

Article

Towards HCXAI: Explainability Preferences of Healthcare Professionals [†]

Mishell Cadena-Yanez ^{1,2,*} , Angela Bernardini ^{1,2}  and Marisol Gómez ² 

¹ Navarra Artificial Intelligence Research Center (NAIR Center), C. del Sadar, s/n, 31006 Pamplona, Spain

² Department of Statistics, Computer Science and Mathematics, Institute for Advanced Materials and Mathematics (INAMAT²), Campus Arrosadía, Public University of Navarre (UPNA), 31006 Pamplona, Spain

* Correspondence: mishell.cadena.yanez@naircenter.com

[†] This paper is an extended version of our paper accepted at HAICAI 2026, currently in press in the conference proceedings.

Abstract

This study examined how healthcare professionals (HCP) perceive and trust AI in clinical settings, and what kinds of explanations they need to use it effectively. Using interviews with six HCPs and a questionnaire ($N = 41$), the findings show that trust was higher for administrative tasks and lower for direct clinical decisions, regardless of their prior AI or clinical experience or AI literacy. Clinical validation and algorithmic bias ranked above explainability as trust factors, indicating that explainability is not a primary trust-building mechanism. However, HCP consistently demanded explanations as a tool to support their own clinical reasoning, preferring to receive them across all clinical contexts rather than only under disagreement with the system, and valued grounding in medical evidence and consistency with clinical protocols over clarity or simplicity. These findings argue for an HCXAI design approach that treats explainability as a critical-reasoning tool rather than a primary trust mechanism.

Keywords: explainable artificial intelligence (XAI); healthcare professionals; clinical decision support; local explanations

1. Introduction

Insufficient trust among healthcare professionals (HCP) has been identified as one of the primary barriers to the adoption of AI systems in clinical practice [1,2]. Trust in AI healthcare systems is a multidimensional construct influenced by robustness, fairness, and explainability (XAI) [3]. Shevtsova (2024) [2] further argues that trust and acceptance can be fostered along four axes: human-related (i.e., the type of institution AI professionals originate from), technology-related (i.e., the explainability and transparency of AI application processes and outcomes), ethical and legal (i.e., data use transparency), and additional factors (i.e., the environmental sustainability of AI applications). Within these axes, it is useful to distinguish between factors that operate prior to clinical use and those that operate at the moment of use. Prior to deployment, clinical validation—the process by which an AI system demonstrates adequate performance under real-world conditions—and its formal institutional recognition through regulatory certification, such as the conformity assessment required for high-risk AI systems under Regulation (EU) 2024/1689 [4], constitute external guarantees independent of any individual professional's judgment. Transparency, understood as the systemic openness of a system with respect to its training data, known limitations, and operational criteria [5], operates at this same pre-deployment level. In the



Academic Editors: Phivos Mylonas,
Ioannis Voyiatzis and Nicolas
Tsapatsoulis

Received: 21 May 2026

Revised: 18 June 2026

Accepted: 22 June 2026

Published: 2 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

moment of use, explainability—the property of a system whereby its individual outputs can be rendered interpretable to the end user at the point of clinical decision-making, operationalised through the explanations that XAI techniques generate [5]—functions as a mechanism for calibrating and justifying trust. However, whether this function holds in practice, and what weight it carries relative to other trust factors among healthcare practitioners, remains empirically underexplored.

Although trust has been cited as a primary motivation for deploying XAI for end users [6], the explanations it produces serve a broader set of purposes for end users: they can support enhanced decision quality, facilitate learning and skill development, enable critical evaluation of AI outputs [7,8], or allow clinicians to communicate AI-generated results to colleagues and patients [9]. Reducing explanation design to a trust-generation mechanism therefore risks misaligning it with the actual functional needs of HCP. This misalignment is reflected in current XAI practice. Despite growing interest in XAI for healthcare, existing approaches often fail to place the clinical user at the centre of explanation design [10]. As a result, explanations may be technically sound yet clinically irrelevant—or, conversely, may create a superficial sense of confidence in the system without supporting genuine clinical reasoning [11]. A Human-Centered XAI (HCXAI) approach seeks to bridge this gap by aligning explanation design with the actual needs, context, and workflows of HCP. This requires understanding what HCP need from AI explanations in terms of content, format, and context, as well as what information they require to achieve justified trust—rather than mere acceptance—of AI systems [11–13]. This study adopts a sequential exploratory mixed-methods design, combining semi-structured interviews to identify relevant themes with a quantitative survey administered to HCP. Two research questions guide the study: (RQ1) How do HCP's perceptions of current AI limitations influence their differentiated trust across clinical tasks, and what role does explainability play in justified trust? (RQ2) What explanation needs do HCPs have regarding AI system recommendations, and how does agreement with the system's recommendation influence those needs?

2. Methodology

In the first phase, semi-structured interviews with 6 participants were conducted with healthcare professionals at the University Hospital of Navarra, representing the specialties of General Medicine, Emergency Medicine, Radiology, Geriatrics, Gastroenterology, and Medical Education. Participants were recruited through purposive sampling, targeting hospital-based clinical staff across specialties; participation was voluntary. The interview with the Medical Education professional was adapted to address the use of AI in clinical training contexts. Prior to each interview, participants received written information about the study's purpose and scope and provided informed consent authorizing audio recording and the use of their data for research purposes. Each interview lasted approximately one hour and was conducted in person. The interview guide explored participants' perceptions of AI in clinical settings, their degree of familiarity with AI tools, and their preferences regarding explanation modalities in relation to the specific demands of their clinical activity. The audio-recorded interviews were analyzed using empathy mapping organizing participant responses across four dimensions: what participants “say”, “think”, “do”, and “feel”. Cross-map analysis identified recurring functional patterns across participants concerning system limitations, desired clinical applications, information presentation preferences, required informational content, and intended uses of AI outputs, which informed the thematic block structure of the questionnaire.

