge because of the lab setting. Yet, a final lesson learned might be an encouragement to apply mixed methods to investigate the use and acceptance of these novel systems to prevent partial and distorted inferences by the users.

Overall, our research's ultimate goal will be to define methodologies for crafting AI interfaces aligned with user values, able to give value to human competence while still recognizing the benefits deriving from AI automation offered by AI tools. The lessons learned through our study highlighted several relevant aspects, from the nuanced nature of user preferences to the interplay between system autonomy and user control, and the importance of identifying a trade-off between automation efficiency and explanation supports in AI-assisted medical systems. Designing interfaces that align with these values is paramount for the successful adoption of such medical systems and positive user experiences.

CRediT authorship contribution statement

Giuseppe Desolda: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Giovanni Dimauro:** Writing – review & editing, Resources, Funding acquisition, Conceptualization. **Andrea Esposito:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rosa Lanzilotti:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Maristella Matera:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Massimo Zancanaro:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Giovanni Dimauro, Andrea Esposito, Rosa Lanzilotti, Maristella Matera, Massimo Zancanaro reports financial support was provided by Italian Ministry of Research and University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Professor Matteo Gelardi for the insightful comments about Rhyno-Cyt and our study, as well as for his invaluable support in helping us to recruit medical specialists for the interviews and the study. Furthermore, we would like to thank the study participants for their time.

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

The research of Andrea Esposito is funded by a Ph.D. fellowship within the framework of the Italian "D.M. n. 352, April 9, 2022" - under the National Recovery and Resilience Plan, Mission 4, Component 2, Investment 3.3 - Ph.D. Project "Human-Centered Artificial Intelligence (HCAI) techniques for supporting end users interacting with AI systems", co-supported by "Eusoft S.r.l". (CUP H91I22000410007).

Appendix A. Interview

Warm-up

- What is your specialization?
- How many years have you been a cytologist?
- Where do you work?

Main section

- Who are their patients? What symptoms do they typically encounter?
- What is the process you follow to reach a diagnosis?
- Could you please explain the cellular population process?
- Do you use specific methods of classifying cells?
- Do you use specific tools?
- How likely/frequent are classification errors and how do you prevent them?
- Do you collaborate with other doctors?
- Are there challenges you face during your work?
- Are there any difficulties you encounter during your work?

Further possible questions

- Do you have expectations regarding the use of an AI-based system to carry out your work?
- At what stage of the cell population process do you wish to have an AI-based system?
- To what extent would you trust the AI-based system?
- If you are not sure about an answer provided by the system, what information should it provide to help you distinguish a system error from your own error?
- What factors could increase trust in the system?
- What should be the balance between efficiency and accuracy?
- Do you have any idea what information would clarify how the system works?
- How would you like to interact with the system?
- Would you like to understand how your choices influence the final result and/or make the system itself evolve?

Cool off

- In summary, what are your thoughts on the technological revolution of artificial intelligence and how they believe it will change the work of doctors?
- Additional comments

Appendix B. Participants details

See Table B.6.

Appendix C. Questionnaire answers

See Tables C.7–C.9.

Table B.6
Details of the nine participants. The skills in IT technologies are self-evaluated from 1 to 10 (inclusive). The Italian region groups are identified following the standard Nomenclature of Territorial Units for Statistics (NUTS).

Participant	Gender	Age	Skills in IT technologies	Years of experience in rhinocytology	Geographic area
1	M	70	7	18	Insular Area
2	F	66	4	14	North-East
3	F	49	3	18	Central
4	M	67	9	15	North-West
5	F	29	8	2	South
6	M	56	10	17	North-East
7	M	49	8	20	South
8	M	57	8	20	South
9	F	38	5	10	South

Table C.7
Answers to the NASA-TLX questionnaire.

Participants	Original							Redesigned						
	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration	Workload	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration	Workload
P1	30	20	40	10	20	10	21.7	10	10	10	0	10	10	10.2
P2	10	10	10	0	10	10	8.3	40	10	20	10	40	20	19.8
P3	30	10	20	20	20	20	20.0	30	30	30	10	90	30	32.5
P4	30	30	30	10	30	20	25.0	10	10	20	0	20	10	14.4
P5	50	20	20	80	30	20	36.7	20	20	20	80	30	10	29.6
P6	20	10	20	0	10	10	11.7	30	30	50	0	20	30	22.7
P7	60	50	50	10	60	50	46.7	50	20	50	10	50	10	35.8
P8	60	60	60	0	10	10	33.3	70	30	10	0	10	10	21.7
P9	40	10	10	10	50	10	21.7	40	10	20	20	40	10	21.5

Table C.8
Answers to the UTAUT questionnaire.

