# Fostering Human-AI interaction: development of a Clinical Decision Support System enhanced by eXplainable AI and Natural Language Processing

Laura Bergomi[1,*]

[1]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

**Abstract**

Artificial Intelligence (AI) is increasingly integrated into Decision Support Systems (DSS). The explainability of AI-based systems becomes crucial in sensitive and critical domains, such as healthcare, where ethical considerations and reliability are paramount concerns. In the clinical setting, it is important to evaluate how humans and AI can collaborate on cognitive tasks. Collaboration protocols (HAI-CP) allow for the investigation of the usefulness of AI models and their impact on users (both positive and negative). Although research on the application of these methods is blooming, there is little understanding of the impact on clinical decision-making, especially for eXplainable AI (XAI) systems, due to the lack of user studies. Therefore, the goal of this proposal is to develop a clinical DSS enhanced by XAI and Natural Language Processing (NLP): their synergy can add value to the interaction between users and AI, fostering a more linguistically natural, comprehensible, trustworthy, and supporting interfacing, that blends into the existing workflows. This proposal explores potential solutions to tailor natural language explanations and data visualizations to the end-user, improving the comprehensibility of the reasons behind a decision, and increasing the user's confidence in the decision; investigates and tests possible strategies to "get the patient-in-the-loop"; explores uncertainty quantification and counterfactual approaches, and finally assesses the impact on naturalistic (i.e., real-world) decision-making and long-term effects and biases.

**Keywords**

Clinical decision making, Explainable artificial intelligence, Natural language interaction, Human-AI collaboration protocol

## 1. Introduction

Artificial Intelligence (AI) is increasingly integrated into Decision Support Systems (DSS). To understand how useful and usable an AI-based system actually is, it is important to assess its impact on the decision support workflow. For a DSS to meet the basic requirement of Art. 22 para 1 GDPR (which prohibits "decision based solely on automated processing"), it is important to provide that the human-in-the-loop has substantial evaluative power and so the last word on the outcome of a decision [1]. To ensure that a human can make a thoughtful and reliable decision, the explainability of AI-based systems becomes crucial. In sensitive and critical domains, such as healthcare, where ethical considerations and reliability are paramount

*Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta*

*Corresponding author.

laura.bergomi01@universitadipavia.it (L. Bergomi)

0009-0006-0359-5128 (L. Bergomi)

concerns, AI-based DSSs should provide high-quality explanations. In this context, eXplainable AI (XAI) has become increasingly important as a counterbalancing force to the widespread adoption of complex black box models, that leave users, and even developers, in the dark as to how results were obtained [2]. XAI refers to the development of AI systems that can provide clear, understandable, and interpretable explanations. XAI methods can be described based on several categorizations; what remains essential is to adapt and test their explanations in Human-Artificial Intelligence collaboration protocol (HAI-CP). An HAI-CP is "the instance of a process schema that stipulates the use of AI tools by competent practitioners to perform a certain task or do a certain job" [3], thus allowing to study the interaction of several parameters involving at least the following dimensions : Affordance (functionalities and task automation), Fit (fitting into the existing work practice), Optimization (learning phase tuning), Output (type of result returned) and Target (the characteristics of the intended user).

Despite the application of AI methods in healthcare being a highly active field of research [4], how AI recommendations affect clinical decision-making is still poorly understood due to the lack of user studies [5]. This motivates the present proposal to investigate extensively how AI systems can be integrated into medical practice, being an aid that is easily understood, but also naturally interfaceable (enabling natural language interaction, NLI). The latter issue can be supported by using Natural Language Processing (NLP) techniques to generate explanations in natural language, that are tailored to the end-user. The type of end-user has intentionally not been specified, suggesting that it may be not only a medical professional, but also the patient (one of the original contributions of the research). GDPR (Recital 71) remarks that providing information about the existence of automated processes, the logic behind them, and their potential consequences should nevertheless be provided voluntarily as a good practice to ensure fairness and transparency. This means patients should have access to explanations for clinical decisions, allowing them to engage in discussions about their care and understand the reasoning behind treatments or examinations.

This proposal is part of the Italian project PRIN PNRR 2022 "InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision-making" (CUP: H53D23008090001 funded by the European Union - Next Generation EU), whose aim is to explore a model-agnostic perspective on the development and evaluation of AI-based (medical) DSSs. Our focus is on the interaction between an AI system and the human user, i.e., understanding to what extend the advice (and the way it is presented) can influence, be understood, and used by users; the interaction features and effects.

