machine learning & knowledge extraction

MDPI

Article

Painting the Black Box White: Experimental Findings from Applying XAI to an ECG Reading Setting

Federico Cabitza [1,2,*], Andrea Campagner [2], Chiara Natali [1], Enea Parimbelli [3,4], Luca Ronzio [5] and Matteo Cameli [6]

[1] Department of Computer Science, Systems and Communication, University of Milano-Bicocca, 20126 Milan, Italy
[2] IRCCS Istituto Ortopedico Galeazzi, 20161 Milan, Italy
[3] Department of Electric, Computer and Biomedical Engineering, University of Pavia, 27100 Pavia, Italy
[4] Telfer School of Management, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[5] Department of Medicine and Surgery, University of Milano-Bicocca, 20126 Milan, Italy
[6] Department of Medicine, Surgery and Neuroscience, University of Siena, 53100 Siena, Italy
* Correspondence: federico.cabitza@unimib.it; Tel.: +39-02-6448-7888

Abstract: The emergence of black-box, subsymbolic, and statistical AI systems has motivated a rapid increase in the interest regarding explainable AI (XAI), which encompasses both inherently explainable techniques, as well as approaches to make black-box AI systems explainable to human decision makers. Rather than always making black boxes transparent, these approaches are at risk of painting the black boxes white, thus failing to provide a level of transparency that would increase the system's usability and comprehensibility, or even at risk of generating new errors (i.e., white-box paradox). To address these usability-related issues, in this work we focus on the cognitive dimension of users' perception of explanations and XAI systems. We investigated these perceptions in light of their relationship with users' characteristics (e.g., expertise) through a questionnaire-based user study involved 44 cardiology residents and specialists in an AI-supported ECG reading task. Our results point to the relevance and correlation of the dimensions of trust, perceived quality of explanations, and tendency to defer the decision process to automation (i.e., technology dominance). This contribution calls for the evaluation of AI-based support systems from a human–AI interaction-oriented perspective, laying the ground for further investigation of XAI and its effects on decision making and user experience.

Keywords: explainable AI; decision support systems; ECG; artificial intelligence; XAI

Citation: Cabitza, F.; Campagner, A.; Natali, C.; Parimbelli, E.; Ronzio, L.; Cameli, M. Painting the Black Box White: Experimental Findings from Applying XAI to an ECG Reading Setting. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 269–286. https://doi.org/10.3390/make5010017

Academic Editors: Luca Longo and Andreas Holzinger

Received: 20 December 2022
Revised: 13 February 2023
Accepted: 1 March 2023
Published: 8 March 2023

1. Introduction

We are witnessing a continuous and, indeed, accelerating move from decision support systems that are based on explicit rules conceived by domain experts (so-called expert systems or knowledge-based systems) to systems with behaviors that can be traced back to a potentially huge number of rules that have been automatically learned on the basis of correlative and statistical analyses of large quantities of data. This is the shift from symbolic AI systems to subsymbolic ones, which has made the black-box nature of these latter systems an object of a lively and widespread debate in both technological and philosophical contexts [1,2]. The main assumption motivating this debate is that making subsymbolic systems explainable to human decision makers makes them better and more acceptable tools and supports.

This assumption is widely accepted [3–5], although there are a few scattered voices against it (see e.g., [6–9]). For instance, explanations were found to increase complacency toward the machine advice [10], increase (or not reduce) automation bias [11–14], and groundlessly increase confidence in one's own decision [15,16]. Understanding or participating in this debate, which characterizes the scientific community that recognizes itself in the expression "explainable AI" and in the acronym "XAI", is difficult for the seemingly disarming heterogeneity of definitions of explanation, and the variety of characteristics that are associated with "good explanations" [17], or of the systems that generate them [18].

In what follows, we adopt the simplifying approach recently proposed in [18], where explanation is defined as the metaoutput (that is, an output that describes, enriches, or complements another main output) of an XAI system. From this perspective, good explanations are those that make the XAI system more usable, and therefore a useful support. The reference to usability suggests that we can assess explanations (and explainability) on different levels, by addressing complementary questions, such as Do explanations make the sociotechnical, decision-making setting more effective [19], in that they help decision makers commit fewer errors? Do they make it more efficient, by making decisions easier and faster, or just by requiring fewer resources? Last but not least, Do they make users more satisfied with the advice received, possibly because they have understood it more, and thus made them more confident about their final say?

Although some studies [13,20] have already considered the psychometric dimension of user satisfaction (see, e.g., the concept of causability [21], related to the role of explanations in making advice more understandable from a causal point of view), here we would like to focus on effectiveness (i.e., accuracy) and other cognitive dimensions (than understandability), both regarding the support (e.g., trust and utility) and the explanations received. In fact, explanations can be either clear or ambiguous (cf. comprehensibility), either tautological and placebic [22] or instructive (cf. informativeness), either pertinent or off-topic (cf. pertinence), and, as obvious as it may seem, either correct or incorrect, as any AI output can be. Therefore, otherwise good explanations (that is persuasive, reassuring, comprehensible, etc.) could even mislead their target users. This is the so-called white-box paradox, which we have already begun investigating in previous empirical studies [23,24]. Thus, investigating if and how much users find explanations "good" [25] (and in the next section we will make this term operationally clear) can be related to focusing on the possible determinants of machine influence (i.e., also called dominance), automation bias, and other negative effects related to the output of decision support systems on decision performance and practices.

In what follows, we will present the findings from a user study conceived to investigate the role of AI advice and textual explanations in the medical task of ECG reading and interpretation. In particular, we chose the use case of ECG reading because the (semi)automatic interpretation and classification of biosignals has been gaining traction recently. Indeed, a number of studies have reported the successful application of AI systems in accurately detecting abnormalities and diagnosing cardiovascular diseases [26]. Notable examples include the use of highly accurate deep learning algorithms [27,28], which in some instances achieved performances that are comparable or superior to expert ECG readers. At the same time, the accuracy of such high-performing AI systems is often coupled with a requirement for explainability [29], with the double aim of fostering trust in the AI system and confidence in the final human (AI-supported) decision, as well as to enrich the AI predictions with further information (i.e., the "why" behind such predictions). The XAI techniques applied to biosignals, and more in general time-series data [30], are various in the nature of their output, including rule-based [31], semantic data [32] and feature importance-based [33] explanations. In the following, we will adopt text-based explanations as our baseline approach for a survey experiment involving human ECG readers, in order to assess the impact of AI advice coupled with textual explanations on the task of ECG reading aimed at diagnosis. Finally, we remark that this work must be interpreted within the broader context of a study in which an international and multidisciplinary team

(including data and computer scientists and medical doctors) explores the role of XAI explanations in AI-supported ECG reading. In a previous study, we investigated the role of visual explanations (in terms of saliency maps, see [34]), whereas here we focus on textual explanations.

2. Methods

To investigate how human decision makers perceive explanations, we designed and conducted a questionnaire-based experiment in which we involved 44 cardiologists of varying expertise and competence (namely, 25 residents and 19 specialists) from the Medicine School of the University Hospital of Siena (Italy), in an AI-supported ECG reading task, not connected to their daily care practice. The readers were invited to classify and annotate 20 ECG cases, previously selected by a cardiologist from a random set of cases extracted from the ECG Wave–Maven repository (https://ecg.bidmc.harvard.edu/maven/mavenmain.asp, accessed 5 March 2023) on the basis of their complexity (recorded in the above repository), so as to have a balanced dataset in terms of case type and difficulty. The study participants had to provide their diagnoses both with and without the support of a simulated AI system, according to an asynchronous Wizard of Oz protocol [35]. The support of the AI system included both a proposed diagnosis and a textual explanation to back the former one. We give two examples here of explanations associated with an ECG identified as normal, and a case associated with a pneumothorax condition (a condition of great significance and severity), respectively. The normal ECG includes a heart rate of 60 beats per minute, normal axis (+53°) and intervals within normal limits, normal P wave, QRS complex and repolarization apparently physiological. The pneumothorax includes low voltages in lateral leads, vertical cardiac axis, the presence of nonsignificant anterior and inferior ST-segment elevation. The experiment was performed by means of a Web-based questionnaire set up through the LimeSurvey platform (version 3.23), to which the readers had been individually invited by personal email (see Figure 1).

The ECG readers were randomly divided into two groups, which were equivalent for expertise and were supposed to interact with the AI system differently (see Figure 2). In doing so, we could comparatively evaluate potential differences between a human-first and an AI-first configuration. In both groups, the first question of the questionnaire asked the readers to self-assess their trust in AI-based diagnostic support systems for ECG reading. The same question was also repeated at the end of the questionnaire to evaluate potential differences in trust caused by the interaction with the AI system.

