Do Explanations Matter? EMS Perceptions of AI in Emergency Scenarios

Lanna Labai

University of Haifa, Haifa, Israel, labai.lanna@gmail.com

Artificial intelligence (AI) is increasingly explored for its potential to support emergency medical services (EMS) professionals in high-pressure environments, yet challenges remain regarding trust, usability, and explainability. This study investigates how providing explanations for AI-generated treatment suggestions influences EMS personnel's perceptions of accuracy, comprehensibility, thoroughness, usefulness, and cognitive effort. Using a between-subjects experimental survey design, eight EMS professionals were assigned to one of three conditions: control (no AI), AI-only, or AI with explanations (XAI). Participants assessed AI treatment suggestions across six simulated emergency scenarios. Quantitative analysis employed nonparametric tests to compare perceptions across conditions and examine correlations, while qualitative thematic analysis explored reasons for non-agreement and differences between AI and human-created treatment plans. Results revealed that explanations did not improve perceptions of accuracy, comprehensibility, thoroughness, or usefulness, though they were associated with stronger inter-variable correlations and fewer identified issues in non-agreement explanations. Cognitive effort remained low across conditions, with significant differences only between the control and AI groups. EMS-created treatment plans were more detailed than AI-generated ones, and participants expressed skepticism about the necessity of AI in EMS decision-making. These findings contribute to the growing literature on explainable AI (XAI) by highlighting that explanations do not automatically enhance perceptions and must be carefully designed to reduce cognitive burden and align with user expertise.

CCS CONCEPTS • Human-Computer Interaction → Collaborative interaction

Additional Keywords and Phrases: EMS Perceptions, Decision-making, AI-assisted, Explainability

ACM Reference Format:

Lanna Labai. 2025. *[Do Explanations Matter? EMS Perceptions of XAI in Emergency Scenarios]*. In Research Methods II. September 14, 2025, Haifa, Israel, 13 pages.

1. INTRODUCTION

Artificial intelligence (AI) is increasingly being researched for implementation in emergency medicine (EM) and prehospital emergency care (PEC), showing potential to enhance diagnostic accuracy, optimize triage processes, and improve outcomes through applications like predicting critical conditions or optimizing resource allocation [1], [2]. Despite these promising opportunities, implementing AI in healthcare faces challenges, including mixed perceptions among the public and healthcare professionals [3], [8] and the need for higher quality evidence to confirm its superiority over standard methods [1], [4]. While AI is seen positively for potential efficiency benefits [3], there is a significant lack of trust [3], [8], stemming from concerns over data privacy, patient safety (including accuracy, reliability, and bias), medical liability, and a general lack of understanding or awareness about what AI is and what it can do [3], [5]. Automated decisions can be perceived as inscrutable [3].

Emergency medical services (EMS) operate in high-stakes environments requiring rapid decisions often made under significant uncertainty, particularly regarding understanding the situation and facing limitations compared to hospital staff [9]. Given the critical nature of EMS decisions, where human oversight remains essential, the explainability of AI suggestions is considered crucial for successful integration and building trust among EMS personnel [1], [7]. Understanding *why* an AI makes a specific prediction is essential for clinicians to trust the system, weigh its advice against their own experience, and justify their final decisions [7]. Design features that visualize the factors influencing an AI's suggestion are important for increasing user trust [7].

The motivation behind this research is to understand what will help EMS workers feel more at ease with the idea of using AI technologies in their line of work. Specifically, I want to investigate whether providing explanations alongside AI-generated suggestions for the EMS workers' next course of action in the face of uncertainty can help them feel more confident in relying on those suggestions. By examining this, we hope to gain a better understanding of how AI technologies should be implemented in EMS. Accordingly, the research question guiding this study is:

1. How does providing explanations for the rationale behind AI-generated suggestions affect EMS workers' perceptions of:
    a. The accuracy of the AI tool's suggestion?
    b. The comprehensibility of the AI tool's suggestion?
    c. The thoroughness of the AI tool's suggestion?
    d. The usefulness of the AI tool's suggestion?
    e. The cognitive effort needed to think about the treatment plan for the simulated emergency scenario?
2. What correlations emerge between perceptions of accuracy, comprehensibility, thoroughness, usefulness, and cognitive effort?
3. How do EMS worker-created treatment plans and AI-generated treatment plans compare in terms of the details provided?