The rest of the paper is organized as follows. A brief review of related work is provided in Section 1.1; Section 2 presents the goals of this research, followed by the planned approaches and methods to achieve them in Section 3. Expected results of our research are proposed in Section 4, and finally, the conclusions are described in Section 5 offering some questions I would like feedback on.

### 1.1. Related works

A human-in-the-loop approach has been advocated as essential for proper evaluation of AI for healthcare [6]: explanations are ultimately directed to a human (expert) interacting with the AI system and should be optimized for this. Following this direction, Gaube et al. [7] compare

the detection of abnormalities in chest radiology images by physicians with the support of an AI or a second human agent; Tschandl et al. [8] evaluate, both with and without the use of AI, clinicians skin cancers recognition. Cabitza et al. [3, 9] propose experimental results of applying XAI to knee MRI and ECG studies, intending to compare the effectiveness of HAI-CP (testing different orders of presentation (human first vs. AI first) and availability of explanations (yes vs. no)). In [9] the support of the AI system includes both a proposed diagnosis and a textual explanation to back the former one.

Some surveys (e.g. [10, 11]) have studied the use of natural language techniques in creating explanations, e.g. the synergy of NLP and XAI methods. Sokol and Flach [12] argue that natural language explanations give the process a natural feeling, increasing the reliability of explanations and helping to gain acceptance from a wider range of users. Despite these considerations, a small part of XAI's works uses natural language presentation methods [10]. Works in the literature concerning the involvement of humans in clinical decision-making, such as those cited so far, point to the clinician as the user (i.e., the human-in-the-loop). Only a few examples involve patients, e.g., Donatello et al. [13] address the challenges of a system that supports patients in following a healthy behavior by presenting an XAI system that supports the monitoring of users' behaviors and persuades them to follow a healthy lifestyle.

Ultimately, some works (e.g. [3, 14, 15]) emphasize the need to examine biases in AI and XAI support and focus on AI advice effects.

## 2. Research goals

Current XAI research often overlooks the importance of presentation techniques, leading to challenges for researchers and practitioners in selecting appropriate methods for explainability. The lack of comprehensive studies adds complexity and potential errors to the process [10]. This research will focus primarily on two main dimensions: the output and the target. In the first case (i.e., the output) the goal is to compare the presentation techniques of XAI methods, studying what type of output the end-user prefers and understands. The output of an XAI system can be multimodal (e.g., classes, confidence scores, category lists, visual or textual explanations, etc.). A better understanding of this output allows users to increase the overall acceptability of the system. In the second case (i.e., the target) the goal is to tailor explanations to the intended end-user in terms of profiling (expertise, experience, role, etc.), content, and form. Going towards dialogues that directly engage the user in the explanation process, we can offer rich and personalized interactions that mimic how humans explain their decisions.

- What factors determine the choice of how to represent the explanations obtained from AI-based methods?
- Does XAI support, both in terms of visual aids and textual explanations, have a significant effect on taking better clinical decisions, and reducing errors?

Clinicians may lack the computer science expertise to comprehend algorithmic decision-making processes, highlighting the need for clear communication. As a result, it becomes imperative for clinicians, as decision-makers, to ensure that decisions made through these processes can be clearly communicated to patients, who may have little technical knowledge.

- What is the impact of asking decision-makers to explain their AI-based decisions?
- What is the best way to "get the patient-in-the-loop"?

Regarding the assessment of impact on decision-making, we are asking about counterfactual explanations and biases in AI and XAI support:

- What is the impact of providing decision-makers with similar cases, or providing counterfactual outcomes, or playing the devil's advocate role?
- Does the AI advice affect the users' performance and proficiency-building processes, both in the short and in the long run? Is this influence also relevant according to the user's experience, expertise, and role?
- Are explanations always beneficial or can they induce paradoxically harmful effects?

## 3. Planned approaches and methods

This section outlines the main phases of the research (as shown in the Gantt diagram in Figure 1), the approaches, and methods thought to be used as a starting point in each phase and from which to develop extensions and insights. At each stage, there is the analysis of State-of-Art methods and the implementation of HAI-CPs to test hypotheses and collect reliable and naturalistic (i.e., real-world) decision-making results. Particular attention will be paid to interpreting and ranking the relative strength of claims about the effectiveness of design solutions and their superiority concerning other possible alternatives [16]. The inXAID project will ensure that the frameworks, metrics and methods developed, as well as the results of the experiments performed throughout the duration of the research activity, will be disseminated to the appropriate target communities and audiences. A research period abroad is also planned.