For each ECG case, the readers in the human-first group were first shown the trace of the ECG together with a brief case description, and then they had to provide an initial diagnosis (in free text format). After that this diagnosis had been recorded, these respondents were then shown the diagnosis proposed by the AI; after having considered this latter advice, the respondents could revise their initial diagnosis; then they were shown the textual explanation (motivating the AI advice) and asked to provide their final diagnosis in light of this additional information. In contrast, the participants enrolled in the AI-first group were shown the AI-proposed diagnosis together with the ECG trace and case description; only afterward, they were asked to provide their own diagnosis in light of the received advice. Finally, ECG readers were shown the textual explanation, and asked whether they wanted to revise their initial diagnosis or confirm it.

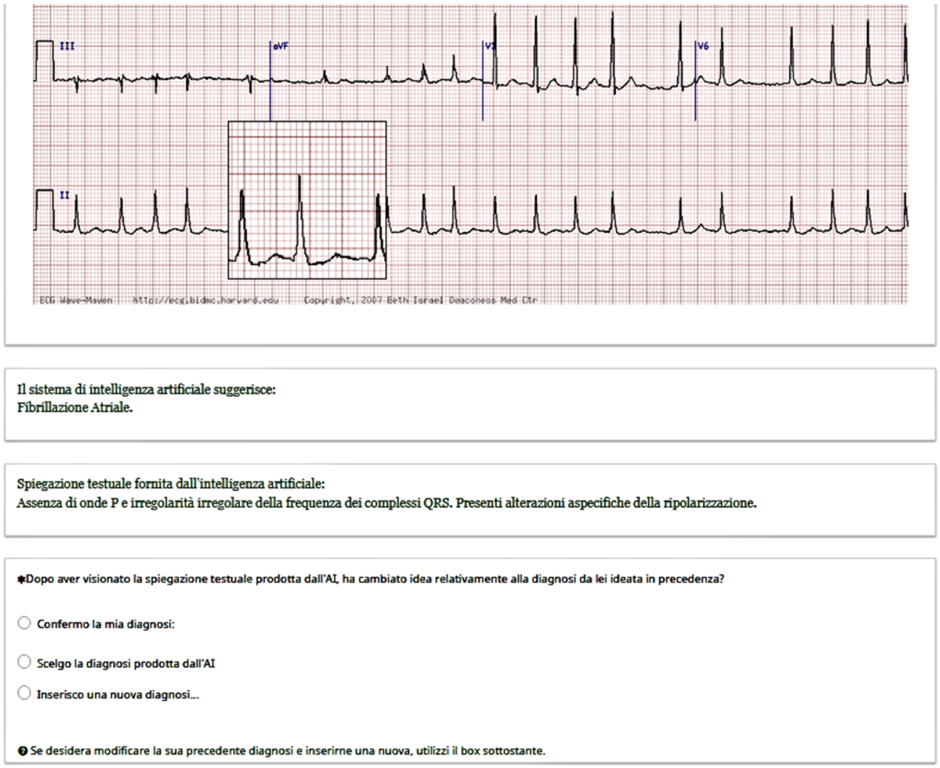

Figure 1. Screenshot taken from one of the pages of the online questionnaire was used in the user study. The top of the image shows part of the ECG shown to participants for a clinical case and, particularly on the left, a magnified portion of the trace from the "magnifying glass" feature that could be activated by simply hovering the pointer over the image. Visible below is the advice given by the AI (in this case, "atrial fibrillation") and in the area below the explanation provided (in this case, "absence of P waves and irregularities in the frequency of the QRS complexes"). At the bottom, one can see the question item by which the user could confirm the diagnosis given previously (on the previous page), or choose the diagnosis provided by the AI, or enter another diagnosis (different from the one entered previously).

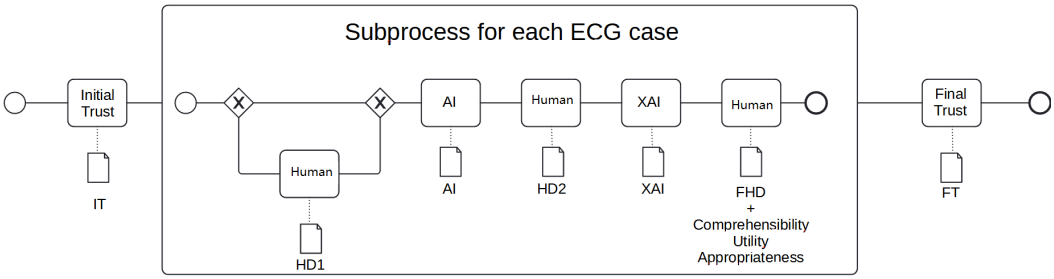

Figure 2. BPMN representation of the study design. Information collected is represented as data objects, coming from collection tasks. Its name is denoted by the name of the main actor. After the initial collection of the perceived "trust in AI" (Initial Trust, IT), the subprocess is repeated for each ECG case, where HD1, AI, HD2, XAI, and FHD items are collected, together with comprehensibility, appropriateness, and utility; these acronyms denote (see the abbreviations list at the end of the article) the first diagnosis humans provide (HD1) before receiving the AI advice (AI), the diagnosis recorded immediately after receiving this AI advice (HD2), the information regarding the XAI support (XAI), and the final and definitive diagnosis (FHD). Finally, a posttest "trust in AI" is collected again (Final Trust, FT).

For each textual explanation, we asked the participants to rate its quality in terms of its comprehensibility, appropriateness, and utility, in this order (so as to reflect a natural sequence through perception, interpretation, and action). In particular, while comprehensibility and utility were considered self-explanatory terms, we pointed out in a written comment that appropriateness, to our research aims, would combine the respondents' perception of pertinence and correctness together. That is, that dimension would reflect the extent to which "the explanation had something to do with the given advice" and, with regard to the latter, "it was plausible and correct." In other words, we asked the participants to judge the quality of the explanation with respect to the advice and to the case at hand by means of two different constructs - appropriateness and utility, respectively.

The accuracy of the simulated AI—that is the proportion of correct diagnostic advice—was 70% with respect to the ECG Wave–Maven gold standard. This rate was considered appropriate because in a previous study [36] we observed a slightly lower average accuracy in a similar population of readers.

To avoid negative priming and ordering bias [37,38], and hence avoid fostering unnecessary distrust in the AI, the first five cases of the questionnaire shown to the participants were all associated with a correct diagnosis and a correct explanation from the XAI support. Although the participants had been told that the explanations were automatically generated by the AI system, like the diagnostic advice, these had been prepared by the same cardiologist who also selected the cases. In particular, 40% of the explanations were incorrect or not completely pertinent to the cases. More precisely, for the five cases classified as simple, all explanations were correct, for the nine cases of medium complexity, four explanations were wrong, and for the remaining six cases denoted as difficult, four explanations were wrong.

Based on the collected data, we then considered the following research questions.

(RQ1): Does the readers' expertise have any effect in terms of either basal trust, difference in trust by the readers, or final trust (RQ1a)? Does the interaction protocol (human-first vs. AI-first) have any effect on differences in trust by the readers, or final trust (RQ1b)?

(RQ2): Is there any difference or correlation between the three investigated psychometric dimensions (i.e., comprehensibility, appropriateness, utility)? A positive answer to this latter question would justify the use of a latent quality construct (defined as the average of the psychometric dimensions) to simplify the treatment of other research questions.

(RQ3): Does the readers' expertise, diagnostic ability (i.e., accuracy), and the adopted interaction protocol have any effect in terms of differences in perceived explanations' quality? Similarly, does the perceived quality of explanations correlate with the basal or final trust? Regarding diagnostic ability, we stratified readers based on whether their baseline accuracy (cf. HD1, see Figure 2) was either higher or lower than the median.

(RQ4): Is there any correlation between the explanations' perceived quality and the readers' susceptibility to technology dominance [39,40]? Although technological dominance is a multifactorial concept, for our practical aims, we express it in terms of the rate of decision change due to exposition to the output of an AI system. Moreover, we distinguished between positive dominance, when changes occur from initial wrong decisions (e.g., diagnoses) to eventually correct ones and negative dominance, for the dual case, when the AI support misleads decision makers.

(RQ5): Finally, does the correctness of the explanation (and of the associated classification) make any difference in terms of either perceived explanations' quality or influence (i.e., dominance)?