This study will use a between-subjects experimental design to investigate how providing explanations for AI-generated suggestions affects EMS workers' perceptions of accuracy, usefulness, comprehensibility, and cognitive load. Participants will be divided into three groups: the control group, the group who will only be presented with the AI's suggestion only ("AI suggestion + no explanation"), and the group that will get an explanation alongside the AI's suggestion ("AI suggestion + explanation"). This will provide deeper insight into how explainability affects trust, understanding, and perceived value.

The contributions are as follows: (1) provides empirical evidence on explainable AI in EMS decision-making; (2) extends XAI research to EMS settings; (3) offers insights into perceptions regarding human–AI interaction in EMS.

The structure of the paper is as follows: Section 2 presents the "related works" of EMS challenges, AI integration in EMS, XAI implementations in medicine, and designing explanations for AI; Section 3 will describe the methodology used in this study including the method, participants, task description, and data analysis; Section 4 will present the results of the study; Section 5 will present the discussion regarding the results and discuss limitations and future work; and Section 6 will conclude the paper.

2. RELATED WORKS

2.1 Challenges in EMS

EMS is a high stakes, dynamic, and complex field to work in. EMS work is physically demanding, often involving long hours, irregular shifts, and heavy lifting in unsafe or unpredictable environments, frequent injuries, and exposure to pathogens and infectious diseases [12, 13, 16]. EMS work also exposes personnel to repeated trauma, violence, and emotionally charged situations such as treating children or witnessing death which contribute to high rates of PTSD, depression, anxiety, and burnout. Insufficient organizational recognition and limited recovery opportunities exacerbate these issues, leaving providers feeling unsupported and stigmatized when seeking help [12, 13, 16]. Furthermore, decision-making and situational awareness in EMS contexts is constrained by high uncertainty [8], incomplete or unreliable information [11, 14, 15], and rapidly changing patient conditions [12]. Studies highlight difficulties in teamwork coordination, time management, and the application of clinical judgment in chaotic environments [16]. Even experienced personnel struggle when forced to improvise under pressure, and team leaders carry a heavy cognitive burden in directing others while processing complex information [12, 14, 15].

2.2 AI Integration in EMS

To address the aforementioned issues, several studies have looked into the possible applications of AI in EMS. In their scoping review, Kirubarajan et al. (2020) found promising applications for acute radiographic imaging and prediction-based diagnoses. Usher et al. (2024) also provided opportunities for AI integration based on research from previous years to address cognitive, physical, emotional, and training-related challenges such as smart glasses that provide real-time advice on treating patients, or adaptive routing systems that provide the best way to get to the scene of the emergency. Researchers are also looking into the possibility of using AI-driven drones to deliver medical supplies to areas that are hard to reach [17]. MADA (the national EMS organization in Israel) has already implemented a speech-to-text program that helps dispatchers document the emergency incidents more accurately by transcribing the calls for them [18]. AI integration in EMS still remains limited.

2.3 Explainability in Medical AI Technologies

Explainability in AI (XAI) can be described as "allowing users to understand how the technology arrives at its predictions or recommendations" [19]. There are two general approaches to XAI: post-hoc explanations that aim to provide explanations for specific decisions; and ante-hoc explanations that are designed into the system [21]. The importance of explainability in AI in medicine comes down to increasing the acceptance of these tools by medical professionals [19]. XAI has wide implementations in medicine, for example, dispatch center cardiac arrest detection, visualization of survival prediction factors for cardiac arrest, AI models for classifying pathological images containing textual justifications, among many more [20].