This questionnaire (Supplementary Material S1), was sent by email to HCPs. The email contained information about the study, data protection information and the link online to $N = 41$ HCP predominantly based in Navarra, Spain, consisted of 24 questions

organized into thematic sections. The questions included preferences, on a Likert scale, multiple choices, and comments. The first section collects the sociodemographic profile and the level of familiarity with AI of the participants. Participants were divided into those *with prior AI experience* ($n_1 = 21$) and *without AI experience* ($n_2 = 20$). HCP *with prior AI experience* answered additional questions about their practical experience. The sections ‘Trust and perception of HCP towards AI systems’, ‘Factors that enable trust in AI systems’, ‘Explainability needs of HCP’ and ‘Availability for follow-up’ form the ‘common block’, answered by all participants (see Figure 1). The questionnaire, designed in accordance with the GDPR and administered in Spanish through a Google Forms link, has been available since January 2026 (https://docs.google.com/forms/d/e/1FAIpQLSdgNQijtHjqn6rHDIA1Ey3iZEqjFiHEuZ-iF3FOL1Y_cOXRA/viewform?usp=sharing&ouid=100270201798660355943 (accessed on 21 June 2026)).

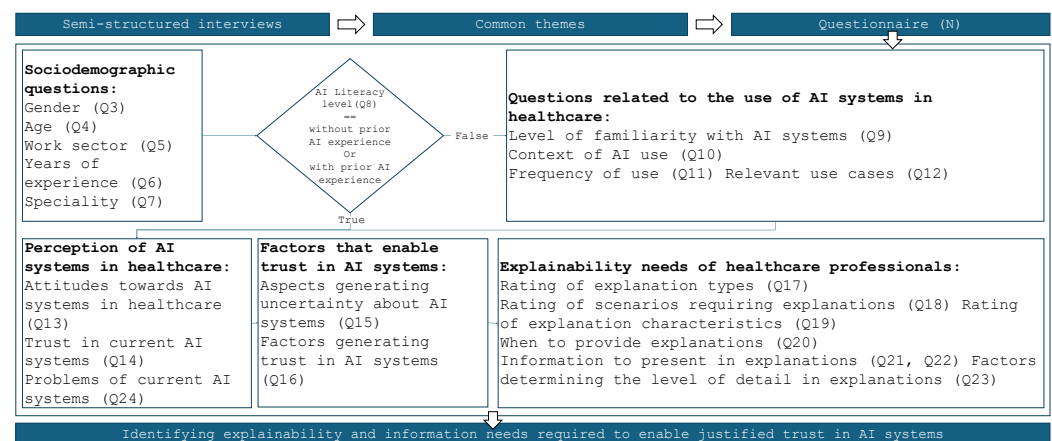


Figure 1. Structure of the questionnaire, grouped according to thematic blocks.

Statistical analyses were performed using Python (version 3.12.3), with the pandas, scipy, and statsmodels libraries. Absolute frequencies and percentages were computed for all sociodemographic and AI-related variables (Q3–Q13), including one attitudinal item on perceptions of AI systems. For Likert-scale items, top-two-box scores (collapsing “agree” and “strongly agree” response options) were used for descriptive reporting, while the full five-point ordinal scale was retained for inferential tests. To assess whether responses differed across sociodemographic or AI literacy subgroups, Mann-Whitney U and Chi-square tests were applied to the relevant variables (gender, years of experience, familiarity with AI systems, and context of use). No significant between-group differences were found (all $p > 0.05$). Speciality was not analysed due to the small number of participants per group. To assess whether response options within the same Likert-scale item differed significantly from one another, the Friedman test was applied to (Q14, Q16–Q23), followed by pairwise Wilcoxon signed-rank tests as post-hoc comparisons. p -values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) correction for multiple comparisons. The significance threshold was set at $\alpha = 0.05$. Response options that did not differ significantly from one another were assigned to the same significance group, allowing the identification of distinct tiers of endorsement within each item.

3. Results

3.1. Sociodemographic Profile of Participants

The sample (N) was predominantly female (65.9%) and relatively young, with 61% under 40 years of age. Most worked in the public sector (56.1%). Professional experience varied widely, with the largest groups being those with less than 5 years (43.9%) and more than 20 years of experience (36.6%). The most represented specialties were general medicine

(23.7%), nursing (13.1%) and dentistry (13.1%), followed by surgery (10.5%), dermatology (10.5%) and pediatrics (7.9%), among others. Table 1 presents the full data.

Table 1. Sociodemographic profile of participants (N).

Variable	Category	Number of Participants	Proportion (%)
Age (years)	<30	14	34.1
	30–50	14	34.1
	>50	13	31.7
Gender	Female	27	65.9
	Male	14	34.1
Work sector	Public	23	56.1
	Private	11	26.8
	Both	7	17.1
Years of experience	<5	18	43.9
	5–20	8	12.2
	>20	15	36.6
Specialty ^a	General medicine	9	23.7
	Nursing	5	13.1
	Dentistry	5	13.1
	Surgery	4	10.5
	Dermatology	4	10.5
	Pediatrics	3	7.9
	Radiology	2	5.3
	Other	6	15.9

^a Specialty reported by 38 of 41 participants (3 missing); percentages computed over valid responses (38 participants). Other includes: Pathological anatomy, pharmacy, ophthalmology, neurology, gynaecology, and geriatrics (1 each).