Participants	Original							Redesigned						
	Overall	Performance Expectancy	Effort Expectancy	Attitude Toward Using Technology	Behavioral Intention to Use the System	Self-Efficacy	Social Influence	Overall	Performance Expectancy	Effort Expectancy	Attitude Toward Using Technology	Behavioral Intention to Use the System	Self-Efficacy	Social Influence
P1	3.7	3.8	4.5	3.8	4.3	5.0	1.0	2.0	1.0	4.5	1.0	1.0	3.3	1.0
P2	2.2	1.0	4.3	2.0	1.0	3.3	1.5	4.4	4.0	5.0	5.0	4.8	4.8	3.0
P3	4.0	4.0	4.3	4.8	4.3	4.0	2.5	4.7	4.5	5.0	4.8	5.0	5.0	4.0
P4	4.5	4.0	5.0	4.8	4.8	5.0	3.5	4.1	4.8	4.0	5.0	4.3	4.3	3.0
P5	4.5	4.0	5.0	4.3	5.0	4.8	4.0	4.9	5.0	5.0	5.0	5.0	4.8	4.5
P6	3.9	4.0	4.8	3.8	3.0	4.8	3.0	3.2	3.0	4.5	3.3	2.0	4.0	2.0
P7	4.0	4.3	4.0	4.8	4.5	3.5	2.5	4.4	5.0	4.0	4.8	5.0	4.5	3.0
P8	4.1	4.8	5.0	5.0	4.0	5.0	1.0	4.0	4.5	4.8	4.0	3.5	5.0	3.0
P9	3.8	3.8	4.0	4.5	3.0	4.3	3.5	4.2	4.0	5.0	4.3	4.3	4.3	2.5

Table C.9
Answers to the TOAST questionnaire.

Participants	Original			Redesigned		
	Reliability	Transparency	Overall	Reliability	Transparency	Overall
P1	6	5	5.5	4.5	1.8	3.15
P2	5.5	1.8	3.65	6.5	5.8	6.15
P3	5.75	6	5.875	6	6	6
P4	6.75	6	6.375	5.25	5.4	5.325
P5	6	6.2	6.1	6.75	6.6	6.675
P6	6.5	6	6.25	6.5	6	6.25
P7	5.25	5.6	5.425	5.5	6.2	5.85
P8	6	7	6.5	7	6.2	6.6
P9	6	5.2	5.6	5.5	6	5.75

References

[1] Shneiderman B. Human-centered AI. 1st ed.. Oxford University Press; 2022.

[2] Paternò F, Burnett M, Fischer G, Matera M, Myers B, Schmidt A. Artificial intelligence versus end-user development: a panel on what are the tradeoffs in daily automations? In: Human-computer interaction – INTERACT 2021: 18th IFIP TC 13 international conference, bari, Italy, August 30 – September 3, 2021, proceedings, part v. Berlin, Heidelberg: Springer-Verlag; 2021, p. 340–3. http://dx.doi.org/10.1007/978-3-030-85607-6_33.

[3] Shneiderman B. Human-centered artificial intelligence: reliable, safe & trustworthy. Int J Hum–Comput Interact 2020;36(6):495–504. http://dx.doi.org/10.1080/10447318.2020.1741118.

[4] Schmidt A, Herrmann T. Intervention user interfaces: A new interaction paradigm for automated systems. Interactions 2017;24(5):40–5. http://dx.doi.org/10.1145/3121357.

[5] Suresh H, Gomez SR, Nam KK, Satyanarayan A. Beyond expertise and roles: a framework to characterize the stakeholders of interpretable machine learning and their needs. In: Proceedings of the 2021 CHI conference on human factors in computing systems. CHI '21, New York, NY, USA: Association for Computing Machinery; 2021, p. 1–16. http://dx.doi.org/10.1145/3411764.3445088.