| Research activity topic | I year* | | | | | | II year* | | | | | | III year | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bim. 1 | bim. 2 | bim. 3 | bim. 4 | bim. 5 | bim. 6 | bim. 1 | bim. 2 | bim. 3 | bim. 4 | bim. 5 | bim. 6 | bim. 1 | bim. 2 | bim. 3 | bim. 4 | bim. 5 | bim. 6 |
| RP 1: Tailoring XAI and NLP explanations | | | | | | | | | | | | | | | | | | |
| XAI / NLP: deriving the explanation | SoA | TI | | RA | | | | | | | | | | | | | | |
| XAI / NLP: data presentation to the end-user | SoA | TI | ES 1 | | RA | | | | | | | | | | | | | |
| Combination of XAI and NLP | | | TI | | RA | | | | | | | | | | | | | |
| RP 2: Get the patient-in-the-loop | | | | | | | | | | | | | | | | | | |
| Review: patient-AI collaborations | | | | | SoA | | RP | | | | | | | | | | | |
| "Patient-out-the-loop" | | | | | | TI | ES 2 | | RA | | | | | | | | | |
| "Patient aware-of-the-loop" | | | | | | | | ES 3 | | RA | | | | | | | | |
| "Patient-in-the-loop" | | | | | | | | | TI | ES 4 | | RA | | | | | | |
| RP 3: Assessment of impact on decision-making | | | | | | | | | | | | | | | | | | |
| Uncertainty Quantification (UQ) assessment | | | | | | | | | | | SoA | ES 5 | | RA | | | | |
| Counterfactual explanations methods | | | | | | | | | | | SoA | TI | | RA | | | | |
| Reliance patterns and biases | | | | | | | | | | | | | SoA | ES 6 | | RA | | |
| Mitigation strategies | | | | | | | | | | | | | SoA | | | RA | | |
| Further phases | | | | | | | | | | | | | | | | | | |
| Dissemination | | | | | | | | | | | | | | | | | | |
| PhD dissertation writing | | | | | | | | | | | | | | | | | | |

**Figure 1:** Gantt chart of the possible timeline of Research phases (RP) and related activities. * denotes the duration of the inXAID project. Abbreviations: SoA= analysis of State-of-Art methods and approches, TI= Tools Implementation, ES= Evaluation Study with users, RA= Results Analysis, RP= Review Pubblication.

**Research phase 1: Tailoring XAI and NLP explanations.** In this phase, the aim is to investigate (i) techniques for deriving the explanation and (ii) presentation to the end-user. The main XAI techniques to test are related to feature importance analysis (e.g. SHAP [17], T-EBAnO [18]) and the use of surrogate models (e.g. LIME [19], AraucanaXAI [20]). Comparing different models, we can assess which type of data visualization is preferred by users. Among the existing NLP models, we analyze transformers models for two pivotal reasons: they rely on the attention mechanism and they are exceptionally effective for common natural language understanding (NLU) and natural language generation (NLG) tasks [21]. A textual explanation may be presented in different ways to the end-user. Some techniques to test include: saliency, visualization of the importance scores, showing input-output word alignment, highlighting words in input text, or displaying extracted relations or word clouds; rewrite the explanations by changing linguistic register. A possible comparison to explore concerns, on the one hand, is the self-explaining approach, which generates the explanation at the same time as the prediction; on the other hand, the post-hoc approach, which requires that an additional operation be performed after the return of the prediction. The final step regards the combination of XAI and NLP techniques; a possible simple solution could be rendering XAI explanations through NLG or using XAI techniques to explain image, text, and graph classification models (e.g., Layer-wise relevance propagation [22]); or incorporating feedback from human users to improve the explanations generated by the model. This could involve allowing users to provide feedback on the explanations or incorporating user preferences into the explanation process. In the example of [3, 9], it is also important to test alternative fitting of AI advice in existing practice (e.g., human first vs. AI first, availability of explanations only in critical cases, AI advice on request).

**Research phase 2: Get the patient-in-the-loop.** In this phase, the aim is to investigate different strategies to involve the patient in the clinical decision-making process. Firstly, it is necessary to review the existing work in the literature regarding surveys on patient-AI collaborations. Subsequently the idea is to investigate different ways of including patients, possibly going through step: (i) give clinicians help in explaining their decision (patient-out-the-loop); (ii) analyze how patients' behavior changes knowing that the doctor's decision about their health status is based on AI advice (patient aware-of-the-loop); (iii) use conversational chatbots to interact with patients and, thus, collect important information (patient-in-the-loop). In this direction, Bennettot et al. [21] provide an example of how XAI can benefit healthcare, specifically in monitoring diabetic patients. They discuss a virtual coaching system that offers guidance on healthy behaviors based on patient-reported data. When the system detects undesirable behaviors, it generates tailored explanations for clinicians and patients. This approach enables clinicians to adjust treatment decisions as needed, while patients gain confidence, feel supported, and become more engaged in their care decisions. At this level, also for ethical considerations, the inclusion of a professional figure such as a behavioral psychologist, is essential.