3.2. Use of AI Healthcare Systems

Table 2 summarizes the use of AI in healthcare systems. In prior AI experience the most common response was uncertainty about whether the digital tools used incorporated AI (34.1%), followed by use under supervision or with specific training (26.8%), regular use in daily clinical practice (14.6%), no conscious use (14.6%), and specific training in AI applied to healthcare (9.8%). For subsequent analyses, participants were classified as *with prior AI experience*—those who had used medical AI systems under supervision or training, used AI regularly in clinical practice, or held specific AI training (n_1)—or *without prior AI experience*—those who had never consciously used AI systems or were uncertain whether the tools they used incorporated AI (n_2). Among participants *with prior AI experience*, self-assessed AI literacy—encompassing understanding of algorithm functioning, ability to interpret performance metrics, and knowledge of limitations and biases—was concentrated at the low and moderate levels. Regarding type of use, most reported non-direct clinical applications (90.5%), including searching for medical information, report writing, personal consultations, and learning, while 71.4% also reported clinical uses such as early detection, diagnostic support, treatment planning, triage prioritization, and obtaining a second clinical opinion; notably, 61.9% engaged in both categories simultaneously. The co-occurrence of both use types within the same participants suggests heterogeneous interaction patterns with AI, ranging from general-purpose tools to dedicated clinical systems. Regarding frequency of use, weekly use was most common (33.3%), followed by monthly (23.8%), occasional (23.8%), and daily use (19.0%).

Table 2. Use of AI healthcare systems. Prior AI experience is reported for the full sample (N); all remaining variables are reported for participants with prior AI experience only (n_1).

Category	Number of Participants	Proportion (%)
Prior AI experience (N)		
<i>without prior AI experience.</i>	20	48.7
<i>with prior AI experience.</i>	21	51.2
AI literacy (n_1)		
Practical use of AI in own specialty	None: 1 · Low: 8 · Moderate: 5 High: 5 · Very High: 2	None: 4.8% · Low: 38.1% · Moderate: 23.8% High: 23.8% · Very High: 9.5%
Understanding of AI algorithm functioning	None: 2 · Low: 9 · Moderate: 6 High: 4 · Very High: 0	None: 9.5% · Low: 42.9% · Moderate: 28.6% High: 19.0% · Very High: 0.0%
Ability to interpret performance metrics ^a	None: 1 · Low: 8 · Moderate: 10 High: 1 · Very High: 0	None: 5.0% · Low: 40.0% · Moderate: 50.0% High: 5.0% · Very High: 0.0%
Knowledge of model limitations and biases ^a	None: 1 · Low: 7 · Moderate: 9 High: 2 · Very High: 1	None: 5.0% · Low: 35.0% · Moderate: 45.0% High: 10.0% · Very High: 5.0%
Type of use (n_1)		
Clinical and non-clinical	13	61.9
Non-clinical only	6	28.6
Clinical only	2	9.5
Frequency of use (n_1)		
Daily	4	19.0
Weekly	7	33.3
Monthly	5	23.8
Rarely	5	23.8

^a One participant *with prior AI experience* did not respond to these two items. Clinical use includes: early detection, diagnostic support, treatment planning, triage prioritisation, and second clinical opinion. Non-clinical use includes: medical information search, report writing, personal consultations, and learning.

3.3. Trust and Perception of HCP Towards AI Systems

In this sample, HCPs reported their general attitudes towards healthcare AI systems (Q13). The need to increase AI literacy received the strongest endorsement, with 87% of respondents agreeing or strongly agreeing. This was followed by the call for greater regulation of AI systems in healthcare (74% agreement) and greater transparency and oversight in AI development (80% agreement). The expectation that new ways of working with AI will emerge was also broadly shared (88% agreement). In contrast, the idea that AI will replace the role of healthcare staff was clearly rejected, with 68% of respondents disagreeing or strongly disagreeing. Opinions on the potential impact of AI on critical thinking and decision-making were more divided: while 57% agreed that AI will affect these processes, 23% disagreed and 20% remained neutral. Trust in AI recommendations varied significantly across task types (Q14). Tasks involving direct clinical decision-making—including disease diagnosis, anomaly detection in medical imaging, clinical severity or risk estimation, and treatment recommendations—received consistently lower trust ratings than non-decisional support tasks such as clinical report writing, support in medical consultations, and healthcare resource management. To examine whether this difference was statistically significant, tasks were grouped into two categories: decision making task and support tasks. A Wilcoxon signed-rank test confirmed a significant difference between the two groups ($p < 0.001$), with administrative tasks receiving systematically higher trust

ratings than direct clinical tasks. Figure 2 summarizes the general perceptions and trust of participants in AI healthcare systems.