[6] Russell S. Human compatible: artificial intelligence and the problem of control. New York?: Viking; 2019.

[7] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. NIPS'17, Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 4768–77.

[8] Cai CJ, Reif E, Hegde N, Hipp J, Kim B, Smilkov D, Wattenberg M, Viegas F, Corrado GS, Stumpe MC, Terry M. Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI conference on human factors in computing systems. CHI '19, New

York, NY, USA: Association for Computing Machinery; 2019, p. 1–14. http://dx.doi.org/10.1145/3290605.3300234.

[9] Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Coughlin JF, Guttag JV, Colak E, Ghassemi M. Do as AI say: Susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021;4(1):31. http://dx.doi.org/10.1038/s41746-021-00385-9.

[10] Aquino YSJ, Rogers WA, Braunack-Mayer A, Frazer H, Win KT, Houssami N, Degeling C, Semsarian C, Carter SM. Utopia versus dystopia: professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. Int J Med Inform 2023;169:104903. http://dx.doi.org/10.1016/j.ijmedinf.2022.104903.

[11] Costabile MF, Desolda G, Dimauro G, Lanzilotti R, Loiacono D, Matera M, Zancanaro M. A human-centric AI-driven framework for exploring large and complex datasets. In: Barricelli BR, Fischer G, Fogli D, Mørch A, Piccinno A, Valtolina S, editors. Proceedings of the 6th international workshop on cultures of participation in the digital age: AI for humans or humans for AI? CEUR workshop proceedings, vol. 3136, Aachen: CEUR-WS; 2022, p. 9–13.

[12] Dimauro G, Ciprandi G, Deperte F, Girardi F, Ladisa E, Latrofa S, Gelardi M. Nasal cytology with deep learning techniques. Int J Med Inform 2019;122:13–9. http://dx.doi.org/10.1016/j.ijmedinf.2018.11.010.

[13] Dimauro G, Girardi F, Gelardi M, Bevilacqua V, Caivano D. Rhino-cyt: a system for supporting the rhinologist in the analysis of nasal cytology. In: Huang D-S, Jo K-H, Zhang X-L, editors. Proceedings of the 14th international conference on intelligent computing theories and application. ICIC 2018, Lecture notes in computer science, vol. 10955, Cham: Springer International Publishing; 2018, p. 619–30. http://dx.doi.org/10.1007/978-3-319-95933-7_71.

[14] Giacomello E, Lanzi PL, Loiacono D, Nassano L. Image embedding and model ensembling for automated chest X-Ray interpretation. In: 2021 international joint conference on neural networks. IJCNN, Shenzhen, China: IEEE; 2021, p. 1–8. http://dx.doi.org/10.1109/IJCNN52387.2021.9534378.

[15] Santoni De Sio F, Van Den Hoven J. Meaningful human control over autonomous systems: a philosophical account. Front Robot AI 2018;5:15. http://dx.doi.org/10.3389/frobt.2018.00015.

[16] Schmidt A. Interactive human centered artificial intelligence: a definition and research challenges. In: Proceedings of the international conference on advanced visual interfaces. AVI '20, New York, NY, USA: Association for Computing Machinery; 2020, p. 1–4. http://dx.doi.org/10.1145/3399715.3400873.

[17] Liao QV, Varshney KR. Human-centered explainable AI (XAI): from algorithms to user experiences. 2022, http://dx.doi.org/10.48550/arXiv.2110.10790, arXiv:2110.10790.

[18] Ardito C, Buono P, Costabile MF, Lanzilotti R, Piccinno A. End users as co-designers of their own tools and products. J Vis Lang Comput 2012;23(2):78–90. http://dx.doi.org/10.1016/j.jvlc.2011.11.005.

[19] Fischer G, Fogli D, Piccinno A. Revisiting and broadening the meta-design framework for end-user development. In: Paternò F, Wulf V, editors. New perspectives in end-user development. Cham: Springer International Publishing; 2017, p. 61–97. http://dx.doi.org/10.1007/978-3-319-60291-2_4.

[20] Holzinger A. Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inform 2016;3(2):119–31. http://dx.doi.org/10.1007/s40708-016-0042-6.

[21] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2018;51(5). http://dx.doi.org/10.1145/3236009.