The abovementioned research questions were evaluated through a statistical hypothesis testing approach. In particular, correlations were evaluated by means of Spearman ρ (and associated p-values), so as to properly take into account for monotone (and not necessarily linear) relationships between ordinal variables and continuous ones. In regard to research questions 1 and 3, on the other hand, paired comparisons were performed by applying the Wilcoxon signed-rank test, whereas unpaired comparisons were performed by applying the Mann–Whitney U test. In both cases, effect sizes were evaluated through the rank biserial correlation (RBC). In all cases, to control the false discovery rate due to multiple hypothesis testing, we adjusted the observed p-values by using the Benjamini–Hochberg procedure. Significance was evaluated at the 95% confidence level. All statistical analysis and data processing was implemented by using the Python language (v. 3.10.6) and the libraries scipy (v. 1.9.1), pingouin (v. 0.5.2), pandas (v. 1.4.4), and scikit-posthocs (v. 0.7.0), as well as the libraries matplotlib (v. 3.5.2) and seaborn (v. 0.11.2) for generating the data visualizations.

3. Results and Discussion

After having closed the survey, we collected a total of 1352 responses from the 44 ECG readers involved, of which 21 had been enrolled in the human-first protocol and the remaining 23 in the AI-first protocol.

3.1. RQ1—Effect of Expertise and Interaction Protocol on Trust

In regard to RQ1a ("Does the readers' expertise have any effect in terms of either basal trust, difference in trust by the readers, or final trust?"), the results concerning the differences in self-perceived trust are reported in Figure 3, stratified by expertise (on the left) and interaction protocol (on the right). The difference between initial and final trust was significant for novice readers (adjusted p, 0.004; RBC, 0.92), but not for the expert ones (adjusted p, 0.407; RBC, 0.35). Furthermore, even though the difference in initial trust between novice and expert readers was not significant (adjusted p, 0.439; RBC, 0.15), the difference in final trust was instead significant (adjusted p, 0.009; RBC, 0.54), with the novice readers reporting, on average, a higher final trust than the expert ones. This is in line with previous studies in the field of human–AI interaction [23,24,41,42], which showed how novice readers are more willing to accept the support of an AI-based system, and better appreciate its output. An explanation for this widely reported observation can be traced back to the literature in the Theory of Technological Dominance (TTD) [39], in which a previous finding from Noga and Arnold [43] identified user expertise as one of the main determinants of dominance and reliance, of which trust is a determinant. Although a tenet of TTD holds that decision aids are especially beneficial to professionals thanks to a bias mitigation effect [44], the study by Jensen et al. [45] displayed a diverging beneficial effect of decision support, with novices benefiting more in comparison to experts who, in turn, often discounted the aid's support. This is in line with our findings, which point to more experienced decision makers possibly being less favorably impacted by such systems, possibly due to a lower level of familiarity or a higher prejudice against the machine (see also [46]).

As for RQ1b ("Does the interaction protocol, human-first vs. AI-first, have any effect on differences in trust by the readers, or final trust?"), we point out that we did not observe any significant effect due to the interaction protocol on initial trust. This result does not come unexpectedly, and was in fact desirable. Indeed, readers were randomly assigned to one of the two considered protocols. Interestingly, the interaction protocol seemingly did not have an effect on final trust either, despite the fact that the adoption of different interaction protocols may have a significant effect in terms of the overall accuracy of the hybrid human–AI team [47]. In fact, differences between the two cohorts were not significant, neither in terms of initial trust (adjusted p, 0.090; RBC, 0.33) nor in terms of final trust (adjusted p, 0.787; RBC, 0.05), although in the first case the effect size was medium. In this sense, it appears that even though the order of the AI intervention was certainly

impactful in regard to the overall diagnostic performance (this is an effect that is not entirely trivial and calls for further study to understand its causes), it might be perceived by the user as a secondary element compared to other trust-inducing or trust-hindering factors. Nonetheless, we found that both interaction protocols had a large effect on trust difference. In particular, human-first protocols led to a significant increase between the initial and final trust (adjusted p: 0.016, RBC: 0.81), but not so for the AI-first cohort (adjusted p, 0.078; RBC, 0.60), even though both effect sizes were large. A possible explanation for this difference can be found in the accuracy level of the AI system, which in this user study was 70%, i.e., well over the average accuracy of the readers and thus likely to lead to a positive interaction for the study participants and hence an increased sense of trust in the decision support. An accuracy rate of 70%, far from being infallible, was nonetheless effective in increasing trust. This is in line with [48], according to which AI systems should be optimized for team performance rather than accuracy per se, going far as affirming that "predictable performance may be worth a slight sacrifice in AI accuracy" whereas the most accurate AI "may not lead to highest team performance".

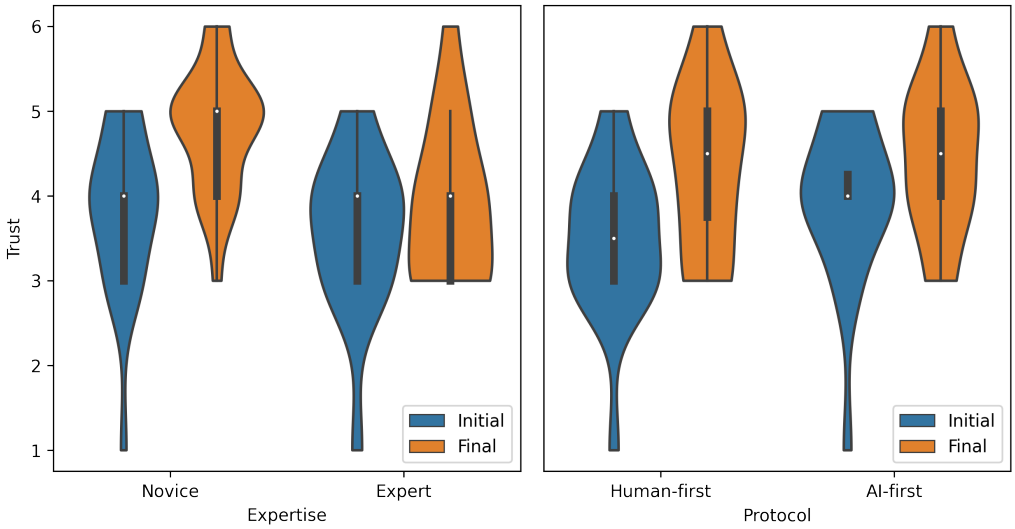

Figure 3. Violin plots of the distributions of the initial (pretest, IT in Figure 2) and final (posttest, FT in Figure 2) trust scores reported by the study participants, stratified by (**left**) readers' expertise (novices vs. experts), and (**right**) interaction protocol (human-first vs. AI-first).

3.2. RQ2—Correlation between the Psychometric Dimensions

Concerning RQ2 ("Is there any difference or correlation between the three investigated psychometric dimensions?"), we briefly note that, as reported in Figure 4, the three psychometric dimensions that we investigated (i.e., comprehensibility, appropriateness, utility) were indeed strongly correlated between each other (appropriateness vs. comprehensibility, ρ, 0.86; appropriateness vs. utility, ρ, 0.82; comprehensibility vs. utility, ρ, 0.80), and all of the correlations were significant (adjusted p-values < 0.001). This result, although interesting, is not totally unexpected because, intuitively, an appropriate and comprehensible explanation is likely to be found also useful. Conversely, for an explanation to be useful it should be at least also be comprehensible. Notably, the observed value of the Cronbach α (which measures the internal consistency of the questionnaire, in regard to the three psychometric items) was 0.93, which is higher than Nunnally's reliability threshold for applied studies (i.e., 0.8; see [49]). Thus, the internal consistency of our test was sufficiently high to guarantee its reliability, but not so much as to suggest redundancy and hence undermine its validity [50]. In particular, we believe that these results justify the aggregation of the three psychometric dimensions into a latent quality construct (defined as the average between appropriateness, comprehensibility, and utility), which was then considered in the statistical analysis.

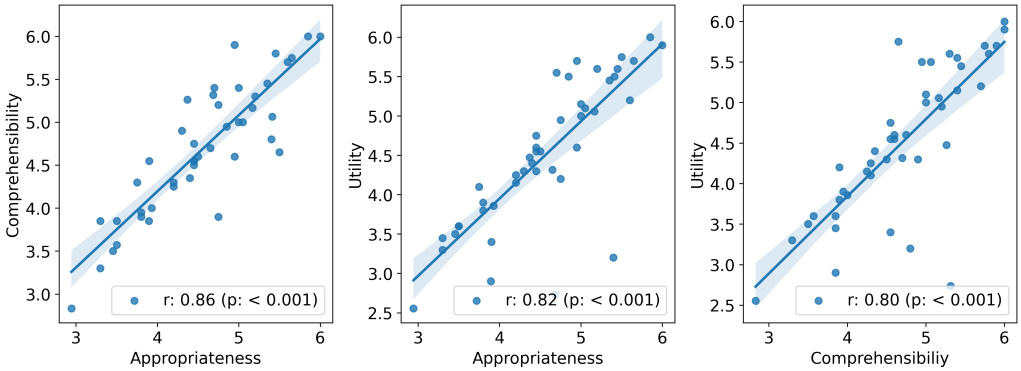

Figure 4. Scatter plots of the correlations observed between the appropriateness, comprehensibility, and utility scores associated with each explanation provided by the XAI module of the AI support.