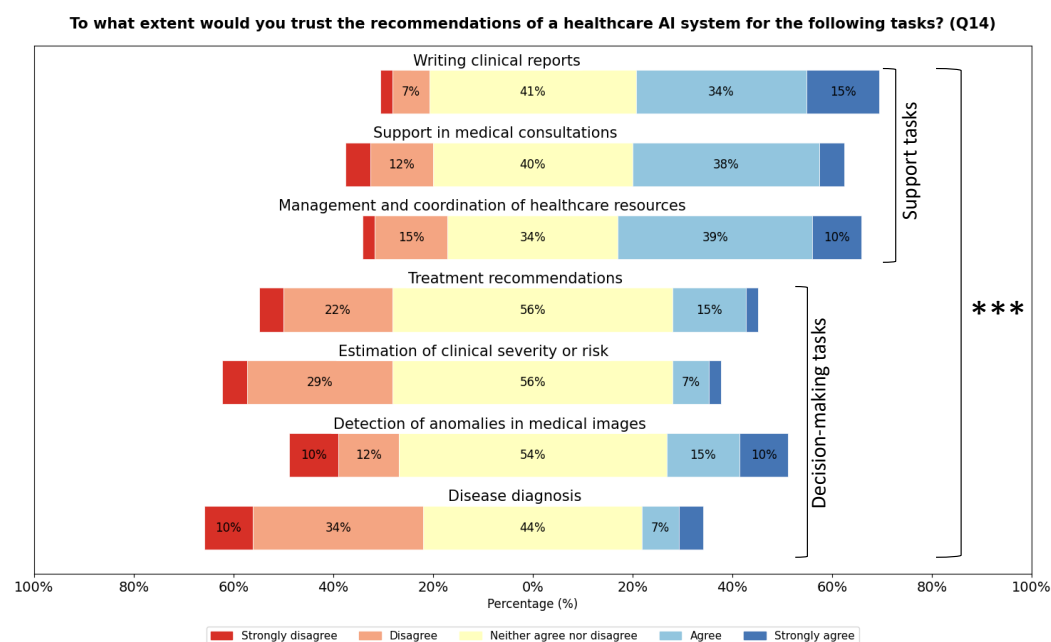


Figure 2. Level of trust in AI systems (Q14). *** indicates a statistically significant difference between support tasks and decision-making tasks ($p < 0.001$).

As shown in Figure 3, among the main concerns reported by HCPs regarding the information provided by AI systems (Q24), the most prominent were overly generic responses, lack of clinical context, and insufficient detail to support clinical decision-making. These concerns were also present among participants *with prior AI experience*; however, those reporting non-direct clinical uses more frequently identified an excess of irrelevant information. This may suggest that participants in this group were referring primarily to the limitations of generative AI tools—such as large language models—rather than to dedicated clinical AI systems, given the nature of the tasks they reported performing.

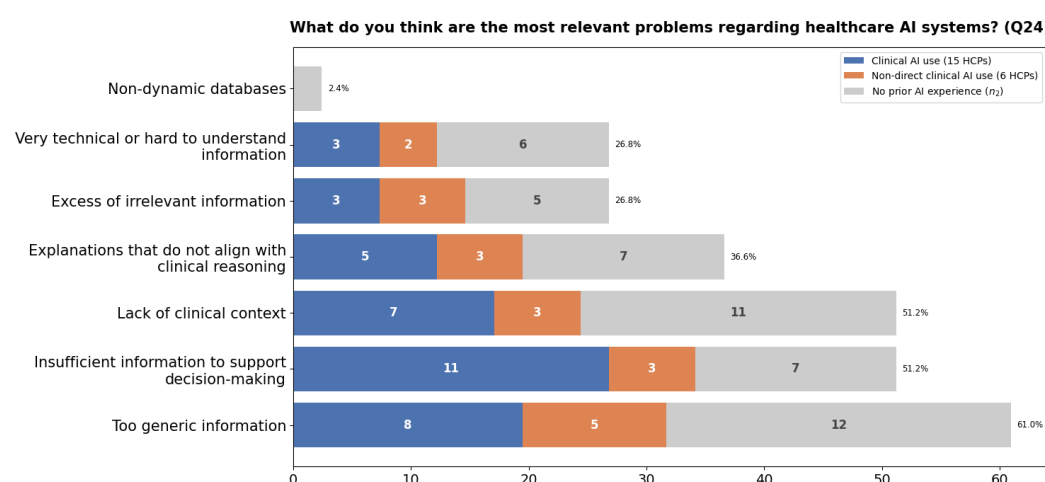


Figure 3. Concerns reported by HCPs regarding information provided by AI systems (Q24). Numbers inside the bars indicate the number of participants who selected each option.

3.4. Factors That Enable Trust in AI Systems

The correspondence between Q15, on aspects generating uncertainty, and Q16, on factors enabling trust, suggests that HCPs' information needs directly target their main

sources of distrust. Clinical certification or validation emerged as the highest-rated factor (Top-2-Box = 86%), forming a distinct significance group together with algorithmic bias and performance metrics (group a/ab), all of which differed significantly from understanding of general system functioning—the lowest-rated item (Top-2-Box = 61%, group c). Training data quality and explanation of individual recommendations occupied an intermediate position (group b; Top-2-Box = 73% and 66% respectively). Explainability-related items—explanation of individual recommendations and understanding of general system functioning—were consistently endorsed at lower levels than institutional reliability factors, although still positively rated by the majority of participants. Table 3 summarizes the factors that influence HCPs' trust in AI systems.

Table 3. Q15: percentage of participants selecting each factor as a source of uncertainty (*N*, multiple selection). Q16: top-2-box score of importance of system characteristics for trust. In significance groups (Q16) items sharing a letter do not differ significantly.

Factor	Q15 %	Q16 Top-2-Box %	Sig. Group
Clinical certification or validation	61.0	86	a
Algorithmic bias and system limitations	58.5	81	ab
Performance metrics and case-specific certainty [†]	48.8	78	ab
Training data source and quality	36.6	73	b
Explanation of individual recommendations	26.8	66	b
Understanding of general system functioning	31.7	61	c

[†] Q16 top-2-box value is the average of two related items: general performance metrics (73%) and case-specific certainty level (83%); significance group reflects post-hoc results for the individual items.

3.5. Explainability Needs of HCP

Results are presented following the framework of Barda et al. [12], which organized explainability needs into two main dimensions: the Context of Use, which captures the factors that determine when and why the professional needs an explanation (Q18, Q20, Q21, Q22), and the Explanation Design, which defines what information it should contain and how it should be presented (Q17, Q19, Q23). Regarding the scenarios in which HCPs considered explanations most useful (Q18), a statistically significant global difference across scenarios ($p = 0.040$); however, no individual pairwise differences survived correction (all $p > 0.05$). Descriptively, however, explanations were most strongly endorsed in situations of high uncertainty or clinical risk: uncertainty in the initial diagnosis (70%), high-severity diagnosis or treatment (66%), and high risk associated with the clinical decision (62%). Routine screening cases and the use of explanations to communicate decisions to patients received slightly lower endorsement.