[22] Liao QV, Gruen D, Miller S. Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI conference on human factors in computing systems. CHI '20, New York, NY, USA: Association for Computing Machinery; 2020, p. 1–15. http://dx.doi.org/10.1145/3313831.3376590.

[23] Bertrand A, Belloum R, Eagan JR, Maxwell W. How cognitive biases affect XAI-assisted decision-making: a systematic review. In: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society. Oxford United Kingdom: ACM; 2022, p. 78–91. http://dx.doi.org/10.1145/3514094.3534164.

[24] Cabitza F, Campagner A, Ronzio L, Cameli M, Mandoli GE, Pastore MC, Sconfienza LM, Folgado D, Barandas M, Gamboa H. Rams, hounds and white boxes: investigating human–AI collaboration protocols in medical diagnosis. Artif Intell Med 2023;138:102506. http://dx.doi.org/10.1016/j.artmed.2023.102506.

[25] Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc ACM Hum-Comput Interact 2021;5(CSCW1):1–21. http://dx.doi.org/10.1145/3449287.

[26] Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: why did you say that? visual explanations from deep networks via gradient-based localization. 2016, CoRR abs/1610.02391. arXiv:1610.02391.

[27] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the 9th international conference on learning representations. ICLR 2021, Virtual Event, Austria: OpenReview.net; 2021, abs/2010.11929.

[28] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems. Vol. 30, Curran Associates, Inc.; 2017.

[29] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206–15. http://dx.doi.org/10.1038/s42256-019-0048-x.

[30] Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, Guidotti R, Del Ser J, Díaz-Rodríguez N, Herrera F. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion 2023;99:101805. http://dx.doi.org/10.1016/j.inffus.2023.101805.

[31] Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, Sharp RR. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021;4(1):140. http://dx.doi.org/10.1038/s41746-021-00509-1.

[32] Lai V, Chen C, Smith-Renner A, Liao QV, Tan C. Towards a science of human-AI decision making: an overview of design space in empirical human-subject studies. In: 2023 ACM conference on fairness, accountability, and transparency. Chicago IL USA: ACM; 2023, p. 1369–85. http://dx.doi.org/10.1145/3593013.3594087.

[33] Nogueira RG, Abdalkader M, Qureshi MM, Frankel MR, Mansour OY, Yamagami H, Qiu Z, Farhoudi M, Siegler JE, Yaghi S, et al. Global impact of COVID-19 on stroke care. Int J Stroke 2021;16(5):573–84. http://dx.doi.org/10.1177/1747493021991652.

[34] Maassen O, Fritsch S, Palm J, Deffge S, Kunze J, Marx G, Riedel M, Schuppert A, Bickenbach J. Future medical artificial intelligence application requirements and expectations of physicians in german university hospitals: web-based survey. J Med Internet Res 2021;23(3):e26646. http://dx.doi.org/10.2196/26646.

[35] Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res 2022;24(8):e36823. http://dx.doi.org/10.2196/36823.

[36] Sheu R-K, Pardeshi MS. A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system. Sensors 2022;22(20):8068. http://dx.doi.org/10.3390/s22208068.

[37] Holzinger A, Muller H. Toward human–AI interfaces to support explainability and causability in medical AI. Computer 2021;54(10):78–86. http://dx.doi.org/10.1109/MC.2021.3092610.

[38] Procter R, Tolmie P, Rouncefield M. Holding AI to account: challenges for the delivery of trustworthy AI in healthcare. ACM Trans Comput-Hum Interact 2023;30(2):31:1–34. http://dx.doi.org/10.1145/3577009.

[39] Gelardi M. Atlas of nasal cytology. 14th ed.. Edi.Ermes s.r.l.; 2012.

[40] Zimmerman J, Forlizzi J, Evenson S. Research through design as a method for interaction design research in hci. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI '07, New York, NY, USA: Association for Computing Machinery; 2007, p. 493–502. http://dx.doi.org/10.1145/1240624.1240704.

[41] Holzinger A. Rapid prototyping for a virtual medical campus interface. IEEE Softw 2004;21(1):92–9. http://dx.doi.org/10.1109/MS.2004.1259241.

[42] Setzu M, Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F. Glocalx - from local to global explanations of black box AI models. Artificial Intelligence 2021;294:103457. http://dx.doi.org/10.1016/j.artint.2021.103457.