**Research phase 3: Assessment of impact on decision-making.** In this phase, the aim is to assess how, how much, and when, an AI-based system influences the cognitive processes involved in clinicians' interpretation of AI support. An explanation, whether true or fictitious, could be correlated by Uncertainty Quantification (UQ) assessment. In this context, we can

apply different approaches, also integrated in NLP models [18, 23]. On the other hand, some methods explain the model by providing information on feature-perturbed versions of the analyzed instance. These methods fall into the counterfactual explanations methods [21]. Some common approaches to test could be [24]: add or delete information to what users know about the facts; create counterfactuals that imagine how the outcome could have been better or worse; construct explanations by ensuring they identify cause-effect or reason-action relations between events; semi-factual about how the outcome could have been the same "even if" the action had been different. Identifying the cognitive processes involved in physicians' interpretation of AI support also precludes analysis of how these same processes may change over time. It can be useful to realize a mapping of reliance patterns and biases affecting human decision-making in the short and long run (e.g., preference for usability over performance, more attention to false negatives than false positives, deskilling/ upskilling etc.); moreover, summarize promising mitigation strategies and research directions to support human critical thinking (e.g., delay showing the AI's prediction and/or explanations, give arguments for non-predicted outcomes, enable to actively explore the data, etc.).

## 4. Expected results

The main objective of this research proposal is to make a significant contribution towards advancing the methodological leveraging of XAI methods and natural language explanations, fostering a more linguistically natural, comprehensible, trustworthy, and supporting interfacing among AI systems and human users, that fits into the existing work practices. On the clinical front, the project's main objective is to enhance decision-making processes by expanding output solutions and taking into consideration aspects that are currently unexplored. These aspects include the contribution of socio-technical elements, experience, expertise, habits, and environmental and temporal information towards cognitive processes. It is expected that "getting the patient-in-the-loop" of decision-making can be an original strength for AI collaborative systems; so that it can also be supportive to the patient, improving well-being. Preliminary results and contributions to date are related to the test of ALFABETO [25] (whose aim is to aid clinicians during COVID-19 patients' hospital admission through the application of machine learning approaches exploiting clinical and chest x-ray features) in a clinical survey. In this framework, different predictions and explanations are proposed to clinicians: from an interpretable model (i.e., ALFABETO original Bayesian network) and from a black box model (i.e., Gradient boosting), in this latter case, with the explanations of two different XAI approaches (SHAP and Araucana XAI).

## 5. Research challenges and future direction

The research is still at an early stage, and thus, there are several challenges that we should cope with. In XAI, some works have claimed that explainability may come at the price of losing predictive performance. Studying such possible trade-offs is an important research area, but one that cannot advance until standardized metrics are developed for evaluating the quality of explanations [26]. Indeed, an open debate in the NLG community is about finding the right

way to measure the goodness of generated explanations [11]. The main issues revolve around whether to rely only on automatic metrics (e.g., ROUGE, BLEU), or instead, how to properly perform human evaluations. That said, although human evaluation remains the gold standard for overall system quality assessment, using it at every stage of the development process would be too costly and slow. Another challenge may be the interactive integration of desired behavior, fairness, correctness, and reliability; as well as verifying and integrating knowledge in each step of decision-making process, instead of XAI producing a single description of a static system [10].

From these general considerations, and more, arise the following questions, on which I would appreciate feedback by the DC mentors, as I believe they can help improve my research path:

- What metrics would be better to investigate (automatic or human) to evaluate the goodness of explanations, especially in natural language?
- How to engage the user interactively, hoping for fruitful and continuous use of the provided DSS?
- What other human actors (besides clinicians and patients) could we include?
- What aspects should we not forget to consider? (e.g., from an ethical and legal perspective)
- What approaches and potential collaborations should we consider to address the challenges of developing a reliable and explainable DSS using clinical data?

I would like to express my willingness to receive constructive criticism and suggestions on this work and to clarify any points that may need further explanation.

## Acknowledgments

I would like to express my gratitude to my supervisor, Enea Parimbelli, for his invaluable assistance in developing this project and its implementation in the future. Authors acknowledge funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision-making. CUP: H53D23008090001 funded by the European Union - Next Generation EU.

## References

[1] D. Schneeberger, et al., The european legal framework for medical ai, in: Machine Learning and Knowledge Extraction, 2020, pp. 209–226. doi:https://doi.org/10.1007/978-3-030-57321-8_12.