In the case of scenarios demanding detailed explanations (Q20), we identified four significance groups across items ($p < 0.001$, see Table 4). Complex or atypical cases (73%) and low or uncertain system performance (70%) formed the highest-ranked group, rated significantly above all remaining items. Receiving explanations always, independently of the clinical case (62%), formed an intermediate group, not differing significantly from the top group but rated above justifying decisions to other professionals. Justifying the decision to other healthcare professionals (53%) was rated significantly below complex cases and low performance but above the remaining items. Restricting explanations only to cases of disagreement with the system (27%) and depending on available time (20%) were both clearly rejected, differing significantly from all positively endorsed items. Never trusting explanations and relying solely on AI predictions received the lowest ratings (4%).

Table 4. Perceived need for detailed explanations in healthcare AI systems (Q20). Items sharing a letter do not differ significantly.

Context for Detailed Explanation	Md	Top-2-Box %	Sig.
When the case is complex or atypical	4	73	a
When the system's performance is low or uncertain	4	70	a
Always, regardless of the clinical case	4	62	ab
When I need to justify the decision to other HCPs	4	53	b
Only when I disagree with the AI recommendation	3	27	c
Depends on the available time	3	20	c
Never, I trust the AI system's prediction	2	4	d

Items ordered by top-2-box score (descending).

These results are related to those obtained in Q21 and Q22, see Figure 4, where the information demanded was largely similar regardless of whether participants agreed or disagreed with the system's recommendation. McNemar tests showed no statistically significant differences between scenarios for any of the items (all $p > 0.05$). In both cases, local explanations of why the system reached the recommendation were the most requested information (54% in disagreement, 61% in agreement), followed by performance metrics (49%/59%) and information about system consistency in similar populations (46%/51%). Descriptive trends, although non-significant, were consistent with differentiated reasoning strategies: counterfactual exploration (46% vs. 34%) and uncertainty information (49% vs. 44%) were more frequently requested under disagreement, while examples of system failures (44% vs. 37%), rules visualization (51% vs. 44%) and population consistency (51% vs. 46%) were prioritised under agreement.

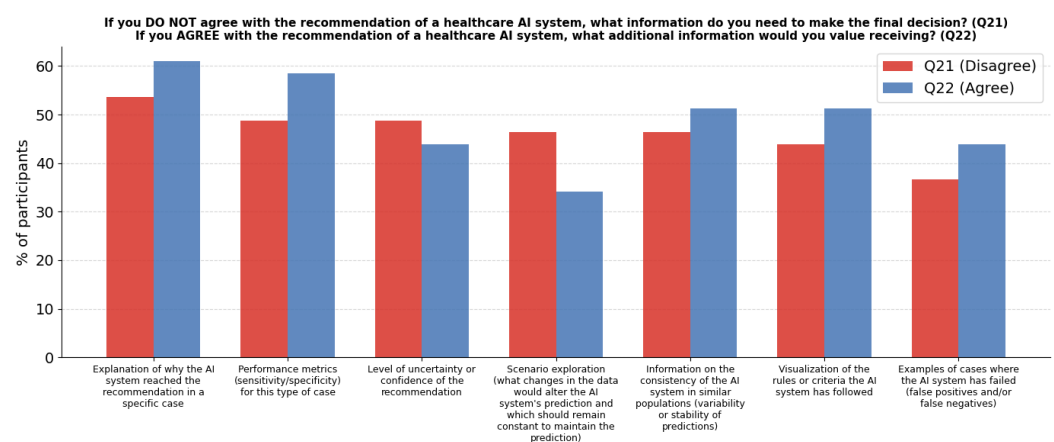


Figure 4. Explanations under agreement and disagreement with the AI system (Q21, Q22).

As for the factors determining the level of detail of explanations (Q23), see Figure 5, the item reflecting rejection of adaptive explanations was excluded from the inferential analysis as it reflects a different construct. AI system knowledge, case urgency, and diagnostic or treatment experience level were similarly endorsed as relevant determinants (group a), with distributions showing an increasing trend towards greater importance. Agreement with the AI system's recommendation received significantly lower ratings than the other three factors (group b), with the majority of participants placing it at the lower end of the importance scale.

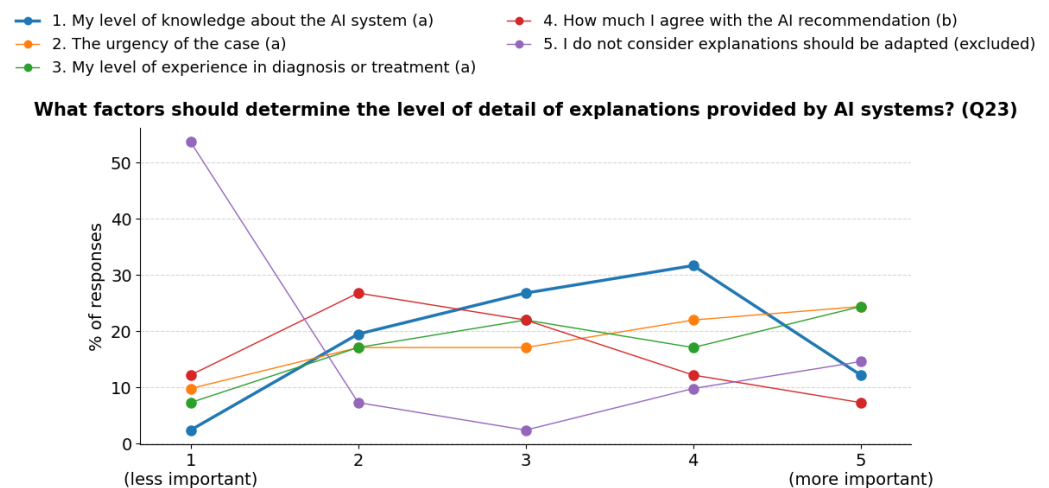


Figure 5. Factors that determine the level of detail of AI-generated explanations, as rated by HCPs (Q23).