[43] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco California USA: ACM; 2016, p. 1135–44. http://dx.doi.org/10.1145/2939672.2939778.

[44] Wang L, Zhang X, Su H, Zhu J. A comprehensive survey of continual learning: theory, method and application. 2023, http://dx.doi.org/10.48550/arXiv.2302.00487, arXiv:2302.00487.

[45] Norman DA. The design of everyday things. revised and expanded ed.. Cambridge, MA London: The MIT Press; 2013.

[46] Dove G, Halskov K, Forlizzi J, Zimmerman J. UX design innovation: challenges for working with machine learning as a design material. In: Proceedings of the 2017 CHI conference on human factors in computing systems. CHI '17, New York, NY, USA: Association for Computing Machinery; 2017, p. 278–88. http://dx.doi.org/10.1145/3025453.3025739.

[47] Holmquist LE. Intelligence on tap: Artificial intelligence as a new design material. Interactions 2017;24(4):28–33. http://dx.doi.org/10.1145/3085571.

[48] Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett PN, Inkpen K, Teevan J, Kikin-Gil R, Horvitz E. Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI conference on human factors in computing systems. CHI '19, New York, NY, USA: Association for Computing Machinery; 2019, p. 1–13. http://dx.doi.org/10.1145/3290605.3300233.

[49] Google PAIR. People + AI guidebook. 2019, https://pair.withgoogle.com/guidebook.

[50] Cabitza F, Campagner A, Malgieri G, Natali C, Schneeberger D, Stoeger K, Holzinger A. Quod erat demonstrandum? - towards a typology of the concept of explanation for the design of explainable AI. Expert Syst Appl 2023;213:118888. http://dx.doi.org/10.1016/j.eswa.2022.118888, URL https://www.sciencedirect.com/science/article/pii/S0957417422019066.

[51] Miller T. Explanation in artificial intelligence: insights from the social sciences. Artificial Intelligence 2019;267:1–38. http://dx.doi.org/10.1016/j.artint.2018.07.007.

[52] Leichtmann B, Humer C, Hinterreiter A, Streit M, Mara M. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. Comput Hum Behav 2023;139:107539. http://dx.doi.org/10.1016/j.chb.2022.107539.

[53] Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum-Comput Stud 2021;146:102551. http://dx.doi.org/10.1016/j.ijhcs.2020.102551.

[54] Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. In: Proceedings 1996 IEEE symposium on visual languages. 1996, p. 336–43. http://dx.doi.org/10.1109/VL.1996.545307.

[55] Famiglini L, Campagner A, Cabitza F. Towards a rigorous calibration assessment framework: advancements in metrics, methods, and use. In: Gal K, Nowé A, Nalepa GJ, Fairstein R, Rădulescu R, editors. Frontiers in artificial intelligence and applications. IOS Press; 2023, http://dx.doi.org/10.3233/FAIA230327.

[56] Hart SG. Nasa-task load index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 2006;50(9):904–8. http://dx.doi.org/10.1177/154193120605000909.

[57] Lazar J, Lazar JH, Hochheiser H. Research methods in human-computer interaction. Elsevier; 2017.

[58] Williams MD, Rana NP, Dwivedi YK. The unified theory of acceptance and use of technology (UTAUT): A literature review. J Enterp Inf Manag 2015;28(3):443–88. http://dx.doi.org/10.1108/JEIM-09-2014-0088.

[59] Wojton HM, Porter D, Lane ST, Bieber C, Madhavan P. Initial validation of the trust of automated systems test (TOAST). J Soc Psychol 2020;160(6):735–50. http://dx.doi.org/10.1080/00224545.2020.1749020.

[60] Friedman B, Hendry D. Value sensitive design: shaping technology with moral imagination. Cambridge, Massachusetts: The MIT Press; 2019.

[61] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed.. Routledge; 2013, http://dx.doi.org/10.4324/9780203771587.

[62] Chaddad A, Katib Y, Hassan L. Future artificial intelligence tools and perspectives in medicine. Curr Opin Urol 2021;31(4):371–7. http://dx.doi.org/10.1097/MOU.0000000000000884.

[63] Sheth D, Giger ML. Artificial intelligence in the interpretation of breast cancer on MRI. J Magn Reson Imaging 2020;51(5):1310–24. http://dx.doi.org/10.1002/jmri.26878.