3.3. RQ3—Perceived Explanations' Quality with Respect to Expertise, Accuracy, Interaction Protocol, and Possible Correlation with Basal or Final Trust

The aforementioned perceived quality level was analyzed in RQ3 first with respect to user expertise, their accuracy, and the interaction protocol ("Do the readers' expertise, their diagnostic ability, as well as the adopted interaction protocol, have any effect in terms of differences in perceived explanations' quality?"). The results are reported below in Figure 5.

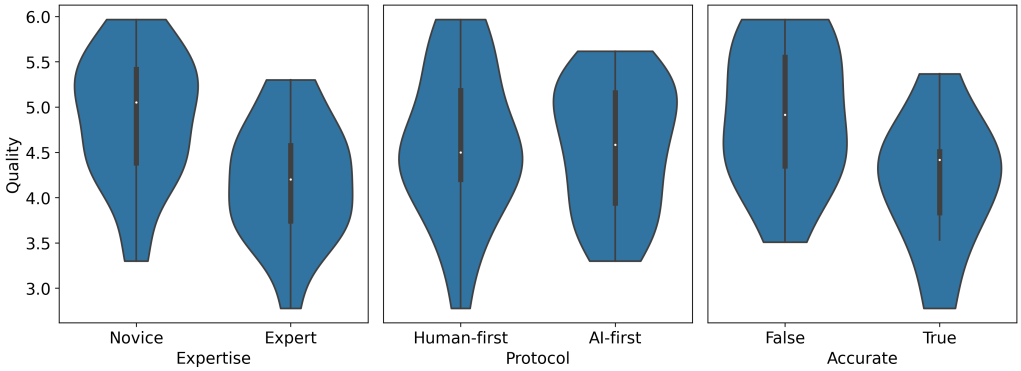

Figure 5. Violin plots of the explanations' quality, stratified by (**left**) readers' expertise, (**center**) interaction protocol, and (**right**) readers' baseline accuracy.

As expected, the difference in explanations' quality between human-first and AI-first interaction protocols (adjusted p, 0.981; RBC, 0.01) was not significant and was associated with only a negligible effect. Indeed, as mentioned previously in regard to trust, readers were assigned randomly to the two interaction protocols which, aside from when the AI advice was shown, were essentially equivalent in terms of the given explanations. By contrast, even though the difference in explanations' quality with respect to readers' baseline accuracy was similarly nonsignificant (adjusted p, 0.155), the relationship between the two variables was associated with a medium-to-large effect size (0.36). Furthermore, the difference in explanations' quality between novice and expert readers (adjusted p: 0.012, RBC: 0.51) was associated with a large effect size and was also statistically significant. These results highlight how readers' proficiency in the ECG reading task (as measured by either self-reported expertise or, more quantitatively, by basal accuracy) might have a significant effect on the perception of explanatory advice. A possible explanation for this observation, which was already mentioned above in reference to trust and previously observed in the literature [51], might be related to an increased acquaintance with AI and XAI systems for the less expert readers (who were also less accurate). Furthermore, less expert readers (e.g., the students and residents) might have found the explanations' quality higher due to their perceived usefulness in helping them identify characteristics of interest in an ECG

that they were not able to interpret alone. More experienced or more accurate readers, who by definition were more well-versed in the interpretation of ECG diagrams, might have missed this novelty element of explanations, which might have led to a lower, on average, perceived quality. Importantly, it appears that the users' initial attitude had little influence on the perceived quality of explanations, as investigated in the second part of the research question ("Does the perceived quality of explanations correlate with the basal or final trust?"). As reported in Figure 6, the explanations' quality was weakly correlated with the readers' basal trust in AI-based support systems (Spearman ρ: 0.27, adjusted p: 0.086), but significantly and strongly correlated with final trust (Spearman ρ: 0.71, adjusted $p < 0.001$). We conjecture this to be a consequence of the fact that the explanations were evaluated for their intrinsic value, rather than for a halo effect (due to trust) [52].

As for final trust, by contrast, the observed improvement is likely due to a simple reason: the support was perceived as a worthy addition to the decision-making process. Due to how the user experiment was constructed, we cannot decouple the contribution of XAI from plain AI. However, given that the accuracy of the decision support was higher than that of the readers but overall not particularly high (equaling 70%), we conjecture that the explanations themselves were an important contributing factor to the trust increase, for its novelty and appropriability [53] when compared to a simple categorical advice. Because the increase in final trust is highly dependent, and significantly so, on explanation quality, this finding reinforces the idea that explanations do influence trust. We believe that this relevant finding also provides an alternative and complementary explanation of the observed effect of readers' expertise on final trust and trust difference. Indeed, this latter effect could be explained as arising from the fact that less experienced readers rated more favorably the quality of explanations than the more expert readers. In light of the strong relationship between explanations' quality and final trust, this might explain why we observed a larger increase in trust for novices than for experts.

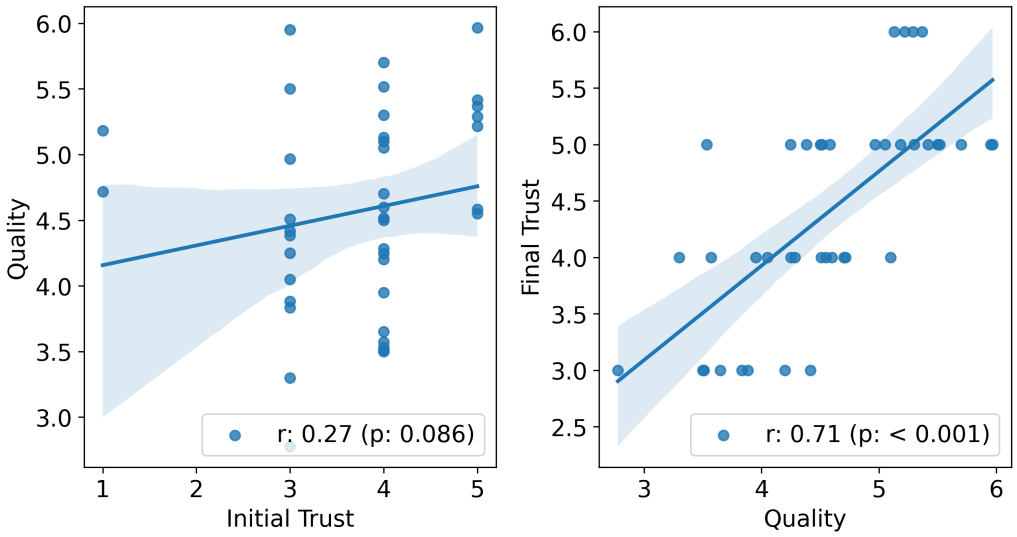

Figure 6. Scatter plots of the correlations observed between the scores of explanations' perceived quality and (**left**) initial (pretest) trust, (**right**) final (posttest) trust.

3.4. RQ4—Correlation between Perceived Quality of Explanation and Technology Dominance

The two final research questions—RQ4 and RQ5—further delve into the effect of explanations, by adopting the lens of the theory of technology dominance [39,40] through which we investigated possible correlations between the readers' susceptibility to this phenomenon (which, in this article, was operationalized through the rate of decision change due to exposition to the output of an AI system. More precisely, we considered the number of decision changes between HD2 and FHD, as we were interested in the

differential dominance induced by explanations) and explanations' perceived quality and actual correctness.

The correlation between dominance (distinguishing between positive and negative dominance) and the perceived explanations' quality is reported in Figure 7. Quality was moderate to strong and significantly correlated with dominance (Spearman ρ, 0.57; adjusted p, 0.007) and also with the positive component of dominance (Spearman ρ, 0.52; adjusted p, 0.025); by contrast, it was only moderately correlated with negative dominance (Spearman ρ, 0.39; adjusted p, 0.077). Furthermore, this latter correlation was not significant.

As clearly shown above in Figure 7, explanation quality strongly influences dominance, and especially positive dominance, and also significantly so.