Regarding Explanation Design, all five types of explanation proposed in Q17 were positively rated, with top-2-box endorsement above 60% in every case and no statistically significant differences between them. Example-based explanations (visualisation of similar cases correctly identified by the system, 75%) and local explanations (why the system reached a specific recommendation, 73%) showed the highest descriptive endorsement, followed by counterfactual exploration and rules visualisation (70%). Global importance-based explanations received the lowest endorsement (61%), though still positive. The most valued characteristics in AI-generated explanations (Q19), see Figure 6, were grounding in medical evidence, which ranked significantly above all other characteristics (group a), followed by consistency with clinical protocols and technical precision (group b). Clarity and simplicity, speed of understanding, and adaptability received the lowest rankings (group c).

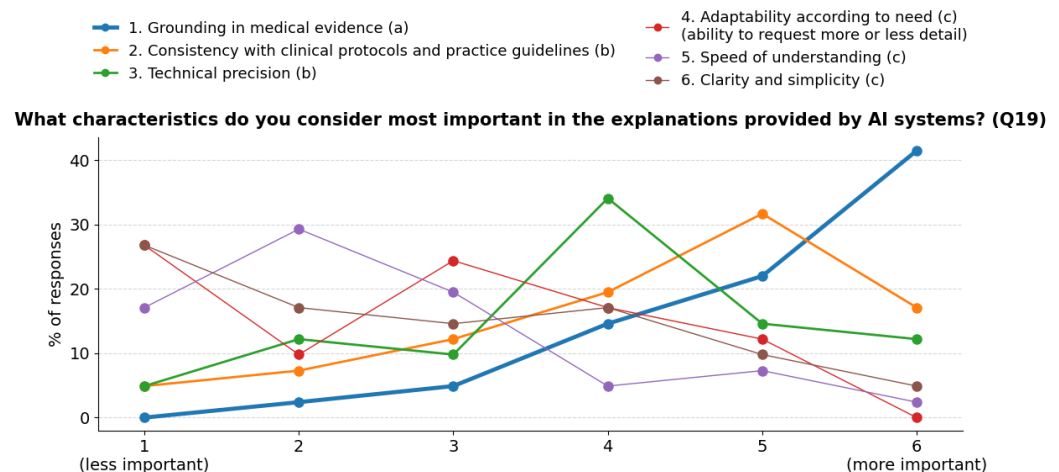


Figure 6. Importance rankings of explanation characteristics as rated by HCPs (Q19).

4. Discussion

4.1. RQ1: How Do HCPs' Perceptions of Current AI Limitations Influence Their Differentiated Trust Across Clinical Tasks, and What Role Does Explainability Play in Justified Trust?

Trust depended on what was at stake in the tasks. It ran high for administrative and documentary support task and fell for direct clinical decision-making, consistent with [2], where the threshold a recommendation must clear rises with the risk of the decision. It

did not appear to depend on clinical experience [5] or AI literacy, which stayed at low-to-moderate levels among the 21 participants who had used such systems. What separated the tasks instead was the information the tools return: participants objected most to output that was too generic (61%), short on clinical context (51.2%), or too thin to act on (51.2%), and they reported these problems for clinical and non-clinical work alike. The reservation is concrete and points to a gap between what the systems provide and what a clinical decision needs. These concerns were reported for both clinical and non-clinical only, regardless of whether the application is a task-specific clinical AI system or a general-purpose large language model. It should be noted, however, that only 6 participants in our sample reported exclusively non-clinical AI use.

Asked what would reduce their uncertainty, participants looked first to external guarantees. Clinical certification and the presence of algorithmic bias were the leading sources of doubt, ahead of interpreting individual predictions or understanding how the model works inside; the importance ratings ordered the same way, with certification set apart from training-data quality, local explanation and general system understanding. The strong support for tighter regulation and for raising AI literacy fits this. HCPs were not, on the whole, asking for a view into the algorithm in order to trust it; they were asking for the institutional and regulatory backing that makes its use legitimate. This is consistent with how clinicians already work, relying on certified devices and validated instruments without auditing their internal workings. It should be noted that the sample is geographically concentrated and that email recruitment may have overrepresented professionals with a prior interest in AI and greater awareness of regulatory frameworks. However, reliance on clinically certified tools is a structural feature of medical practice and the pattern observed is consistent with how clinicians engage with validated instruments across healthcare systems more broadly.

Explainability is not displaced by this, but its place changes. Validation before deployment establishes that a system performs adequately and within known limits [5], which lets the clinician use it as a collaborator rather than audit it case by case. What it cannot establish is whether a given output fits the patient in front of the clinician. That judgment needs transparency about the system's limits and, in atypical cases, local explanations to check whether the model's reasoning carries over to the individual. At this point explainability calibrates trust rather than creating it, alongside evidence that explanations can improve decisions and reduce diagnostic error [5,14,15]. Whether that calibrating role is better measured through collaboration outcomes, such as diagnostic accuracy or reduced over-reliance, than through trust collected after the explanation is a question current evaluation frameworks have not settled [16].

4.2. RQ2: What Explanation Needs Do HCPs Have Regarding AI System Recommendations, and How Does Agreement with the System's Recommendation Influence Those Needs?

The literature has proposed explanations partly as a way to communicate AI results to patients and colleagues [9,11], but our participants endorsed that use only weakly. Across the scenarios in Q18 the global difference was significant ($p = 0.040$) while no pairwise contrast survived correction, and the items about communicating results to patients or colleagues sat below those involving clinical uncertainty or high-risk decisions. The same pattern appeared in Q20: detailed explanations were wanted most for complex or atypical cases (73%) and for low or uncertain system performance (70%), both rated significantly above justifying a decision to other professionals (53%). Using an explanation to justify an AI output to others would hand the system a task that stays with the clinician, the interpretation of a result for a particular patient and its communication onward.