[2] C. Combi, et al., A manifesto on explainability for artificial intelligence in medicine, Artificial Intelligence in Medicine 133 (2022). doi:https://doi.org/10.1016/j.artmed.2022.102423.

[3] F. Cabitza, et al., Rams, hounds and white boxes: Investigating human–ai collaboration protocols in medical diagnosis, Artificial Intelligence in Medicine 138 (2023) 102506. doi:https://doi.org/10.1016/j.artmed.2023.102506.

[4] V. L. Patel, et al., The Coming of Age of Artificial Intelligence in Medicine, Artificial intelligence in medicine 46 (2009) 5–17. doi:10.1016/j.artmed.2008.07.017.

[5] F. Cabitza, et al., Quod erat demonstrandum? - towards a typology of the concept of explanation for the design of explainable ai, Expert Systems with Applications 213 (2023). doi:https://doi.org/10.1016/j.eswa.2022.118888.

[6] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop? 3 (2016) 119–131. doi:10.1007/s40708-016-0042-6.

[7] S. Gaube, et al., Do as ai say: susceptibility in deployment of clinical decision-aids, NPJ digital medicine 4 (2021) 31. doi:10.1038/s41746-021-00385-9.
[8] P. Tschandl, et al., Human-computer collaboration for skin cancer recognition, Nature Medicine 26 (2020) 1229–1234. doi:10.1038/s41591-020-0942-0.
[9] F. Cabitza, et al., Painting the black box white: Experimental findings from applying XAI to an ECG reading setting, Machine Learning and Knowledge Extraction 5 (2023) 269–286. doi:10.3390/make5010017.
[10] E. Cambria, et al., A survey on xai and natural language explanations, Information Processing & Management 60 (2023) 103111. doi:10.1016/j.ipm.2022.103111.
[11] K. Qian, et al., Xnlp: A living survey for xai research in natural language processing, in: 26th International Conference on Intelligent User Interfaces-Companion, 2021, pp. 78–80. doi:10.1145/3397482.3450728.
[12] K. Sokol, P. Flach, Conversational explanations of machine learning predictions through class-contrastive counterfactual statements, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 5785–5786. doi:10.24963/ijcai.2018/836.
[13] I. Donadello, M. Dragoni, C. Eccher, Persuasive explanation of reasoning inferences on dietary data, in: PROFILES/SEMEX@ISWC, volume 2465, 2019, pp. 46–61.
[14] A. Bertrand, et al., How cognitive biases affect xai-assisted decision-making: A systematic review, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2022, pp. 78–91. doi:10.1145/3514094.3534164.
[15] F. Cabitza, Biases affecting human decision making in AI-supported second opinion settings, in: Modeling Decisions for Artificial Intelligence, Springer International Publishing, 2019, pp. 283–294. doi:10.1007/978-3-030-26773-5_25.
[16] L. Famiglini, et al., Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems, Computers in Biology and Medicine 170 (2024). doi:10.1016/j.compbiomed.2024.108042.
[17] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017. doi:10.48550/arXiv.1705.07874.
[18] F. Ventura, et al., Trusting deep learning natural-language models via local and global explanations, Knowledge and Information Systems 64 (2022) 1863–1907. doi:10.1007/s10115-022-01690-9.
[19] M. T. Ribeiro, et al., "why should i trust you?": Explaining the predictions of any classifier, 2019. doi:10.48550/arXiv.1602.04938.
[20] E. Parimbelli, et al., Why did AI get this one wrong? — tree-based explanations of machine learning model predictions, Artificial Intelligence in Medicine 135 (2023). doi:10.1016/j.artmed.2022.102471.
[21] A. Bennetot, et al., A practical guide on explainable AI techniques applied on biomedical use case applications, 2022. doi:10.48550/arXiv.2111.14260.
[22] S. Bach, et al., On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015). doi:10.1371/journal.pone.0130140.
[23] S. H. Tanneru, et al., Quantifying uncertainty in natural language explanations of large language models, 2023. doi:10.48550/arXiv.2311.03533.
[24] R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning, IJCAI-19 (2019) 6276–6282. doi:10.24963/ijcai.2019/876.
[25] G. Nicora, et al., Bayesian networks in the management of hospital admissions: A comparison between explainable ai and black box ai during the pandemic, Journal of Imaging 10 (2024). doi:10.3390/jimaging10050117.
[26] M. Danilevsky, et al., A survey of the state of explainable AI for natural language processing, 2020. doi:10.48550/arXiv.2010.00711.