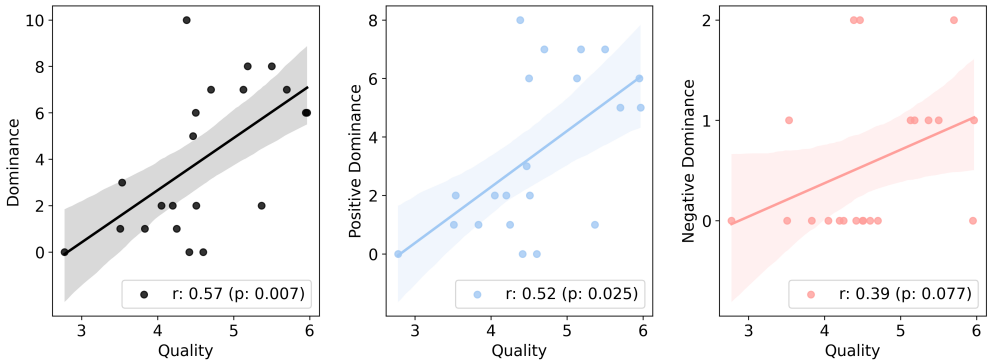

Figure 7. Scatter plots of the correlations between dominance (both positive and negative) and explanations' quality. Dominance is defined as the number of decision changes occurring between HD2 and FHD.

3.5. RQ5—Relationship between the Correctness of Explanations and Dominance

In Figure 8, we introduced the relationship between the the correctness of the associated classifications and their dominance, as well as the effect the perceived quality of these latter, and their dominance, which aid us in answering our fifth research question ("Does the correctness of the explanation make any difference in terms of either perceived explanations' quality or dominance?").

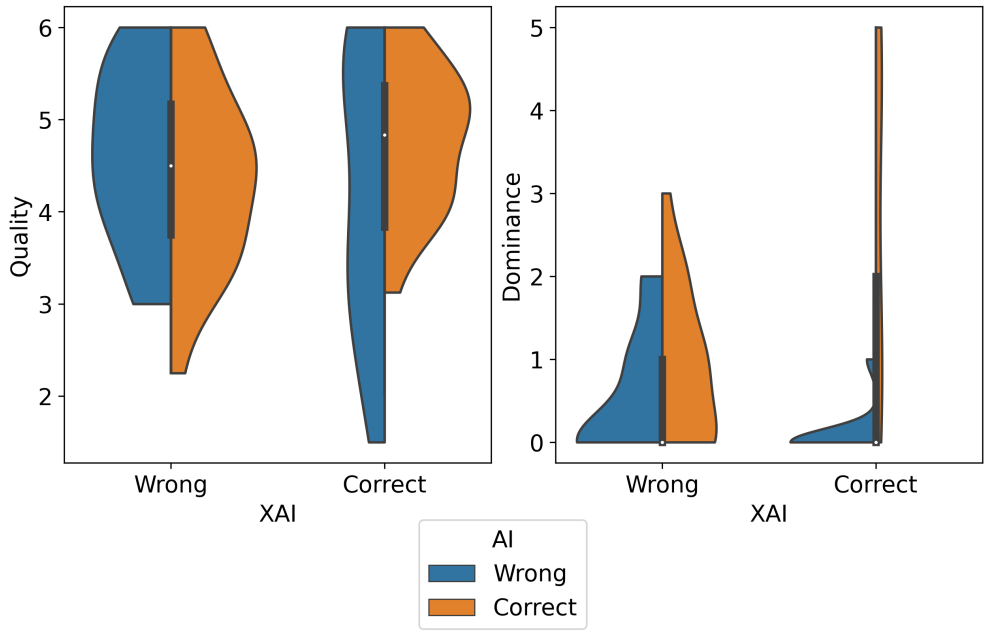

Figure 8. Violin plots showing the effect of the classifications' and explanations' correctness on (**left**) perceived explanations' quality and (**right**) dominance.

Both the average perceived quality and dominance increased when the classification or explanations were correct, as compared to when these latter ones were wrong. The differences in terms of quality were not significant (adjusted p-values, AI, 0.791; XAI, 0.791), and the associated effect sizes were small (RBC, AI, 0.03; XAI, 0.06). Similarly, the difference on dominance due to the correctness was not significant (p, 0.791) and associated with a small effect size (0.07). By contrast, the difference in dominance due to the correctness of the classification was significant (p, < 0.001) and associated with a medium-to-large effect size (RBC, 0.52). For this reason, we further investigated the combined effect of the explanation and classification on dominance. The four cases were significantly different in terms of dominance (Kruskal test, p, < 0.001). The results of the post-hoc comparison between the combination (in terms of adjusted p-values and effect sizes) is represented in Figure 9. All effect size were medium-to-large or large, whereas only the combination of a correct explanations with a correct classification reported a significant increase in dominance (as compared with the other possible combinations).

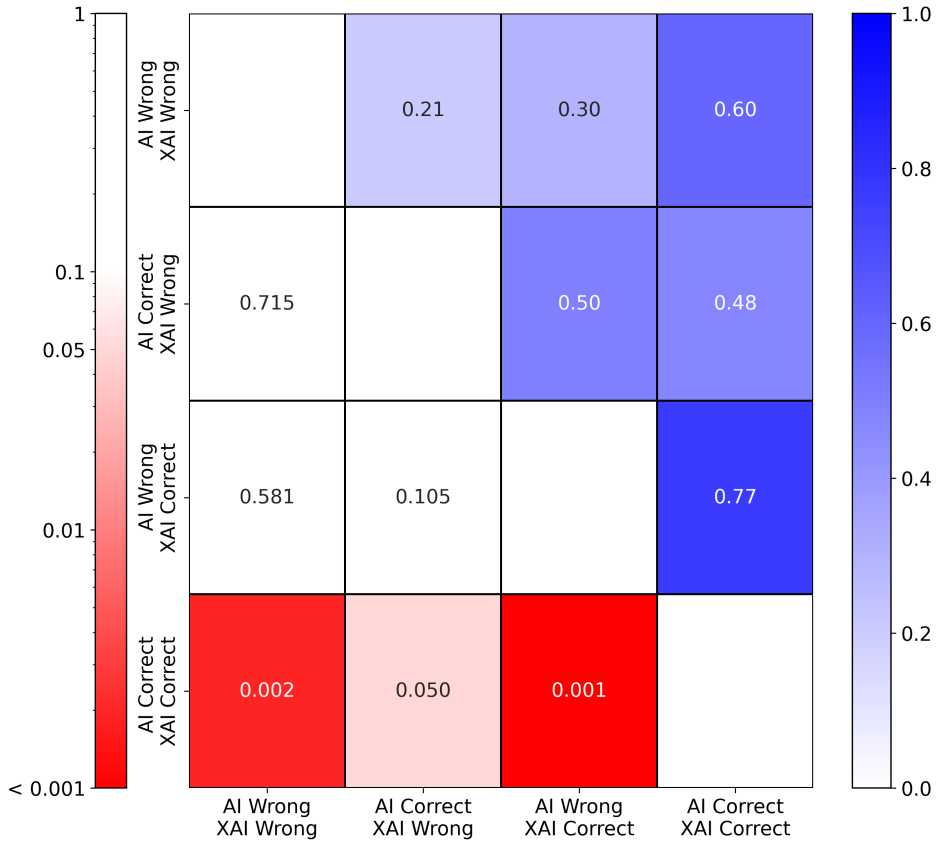

Figure 9. Matrix of the pairwise comparisons of the effects of the classifications' and explanations' correctness on dominance. Cells under the diagonal denote the p-value (Nemenyi post-hoc test) for the given pair of configurations, whereas cells above the diagonal report the corresponding effect size (RBC). Brighter shades of red (resp. blue) denote significance (resp. strength of the effect).

We believe this finding to be of particular interest because it confirms that high-quality explanations, especially when associated with correct classifications, can increase the persuasion potential of an XAI system—and especially so for the good (see, in particular, the rightmost panel in Figure 8 and the significance results in Figure 9). Nonetheless, as can be noticed from the rightmost panel in Figure 7, perceived quality can influence the users for the worse, as highlighted by the fact that explanations' quality was moderately associated with negative dominance (hence, opinion changes from a correct to an incorrect diagnosis), and especially so when a misleading explanations was associated with an otherwise correct classification (see Figure 8).

As observed in [54], which discusses the results of a user study on ML recommendations for clinician treatment selection, explainability may lead to lower accuracy in case of incorrect recommendations compared to black-box systems, with the introduction of explanations proving ineffective or insufficient in preventing dominance. A possible explanation for this effect can be traced back to the imperfect ability of the readers' in discriminating a correct explanation from a wrong one in terms of perceived quality (see the left panel in Figure 8). In turn, such an effect could motivate the emergence of biases and cognitive effects associated with automation, and especially so those that are directly related to the role of XAI, as in the case of the white-box paradox [23,24]. Despite the relevance of these results, we remark, however, that further research should address whether this correlation also holds in the case of placebic information (i.e., not semantically sensible, nor structurally consistent, but still perceived as valuable) as described by Langer [22], leveraging this concept from the psychological sphere of interpersonal communication to that of human–computer interaction.