The demand for explanations did not depend on whether the clinician agreed with the system. Q21 and Q22 showed no significant differences (McNemar) in the information

requested under agreement and disagreement, with local explanations most frequently requested in both (61% and 54%). What shifted was the kind of explanation: disagreement drew more interest in counterfactual reasoning [17] (46% vs. 34%), which helps clarify why the system favored one decision over another. This runs against designs that suppress explanations once the clinician agrees, on the assumption that they only add noise in screening [18,19]. In disagreement an explanation can act as a cognitive-forcing step and curb over-reliance [20]; in agreement it can show whether the model reached its recommendation by the same route as the clinician, or surface something the clinician overlooked. That argues for keeping explanations available under agreement at least early in adoption, even if their value decreases as the tool becomes familiar [21]. Q23 supports reading these needs as clinically rather than relationally driven, since agreement with the system was the least influential factor in how much detail clinicians wanted, well below case urgency, diagnostic experience and knowledge of the system (all $p < 0.05$). What HCPs ask for is support for their own reasoning [7,20], not a substitute for it.

On the form explanations should take, the five types in Q17 were all rated highly (top-2-box above 60%) with no significant differences between them; example-based and local explanations led descriptively (75% and 73%) and global, importance-based explanations trailed (61%), in line with the value of case-specific information reported by [12]. These results characterize which explanation forms clinicians value rather than their effect on decisions; whether those preferences hold across specialties and translate into measurable gains in decision behavior remains to be established in situ. The characteristics clinicians wanted (Q19) are more telling. Grounding in medical evidence ranked significantly above every other property, with consistency with clinical protocols and technical precision next, and all of these above clarity, simplicity and speed of understanding (all $p < 0.05$). Methodological rigor matters to HCPs, but it is not what they ask for first: they want explanations anchored in clinical evidence that do not contradict established knowledge. This sits awkwardly with XAI guidance built around accessibility and ease of interpretation [10] and suggests that an over-simplified explanation can obstruct clinical reasoning rather than aid it, because it fails the evidential standard clinicians apply to any clinical source. The point is sharpened by [22]: workable human-AI collaboration depends on systems validated against clinical evidence, not on interpretability bolted on afterwards, which makes these properties a precondition for the use clinicians are willing to commit to rather than an optional refinement. Whether evidence-grounded explanations actually improve decisions over conventional technical ones, and whether that depends on specialty and on time pressure, is the test these declared preferences still have to pass.

5. Conclusions

This study offers empirical evidence for an HCXAI account in which justified trust rests mainly on institutional clinical validation — this is particularly relevant for decision-making tasks, where healthcare professionals tend to report lower confidence. Certain XAI techniques could contribute to that clinical validation by providing information about what the model attends to, while explanations delivered to HCPs do the secondary work of calibrating that trust at the point of use, so the clinician can fold a recommendation into their own reasoning. The distinction matters for evaluation: it moves the question from whether an explanation is understood to whether it improves the joint clinician-system decision. The supporting findings are that clinicians wanted explanations regardless of agreement with the system, that they placed clinical grounding ahead of simplicity against the grain of common XAI guidance [10], and that their need for detail tracked the difficulty of the case rather than their AI literacy [23]. Together these point toward HCXAI frameworks in

which explainability is part of how clinical decisions are made rather than a usability layer added on top.

6. Limitations

The sample was drawn entirely from Navarra, Spain, and was too small to stratify by clinical specialty, leaving open whether explanation preferences differ across professional roles. Its composition adds further gaps: the survey sample included no emergency-medicine professionals, which matters because the preference for clinical rigor over simplicity may not survive the time pressure of acute care, and only six participants reported exclusively non-clinical use, a thin basis for the claim that the results extend across use contexts and tool types. Among the clinicians who had used AI systems, self-assessed literacy was low to moderate, so the study says little about how more AI-expert users would respond. Recruitment by email probably favored professionals already interested in AI, which may have tilted the sample toward more receptive attitudes. Two methodological points limit what can be inferred from the results. The between-group comparisons returned no significant differences, but with 41 participants in total and subgroups of around twenty, these tests had little power, and a null result of this kind does not show that the groups are equivalent; the claim that the findings hold across subgroups is therefore provisional rather than established. The measures were also self-reported, drawn from a questionnaire developed for this study rather than a validated instrument and from a self-rated literacy scale, all of which are susceptible to social-desirability bias. Above all, the study recorded what clinicians say they need, not what they do, and whether these stated preferences survive real clinical use, and hold under time pressure and across specialties, is the question that should drive the next study.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ai7070247/s1>.

Author Contributions: Conceptualization, A.B. and M.G.; Methodology, M.C.-Y., A.B. and M.G.; Software, M.C.-Y.; Validation, M.C.-Y.; Formal analysis, M.C.-Y. and M.G.; Investigation, M.C.-Y. and M.G.; Data curation, M.C.-Y.; Writing—original draft, M.C.-Y.; Writing—review & editing, M.C.-Y., A.B. and M.G.; Visualization, M.C.-Y.; Supervision, M.G.; Project administration, M.G.; Funding acquisition, M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Government of Navarra has funded the doctoral researcher's contract within the framework of the 'Industrial Doctorates 2024' funding call. Grant number: 0011-1408-2024-000000.

Institutional Review Board Statement: Ethical review and approval were waived for this study. The data analyzed were obtained from healthcare professionals who voluntarily completed an anonymous survey on AI perceptions. No personal or identifiable information was collected, and the survey contained no questions regarding patients. Participation was entirely voluntary, and participants were informed of the scientific purpose of the study prior to completion.