[64] Goyal H, Mann R, Gandhi Z, Perisetti A, Ali A, Aman Ali K, Sharma N, Saligram S, Tharian B, Inamdar S. Scope of artificial intelligence in screening and diagnosis of colorectal cancer. J Clin Med 2020;9(10):3313. http://dx.doi.org/10.3390/jcm9103313.

[65] Shneiderman B, Plaisant C, Cohen M, Jacobs S, Elmqvist N. Designing the user interface: strategies for effective human-computer interaction. sixth ed., global ed.. Boston Columbus Indianapolis New York San Francisco Hoboken Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo: Pearson; 2018.

[66] Kalyuga S, Ayres P, Chandler P, Sweller J. The expertise reversal effect. Educ Psychol 2003;38(1):23–31. http://dx.doi.org/10.1207/S15326985EP3801_4.

[67] Zielinska OA, Welk AK, Mayhorn CB, Murphy-Hill E. Exploring expert and novice mental models of phishing. Proc Hum Factors Ergon Soc Annu Meet 2015;59(1):1132–6. http://dx.doi.org/10.1177/1541931215591165.

[68] Paas F, Renkl A, Sweller J. Cognitive load theory: instructional implications of the interaction between information structures and cognitive architecture. Instr Sci 2004;32(1/2):1–8. http://dx.doi.org/10.1023/B:TRUC.0000021806.17516.d0.

[69] Akiki PA, Bandara AK, Yu Y. Adaptive model-driven user interface development systems. ACM Comput Surv 2014;47(1):1–33. http://dx.doi.org/10.1145/2597999.

[70] Lavie N, Hirst A, De Fockert JW, Viding E. Load theory of selective attention and cognitive control. J Exp Psychol: Gen 2004;133(3):339–54. http://dx.doi.org/10.1037/0096-3445.133.3.339.

[71] Wang Y, Mahmud J, Liu T. Understanding cognitive styles from user-generated social media content. In: Proceedings of the international AAAI conference on web and social media. Vol. 10, 2021, p. 715–8. http://dx.doi.org/10.1609/icwsm.v10i1.14775, (1).

[72] Pillay H, Boles W, Raj L. Personalizing the design of computer-based instruction to enhance learning. Res Learn Technol 2011;6(2). http://dx.doi.org/10.3402/rlt.v6i2.11004.

[73] Sundar SS. Rise of machine agency: a framework for studying the psychology of human–AI interaction (HAII). J Comput-Mediat Commun 2020;25(1):74–88. http://dx.doi.org/10.1093/jcmc/zmz026.

[74] Horvitz E. Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI conference on human factors in computing systems the CHI is the limit - CHI '99. Pittsburgh, Pennsylvania, United States: ACM Press; 1999, p. 159–66. http://dx.doi.org/10.1145/302979.303030.

[75] Reiter E, Sripada SG, Robertson R. Acquiring correct knowledge for natural language generation. J Artificial Intelligence Res 2003;18:491–516. http://dx.doi.org/10.1613/jair.1176.

[76] Blandford A. HCI for health and wellbeing: challenges and opportunities. Int J Hum-Comput Stud 2019;131:41–51. http://dx.doi.org/10.1016/j.ijhcs.2019.06.007.

[77] Sambasivan N, Veeraraghavan R. The deskilling of domain expertise in AI development. In: CHI conference on human factors in computing systems. New Orleans LA USA: ACM; 2022, p. 1–14. http://dx.doi.org/10.1145/3491102.3517578.

[78] Hoff T. Deskilling and adaptation among primary care physicians using two work innovations. Health Care Manag Rev 2011;36(4):338–48. http://dx.doi.org/10.1097/HMR.0b013e31821826a1.

[79] Troya J, Fitting D, Brand M, Sudarevic B, Kather JN, Meining A, Hann A. The influence of computer-aided polyp detection systems on reaction time for polyp detection and eye gaze. Endoscopy 2022;54(10):1009–14. http://dx.doi.org/10.1055/a-1770-7353.

[80] Cabitza F, Natali C, Famiglini L, Campagner A, Caccavella V, Gallazzi E. Never tell me the odds: investigating pro-hoc explanations in medical decision making. Artif Intell Med 2024;150:102819. http://dx.doi.org/10.1016/j.artmed.2024.102819.