3.6. Limitations and Further Research

Concluding our discussion, we remark that this study is exploratory; hence, its main limitations regard the relatively small sample of cases considered, if not of readers involved. In fact, in regard to participation, this study can leverage the perceptions and opinions of tens of cardiologists of different competencies and expertise. However, the study regards a serious game where the doctors involved knew no harm could be caused to real patients. More generally, our study being focused on a specific task (i.e., ECG reading), we do not have ambition on generalizability of the hard conclusions beyond the setting at hand. Nonetheless, we are confident that our general findings (i.e., the correlations between explanation quality and confidence and trust) can inspire further research in other settings, both in medicine and other domains. In this sense, three main possible areas where further research could extend our results regard the stratification by explanation types, the analysis of the impact of explanations on the readers' confidence, and the generalization of our findings to other settings. On the one hand, explanations should be distinguished according to a reference taxonomy, for instance, those recently proposed in [18,55–57] to see if different types of explanations can have different effects on decision making. We recall that in this study we focused on textual explanations of a justificatory and causal kind [18]. Moreover, explanations can be wrong in different ways. For instance, an explanation can be wrong because it does not regard (or is badly fitted to) either the case at hand or the machine's advice, or because it expresses a wrong way of reasoning. This macrodistinction reflects the typology proposed in [58], in which lapses and mistakes, respectively, regard errors in perception or attention, and the latter ones regard errors in reasoning and the application of domain knowledge. The explanations provided in the study presented in this paper were of various kinds, depending on the case at hand and the ECG to read, including both correct and incorrect explanations. Our explanations were not really produced by an XAI system, but rather by the human expert selecting the cases to present in the survey, various types of errors can affect both the original black-box system, as well as the XAI system producing the explanations (e.g., affecting their fidelity, stability, or other relevant properties as highlighted in [59] for a set of popular model-agnostic post-hoc explainability methods). On the other hand, as mentioned above, we did not collect confidence scores at each human decision step (i.e., HD1, HD2, and FHD, see Figure 2). For this reason, we cannot address the research question of whether explanations would improve confidence in the decision reported or not. Related to this research question, the debate around XAI for medical applications [54,60] has highlighted how explanations coupled with XAI advice can have positive (e.g., improved confidence in the final decision) as well as negative (e.g., persuade decision makers to trust bad AI-generated advice) impacts, and explainability is often highly sought after in AI applications to the medical domain because of its early examples of "expert systems" [61], which were not even based on black-box machine learning algorithms. For these reasons, further research should be aimed at investigating the

confidence construct, and its relationship with perceived user experience and satisfaction and ultimately the increase in the adoption of AI systems that are associated with promises to make healthcare more sustainable and efficient in virtue of their capability to detect otherwise-hidden diagnostic features in biosignals, as in the case of [62]. In this case, an AI reading low-cost and easily acquired ECGs was found to accurately predict subtypes of long QT syndromes that would usually require costly and time-consuming genetic testing to be diagnosed. Finally, even though our findings shed some light on the potential utility of explainable system, we note that these only refer to a specific setting. Future research should be devoted to the evaluation of the generalizability of our findings to other settings and domains.

4. Conclusions

The current interest in XAI is ostensibly and programmatically motivated by the need to make AI systems more transparent, understandable, and thus usable. However, in light of some empirically grounded findings, presented, among others, in the literature on naturalistic decision making [63,64], this interest appears to be more instrumental to the rising prevalence and diffusion of automated decision making (ADM) systems, especially when their use is anticipated in contexts for which the main legislative frameworks (e.g., the EU GDPR) require these systems to also provide reasons for their output whenever this latter can have legal effects (see, e.g., the debate around ADM and human-in-the-loop decision making in the risk assessment domain [19,65,66], which uses ADM. This addresses a requirement for justification, rather than explanation, although these two concepts are often conflated (for a line of reasoning on the difference between explanation and justification, see [18]).

In this light, it is important to note that, metaphorically speaking, providing AI with explainability, that is the capability to properly explain its own output, is more akin to painting the black box of inscrutable algorithms (such as deep learning or ensemble models) white, rather than making them transparent. What we mean with this metaphoric statement is that XAI explanations do not necessarily explain (as by definition or ontological status) but rather describe the main output of systems aimed at supporting (or making) decisions. This is why we described XAI explanations as a metaoutput. As such, explanations can fail to make the output they relate to more comprehensible, or its reasons explicit. They can even be wrong. For example, a previous study [10] noted how users who interacted with a transparent model, while seemingly being more able to simulate its behaviour, still failed to align their predictions to those of the model when it would be beneficial to do so or also to detect and correct even the model's most remarkable errors.

This trivial observation is seldom emphasized, and it should motivate researchers to embrace a wider idea of explainability. In this light, an explainable system would not be a system that merely gives explanations, whether in the form of either statements or images. It would rather be a system that provides users with elements that can help them not only to understand why the system suggests a certain judgment or classification, but also with elements whose aim is to foster the reasoning process of the user by, e.g., estimating the probability that the system is right, appraising the complexity of the automated decision (by going beyond the self-reported confidence scores), assessing the extent to which the training data are representative and reliable, and the similarity of the case under consideration to those available, by letting users focus on the possible commonalities and differences between these.

In short, we then make our own design recommendations and principles highlighted in [67], where explainability does encompass the capability of a system to explain itself and its advice. More broadly, it regards whether the system is usable and understandable. In this light, an adequate, explainable system increases the understandability and interpretability of a case (not of itself) by a specific user working in a specific context, and is capable to improve situated decisions about specific cases.

This is why it is important to make research and development initiatives in the XAI discourse more human-centered [68] and more focused on the relationships [69] and interactions [65] that are established and continuously configured between the practitioners and their tools, which are aimed at supporting cognition, in particular judgment and prospection (that is, the action of looking forward into the future). This requires the adoption of methodologies and methods to assess usability along the traditional dimensions of decision-making effectiveness (that is, accuracy of the hybrid decision making [70]), efficiency (e.g., the extent providing explanations makes decision making more or less time-consuming), and decision-maker satisfaction (the extent to which decision makers find the additional information useful, or at least they feel more confident in their decision after consuming an explanation), as well as the failure modes and the impact these technologies have on both decision-making practices (e.g., considering patterns of reliance and risks of overreliance) and the subjects involved (i.e., the people who are affected by the decisions).

In particular, our findings suggest that we should not take the perceived and actual utility of explainable systems for granted. These qualities should be assessed in an ongoing manner, and in vivo rather than in labo, ensuring that quantitative measures of performance such as error rates, throughput, and execution times do not take undue precedence over the evaluation of user experience [71]. Rather, we should ground our design choice to make AI systems explainable (that is, capable of supplying explanations) on empirical evidence on the basis of the fit (or cognitive congruence) between the user and the artifact (i.e., trust and expectations), user and task (i.e., skill-difficulty match, expertise), and artifact and task to support, as proposed in the theory of dominance [39,40].

By evaluating the quality and usefulness of explanations in relation to user perception and performance, rather than in isolation, our study brought to light some paradoxical effects related to the introduction of explanations into diagnostic AI systems. Therefore our study aims at contributing to the discussion around the necessity of a relational approach to AI design and evaluation. Following Virginia Dignum [69], we also call for greater attention to the dynamics of decision-making settings, as well as to how humans and machines come to interact, and even "collaborate", in so-called hybrid human–artificial intelligence ensembles [70,72]. This leads to our belief that, even more so than from engineering and computer science, the greatest advances for AI are likely to emerge from a multidisciplinary effort gathering relevant contributions from the scholarly fields of cognitive ergonomics [73,74], social psychology [75], human–computer interaction [68], computer-supported cooperative work [76,77], naturalistic decision-making [78] and human factors [79,80].

Author Contributions: Conceptualization: F.C. and A.C.; Methodology: F.C. and A.C.; Software: A.C.; Validation: F.C., A.C., C.N., E.P. and L.R.; Data Curation: L.R. and M.C.; Writing-Original Draft Preparation: F.C., A.C., C.N. and E.P.; Writing-Review and Editing: F.C., A.C., C.N., E.P., L.R. and M.C.; Visualization: F.C. and A.C.; Supervision: F.C. and M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been supported by the Italian Ministry of Health through project "Ricerca Corrente".