Informed Consent Statement: Informed consent was obtained implicitly through voluntary completion of the survey. Prior to participation, all respondents received an information notice explaining that their responses would be used for scientific research purposes.

Data Availability Statement: The questionnaire used in this study is provided in the Supplementary Materials. The de-identified survey response data are available from the corresponding author upon request.

Acknowledgments: During the preparation of this manuscript, the author(s) used Claude (Version 5), for the purposes of Writing improvement and translation assistance. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HCP	Healthcare professionals
XAI	Explainable Artificial Intelligence
HCXAI	Human-Centered XAI
RQ	Research Questions
FDR	False Discovery Rate

References

1. Shamszare, H.; Choudhury, A. Clinicians' Perceptions of Artificial Intelligence: Focus on Workload, Risk, Trust, Clinical Decision Making, and Clinical Integration. *Healthcare* **2023**, *11*, 2308. [CrossRef] [PubMed]
2. Shevtsova, D.; Ahmed, A.; Boot, I.W.A.; Sanges, C.; Hudecek, M.; Jacobs, J.J.L.; Hort, S.; Vrijhoef, H.J.M. Trust in and Acceptance of Artificial Intelligence Applications in Medicine: Mixed Methods Study. *JMIR Hum. Factors* **2024**, *11*, e47031. [CrossRef] [PubMed]
3. Mamun, A.; Barua Soumma, S.; Ghasemzadeh, H. Trustworthy AI in digital health: A comprehensive review of robustness and explainability. *Prog. Biomed. Eng.* **2026**, *8*, 022007. [CrossRef] [PubMed]
4. European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). 2024. Available online: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng> (accessed on 21 June 2026).
5. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef] [PubMed]
6. Biecek, P.; Samek, W. Position: Explain to Question not to Justify. *arXiv* **2024**, arXiv:2402.13914. [CrossRef]
7. Herrera, F. Reflections and attentiveness on eXplainable Artificial Intelligence (XAI). The journey ahead from criticisms to human–AI collaboration. *Inf. Fusion* **2025**, *121*, 103133. [CrossRef]
8. Chen, F.; Zhou, J.; Holzinger, A.; Fleischmann, K.R.; Stumpf, S. Artificial Intelligence Ethics and Trust: From Principles to Practice. *IEEE Intell. Syst.* **2023**, *38*, 5–8. [CrossRef]
9. Hildt, E. What Is the Role of Explainability in Medical Artificial Intelligence? A Case-Based Approach. *Bioengineering* **2025**, *12*, 375. [CrossRef] [PubMed]
10. Wang, D.; Yang, Q.; Abdul, A.; Lim, B.Y. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–15. [CrossRef]
11. Tonekaboni, S.; Joshi, S.; McCradden, M.D.; Goldenberg, A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *arXiv* **2019**, arXiv:1905.05134. [CrossRef]
12. Barda, A.J.; Horvat, C.M.; Hochheiser, H. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 257. [CrossRef] [PubMed]
13. Van Leersum, C.M.; Maathuis, C. Human centred explainable AI decision-making in healthcare. *J. Responsible Technol.* **2025**, *21*, 100108. [CrossRef]
14. Salvi, M.; Seoni, S.; Campagner, A.; Gertych, A.; Acharya, U.R.; Molinari, F.; Cabitza, F. Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *Int. J. Med. Inform.* **2025**, *197*, 105846. [CrossRef] [PubMed]
15. Rahman, J.; Delaforce, A.; Bradford, D.; Li, J.; Magrabi, F.; Cook, D.; Brankovic, A. The Role of Explanations in AI-Generated Alerts: Qualitative Study of Clinical Views on Explainable AI in Predictive Tools. *JMIR Hum. Factors* **2026**, *13*, e81460. [CrossRef] [PubMed]
16. Bauer, J.M.; Michalowski, M. Human-centered explainability evaluation in clinical decision-making: A critical review of the literature. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 1477–1484. [CrossRef] [PubMed]
17. Mertes, S.; Huber, T.; Weitz, K.; Heimerl, A.; André, E. GANterfactual—Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning. *Front. Artif. Intell.* **2022**, *5*, 825565. [CrossRef] [PubMed]
18. Pinto, I.; Olazarán, Á.; Jurío, D.; De la Osa, B.; Sainz, M.; Oscoz, A.; Ballaz, J.; Gorricho, J.; Galar, M.; Andonegui, J. Improving diabetic retinopathy screening using artificial intelligence: Design, evaluation and before-and-after study of a custom development. *Front. Digit. Health* **2025**, *7*, 1547045. [CrossRef] [PubMed]

19. Sayres, R.; Taly, A.; Rahimy, E.; Blumer, K.; Coz, D.; Hammel, N.; Krause, J.; Narayanaswamy, A.; Rastegar, Z.; Wu, D.; et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology* **2019**, *126*, 552–564. [[CrossRef](#)] [[PubMed](#)]
20. Wysocki, O.; Davies, J.K.; Vigo, M.; Armstrong, A.C.; Landers, D.; Lee, R.; Freitas, A. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artif. Intell.* **2023**, *316*, 103839. [[CrossRef](#)] [[PubMed](#)]
21. Sivaraman, V.; Bukowski, L.A.; Levin, J.; Kahn, J.M.; Perer, A. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–18. [[CrossRef](#)]
22. Xiong, W.; Fan, H.; Ma, L.; Wang, C. Challenges of human—machine collaboration in risky decision-making. *Front. Eng. Manag.* **2022**, *9*, 89–103. [[CrossRef](#)]
23. Arbelaez Ossa, L.; Starke, G.; Lorenzini, G.; Vogt, J.E.; Shaw, D.M.; Elger, B.S. Re-focusing explainability in medicine. *Digit. Health* **2022**, *8*, 205520762210744. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.