Data Availability Statement: All code and data is publicly available on GitHub at https://github.com/MUDILab/evaluate-human-ai-interaction accessed on 19 December 2022.

Acknowledgments: The authors are grateful to the anonymous cardiologists and cardiology residents of the Specialization School of the University of Siena, Italy, and in particular to Giulia Mandoli and Maria Concetta Pastore for their coordination and support in the empirical study.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

ADM	Automated Decision Making
BPMN	Business Process Model and Notation (i.e., the OMG standard on process modelling)
AI	Artificial Intelligence
ECG	ElectroCardioGram
EU	European Union
FHD	Final Human Decision
GDPR	General Data Protection Regulation (cf. EU 2016/679)
HD	Human Decision
ML	Machine Learning
RBC	Rank Biserial Correlation
QT	QT interval, i.e., the time from the start of the Q wave to the end of the T wave in an *ECG*
RQ	Research Question
TTD	Theory of Technological Dominance
XAI	eXplainable AI

References

1. Calegari, R.; Ciatto, G.; Omicini, A. On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intell. Artif.* **2020**, *14*, 7–32. [CrossRef]
2. Springer, A.; Hollis, V.; Whittaker, S. Dice in the black box: User experiences with an inscrutable algorithm. In Proceedings of the 2017 AAAI Spring Symposium Series, Stanford, CA, USA, 27–29 March 2017.
3. Cinà, G.; Röber, T.; Goedhart, R.; Birbil, I. Why we do need Explainable AI for Healthcare. *arXiv* **2022**, arXiv:2206.15363.
4. Gerlings, J.; Shollo, A.; Constantiou, I. Reviewing the need for explainable artificial intelligence (xAI). *arXiv* **2020**, arXiv:2012.01007.
5. Goebel, R.; Chander, A.; Holzinger, K.; Lecue, F.; Akata, Z.; Stumpf, S.; Kieseberg, P.; Holzinger, A. Explainable AI: The new 42? In Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Hamburg, Germany, 27 August 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 295–303.
6. De Bruijn, H.; Warnier, M.; Janssen, M. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Gov. Inf. Q.* **2022**, *39*, 101666. [CrossRef]
7. Janssen, M.; Hartog, M.; Matheus, R.; Yi Ding, A.; Kuk, G. Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. *Soc. Sci. Comput. Rev.* **2022**, *40*, 478–493. [CrossRef]
8. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57. [CrossRef]
9. Schemmer, M.; Kühl, N.; Benz, C.; Satzger, G. On the Influence of Explainable AI on Automation Bias. *arXiv* **2022**, arXiv:2204.08859.
10. Poursabzi-Sangdeh, F.; Goldstein, D.G.; Hofman, J.M.; Wortman Vaughan, J.W.; Wallach, H. Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–52.
11. Zhang, Y.; Liao, Q.V.; Bellamy, R.K. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 295–305.
12. Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M.T.; Weld, D. Does the whole exceed its parts? The effect of ai explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–16.
13. Buçinca, Z.; Malaya, M.B.; Gajos, K.Z. 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*, 1–21. [CrossRef]
14. Suresh, H.; Lao, N.; Liccardi, I. Misplaced trust: Measuring the interference of machine learning in human decision-making. In Proceedings of the 12th ACM Conference on Web Science, Southampton, UK, 6–10 July 2020; pp. 315–324.
15. Eiband, M.; Buschek, D.; Kremer, A.; Hussmann, H. The impact of placebic explanations on trust in intelligent systems. In Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–6.
16. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [CrossRef]
17. Finzel, B.; Saranti, A.; Angerschmid, A.; Tafler, D.; Pfeifer, B.; Holzinger, A. Generating Explanations for Conceptual Validation of Graph Neural Networks: An Investigation of Symbolic Predicates Learned on Relevance-Ranked Sub-Graphs. *KI-Künstl. Intell.* **2022**, *36*, 271–285. [CrossRef]
18. 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.* **2022**, *213*, 118888. [CrossRef]

19. Green, B.; Chen, Y. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 90–99.
20. 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. [CrossRef]
21. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1312. [CrossRef] [PubMed]
22. Langer, E.J.; Blank, A.; Chanowitz, B. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *J. Personal. Soc. Psychol.* **1978**, *36*, 635. [CrossRef]
23. Cabitza, F.; Campagner, A.; Simone, C. The need to move away from agential-AI: Empirical investigations, useful concepts and open issues. *Int. J. Hum.-Comput. Stud.* **2021**, *155*, 102696. [CrossRef]
24. Cabitza, F.; Campagner, A.; Famiglini, L.; Gallazzi, E.; La Maida, G.A. Color Shadows (Part I): Exploratory Usability Evaluation of Activation Maps in Radiological Machine Learning. In Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Vienna, Austria, 23–26 August 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 31–50.
25. Parimbelli, E.; Peek, N.; Holzinger, A.; Guidotti, R.; Mittelstadt, B.; Dagliati, A.; Nicora, G. Explainability, Causability, Causality, Reliability: The many facets of "good" explanations in XAI for health. In Proceedings of the Challenges of Trustable AI and Added-Value on Health, Nice, France, 27–30 May 2022; EFMI: Nice, France, 2022.
26. Ebrahimi, Z.; Loni, M.; Daneshtalab, M.; Gharehbaghi, A. A review on deep learning methods for ECG arrhythmia classification. *Expert Syst. Appl. X* **2020**, *7*, 100033. [CrossRef]
27. Huang, J.S.; Chen, B.Q.; Zeng, N.Y.; Cao, X.C.; Li, Y. Accurate classification of ECG arrhythmia using MOWPT enhanced fast compression deep learning networks. *J. Ambient. Intell. Humaniz. Comput.* **2020**, 1–18. [CrossRef]
28. Chen, C.Y.; Lin, Y.T.; Lee, S.J.; Tsai, W.C.; Huang, T.C.; Liu, Y.H.; Cheng, M.C.; Dai, C.Y. Automated ECG classification based on 1D deep learning network. *Methods* **2022**, *202*, 127–135. [CrossRef] [PubMed]
29. Bond, R.; Finlay, D.; Al-Zaiti, S.S.; Macfarlane, P. Machine learning with electrocardiograms: A call for guidelines and best practices for 'stress testing' algorithms. *J. Electrocardiol.* **2021**, *69*, 1–6. [CrossRef]
30. Rojat, T.; Puget, R.; Filliat, D.; Del Ser, J.; Gelin, R.; Díaz-Rodríguez, N. Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey. *arXiv* **2021**, arXiv:2104.00950.
31. Raza, A.; Tran, K.P.; Koehl, L.; Li, S. Designing ECG monitoring healthcare system with federated transfer learning and explainable AI. *Knowl.-Based Syst.* **2022**, *236*, 107763. [CrossRef]
32. Panigutti, C.; Perotti, A.; Pedreschi, D. Doctor XAI: An ontology-based approach to black-box sequential data classification explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), Barcelona, Spain, 27–30 January 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 629–639. [CrossRef]
33. Lundberg, S.M.; Nair, B.; Vavilala, M.S.; Horibe, M.; Eisses, M.J.; Adams, T.; Liston, D.E.; Low, D.K.W.; Newman, S.F.; Kim, J.; et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749–760. [CrossRef]
34. Neves, I.; Folgado, D.; Santos, S.; Barandas, M.; Campagner, A.; Ronzio, L.; Cabitza, F.; Gamboa, H. Interpretable heartbeat classification using local model-agnostic explanations on ECGs. *Comput. Biol. Med.* **2021**, *133*, 104393. [CrossRef] [PubMed]
35. Dahlbäck, N.; Jönsson, A.; Ahrenberg, L. Wizard of oz studies—Why and how. In Proceedings of the 1993 International Workshop on Intelligent User Interfaces, Orlando, FL, USA, 4–7 January 1993; pp. 4–7.
36. Ronzio, L.; Campagner, A.; Cabitza, F.; Gensini, G.F. Unity Is Intelligence: A Collective Intelligence Experiment on ECG Reading to Improve Diagnostic Performance in Cardiology. *J. Intell.* **2021**, *9*, 17. [CrossRef]
37. Nourani, M.; King, J.; Ragan, E. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Hilversum, The Netherlands, 25–29 October 2020; Volume 8, pp. 112–121.
38. Kim, A.; Yang, M.; Zhang, J. When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms. *ACM Trans. Comput.-Hum. Interact.* **2020**. [CrossRef]
39. Arnold, V.; Sutton, S.G. The theory of technology dominance: Understanding the impact of intelligent decision aids on decision maker's judgments. *Adv. Account. Behav. Res.* **1998**, *1*, 175–194.
40. Sutton, S.G.; Arnold, V.; Holt, M. An Extension of the Theory of Technology Dominance: Understanding the Underlying Nature, Causes and Effects. *Causes Eff.* **2022**. Available online: https://www.nhh.no/globalassets/centres/digaudit/activities/sutton-arnold-and-holt-2022-april-an-extension-of-the-theory-of-technology-dominance.pdf (accessed on 30 April 2022). [CrossRef]
41. Glick, A.; Clayton, M.; Angelov, N.; Chang, J. Impact of explainable artificial intelligence assistance on clinical decision-making of novice dental clinicians. *JAMIA Open* **2022**, *5*, ooac031. [CrossRef] [PubMed]
42. Paleja, R.; Ghuy, M.; Ranawaka Arachchige, N.; Jensen, R.; Gombolay, M. The Utility of Explainable AI in Ad Hoc Human-Machine Teaming. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 610–623.
43. Noga, T.; Arnold, V. Do tax decision support systems affect the accuracy of tax compliance decisions? *Int. J. Account. Inf. Syst.* **2002**, *3*, 125–144. [CrossRef]
44. Arnold, V.; Collier, P.A.; Leech, S.A.; Sutton, S.G. Impact of intelligent decision aids on expert and novice decision-makers' judgments. *Account. Financ.* **2004**, *44*, 1–26. [CrossRef]

45. Jensen, M.L.; Lowry, P.B.; Burgoon, J.K.; Nunamaker, J.F. Technology dominance in complex decision making: The case of aided credibility assessment. *J. Manag. Inf. Syst.* **2010**, *27*, 175–202. [CrossRef]
46. Cabitza, F. Biases affecting human decision making in AI-supported second opinion settings. In Proceedings of the International Conference on Modeling Decisions for Artificial Intelligence, Milan, Italy, 4–6 September 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 283–294.
47. Cabitza, F.; Campagner, A.; Ronzio, L.; Cameli, M.; Mandoli, G.E.; Pastore, M.C.; Sconfienza, L.; Folgado, D.; Barandas, M.; Gamboa, H. Rams, Hounds and White Boxes: Investigating Human-AI Collaboration Protocols in Medical Diagnosis. *Artif. Intell. Med.* **2022**, *submitted*. [CrossRef]
48. Bansal, G.; Nushi, B.; Kamar, E.; Horvitz, E.; Weld, D.S. Is the most accurate ai the best teammate? optimizing ai for teamwork. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 11405–11414.
49. Nunnally, J.; Bernstein, I. *Psychometric Theory*, 3rd ed.; McGraw-Hil: New York, NY, USA, 1994.
50. Cho, E.; Kim, S. Cronbach's coefficient alpha: Well known but poorly understood. *Organ. Res. Methods* **2015**, *18*, 207–230. [CrossRef]
51. Gaube, S.; Suresh, H.; Raue, M.; Merritt, A.; Berkowitz, S.J.; Lermer, E.; Coughlin, J.F.; Guttag, J.V.; Colak, E.; Ghassemi, M. Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **2021**, *4*, 31. [CrossRef] [PubMed]
52. Brill, T.M.; Munoz, L.; Miller, R.J. Siri, Alexa, and other digital assistants: A study of customer satisfaction with artificial intelligence applications. *J. Mark. Manag.* **2019**, *35*, 1401–1436. [CrossRef]
53. Yang, J.; Hurmelinna-Laukkanen, P. Benefiting from innovation–Playing the appropriability cards. In *Innovation*; Routledge: Abingdon, UK, 2022; pp. 310–331.
54. Jacobs, M.; Pradier, M.F.; McCoy, T.H.; Perlis, R.H.; Doshi-Velez, F.; Gajos, K.Z. How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. *Transl. Psychiatry* **2021**, *11*, 108. [CrossRef]
55. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
56. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
57. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. [CrossRef]
58. Reason, J. Human error: Models and management. *BMJ* **2000**, *320*, 768–770. [CrossRef] [PubMed]
59. Parimbelli, E.; Buonocore, T.M.; Nicora, G.; Michalowski, W.; Wilk, S.; Bellazzi, R. Why did AI get this one wrong?—Tree-based explanations of machine learning model predictions. *Artif. Intell. Med.* **2023**, *135*, 102471. [CrossRef] [PubMed]
60. Amann, J.; Vetter, D.; Blomberg, S.N.; Christensen, H.C.; Coffee, M.; Gerke, S.; Gilbert, T.K.; Hagendorff, T.; Holm, S.; Livne, M.; et al. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLoS Digit. Health* **2022**, *1*, e0000016. [CrossRef]
61. Shortliffe, E.H.; Davis, R.; Axline, S.G.; Buchanan, B.G.; Green, C.C.; Cohen, S.N. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Comput. Biomed. Res. Int. J.* **1975**, *8*, 303–320. [CrossRef]
62. Bos, J.M.; Attia, Z.I.; Albert, D.E.; Noseworthy, P.A.; Friedman, P.A.; Ackerman, M.J. Use of artificial intelligence and deep neural networks in evaluation of patients with electrocardiographically concealed long QT syndrome from the surface 12-lead electrocardiogram. *JAMA Cardiol.* **2021**, *6*, 532–538. [CrossRef] [PubMed]
63. Klein, G.; Hoffman, R.; Mueller, S. Naturalistic Psychological Model of Explanatory Reasoning: How people explain things to others and to themselves. In Proceedings of the International Conference on Naturalistic Decision Making, San Francisco, CA, USA, 17–21 June 2019.
64. Gunning, D.; Aha, D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* **2019**, *40*, 44–58.
65. Green, B.; Chen, Y. The principles and limits of algorithm-in-the-loop decision making. *Proc. Acm Hum.-Comput. Interact.* **2019**, *3*, 1–24. [CrossRef]
66. Vaccaro, M.; Waldo, J. The Effects of Mixing Machine Learning and Human Judgment: Collaboration between humans and machines does not necessarily lead to better outcomes. *Queue* **2019**, *17*, 19–40. [CrossRef]
67. Mueller, S.T.; Veinott, E.S.; Hoffman, R.R.; Klein, G.; Alam, L.; Mamun, T.; Clancey, W.J. Principles of explanation in human-AI systems. *arXiv* **2021**, arXiv:2102.04972.
68. Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum.–Comput. Interact.* **2020**, *36*, 495–504.
69. Dignum, V. Relational Artificial Intelligence. *arXiv* **2022**, arXiv:2202.07446.
70. Reverberi, C.; Rigon, T.; Solari, A.; Hassan, C.; Cherubini, P.; Cherubini, A. Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci. Rep.* **2022**, *12*, 14952. [CrossRef] [PubMed]
71. Holzinger, A.T.; Muller, H. Toward Human–AI Interfaces to Support Explainability and Causability in Medical AI. *Computer* **2021**, *54*, 78–86. [CrossRef]

72. Dellermann, D.; Calma, A.; Lipusch, N.; Weber, T.; Weigel, S.; Ebel, P. The future of human-ai collaboration: A taxonomy of design knowledge for hybrid intelligence systems. In Proceedings of the Hawaii International Conference on System Sciences (HICSS), Maui, HI, USA, 8–11 January 2019.
73. Andrews, R.W.; Lilly, J.M.; Srivastava, D.; Feigh, K.M. The role of shared mental models in human-AI teams: A theoretical review. *Theor. Issues Ergon. Sci.* **2022**, *2*, 1–47. [CrossRef]
74. Neerincx, M.A.; Waa, J.v.d.; Kaptein, F.; Diggelen, J.v. Using perceptual and cognitive explanations for enhanced human-agent team performance. In Proceedings of the International Conference on Engineering Psychology and Cognitive Ergonomics, Las Vegas, NV, USA, 15 July 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 204–214.
75. Cooke, N.J.; Lawless, W.F. Effective Human–Artificial Intelligence Teaming. In *Systems Engineering and Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 61–75.
76. Liu, B. In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *J. Comput.-Mediat. Commun.* **2021**, *26*, 384–402. [CrossRef]
77. Wang, D.; Churchill, E.; Maes, P.; Fan, X.; Shneiderman, B.; Shi, Y.; Wang, Q. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–6.
78. Klein, G. A naturalistic decision making perspective on studying intuitive decision making. *J. Appl. Res. Mem. Cogn.* **2015**, *4*, 164–168. [CrossRef]
79. Asan, O.; Choudhury, A. Research trends in artificial intelligence applications in human factors health care: Mapping review. *JMIR Hum. Factors* **2021**, *8*, e28236. [CrossRef] [PubMed]
80. Parasuraman, R.; Sheridan, T.B.; Wickens, C.D. A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern.-Part A Syst. Humans* **2000**, *30*, 286–297. [CrossRef]