Because research is lacking regarding XAI implementation challenges in EMS, we can borrow concepts from related fields such as medicine. Due to the sensitive nature of medical work, AI needs to be transparent, autonomous, detect biases, justify its decision-making, and not hurt the doctor-patient relationship [19]. It is relevant so that doctors can gauge the plausibility of its predictions and also for patients who need to receive adequate information to be able to give informed consent to treatments [19]. Shin et al. (2021) found that medical professionals' willingness to trust and adopt AI depends heavily on the clarity of explanations provided and when clinicians can understand why a recommendation was made. Without these explanations, professionals perceive AI as a "black box," raising concerns about accountability and patient safety. These are also true for EMS workers.

2.4 Designing Explanations for AI

According to Lim et al. (2019), intelligibility types are different forms of explanations that help users understand and interact with AI systems depending on their reasoning goals. They include: Inputs - what data the system is using; Outputs - what predictions or classifications it produces; Certainty - how confident it is in the prediction; Why - why a particular outcome occurred; Why Not - why an alternative prediction was not considered; What If - what would happen if inputs were changed; How To - what changes are needed for a desired outcome; and When - under what conditions an outcome occurs. Each type supports different reasoning processes: Inputs and Why/Why Not help users filter causes; Certainty and When explanations support generalization and learning; and What If and How To explanations allow prediction and control. The authors emphasize that designers should select explanation types strategically based on user goals, such as enhancing transparency, improving decision-making, debugging models, or calibrating trust, rather than defaulting to a single form like feature attribution. A subset of these intelligibility types were chosen to be used for this project based on the goal of finding and filtering causes.

3. METHODOLOGY

3.1. Experimental Survey Design

The study method chosen is a between-subjects experimental design to allow us to compare between different AI conditions to observe differences (if any) in perceptions of the AI suggestions across four dimensions. Ideally, this would be carried out in the physical work environments of EMTs and paramedics, but at this time is not possible due to legal, ethical, and resource-related constraints. Therefore, a survey method was chosen to conduct this study.

3.1.1. Survey Development

The questionnaires presented six simulated emergency scenarios and AI-generated suggestions detailing a treatment plan accordingly. The participants will state whether they agree or disagree with the suggested treatment and assess it across five dimensions (perceived accuracy, perceived comprehensibility, perceived thoroughness, perceived usefulness, and perceived cognitive effort; henceforth, they will be referred to without the "perceived", e.g., "perceived accuracy" will simply be "accuracy" and so on). A control condition (no AI) was included in this experiment to understand what treatment plans are used by EMS workers for these hypothetical scenarios. It would also provide a baseline comparison for the AI conditions results, allowing us to determine the impact of the inclusion of the AI suggestion and the AI explanation separately and to assess the AI suggestions. The cognitive effort was compared across the three conditions. Thus, the three conditions are:

1. Condition 1: no AI suggestion (control)
2. Condition 2: AI suggestion + no explanation (AI)

3. Condition 3: AI suggestion + explanation (XAI)

The six emergency scenarios were designed through consultation with a certified EMT. ChatGPT-4.0 was used to flesh these scenarios out to include additional details, and was once again validated through two EMTs in the first pilot round. They were used across all conditions and supplemented with an AI suggestion and an AI explanation according to the conditions. The AI suggestions and explanations were also generated using ChatGPT-4.0, however, were only reviewed by an EMT/paramedic to make sure that they were realistic enough without affecting its essence as an AI response. A table containing the full list of scenarios, suggestions, and explanations can be found in appendix A1.

When generating the prompts, I referred to [22] to understand what kind of content (or "intelligibility types") needs to be included in the explanation. I chose to include the "Inputs" and "Why" intelligibility types since my goal is to find and filter causes for the treatment (thus explaining the rationale behind the suggestion). I also found that including two intelligibility types provides enough information without becoming too long. The prompts used to generate the suggestions and explanations can be found in appendix A2.

The questionnaires were written in Hebrew to accommodate the local EMS workers. They were developed using Google forms. From this point on, the "AI suggestion + no explanation" condition will be referred to as the AI condition and the "AI suggestion + explanation" condition will be referred to as the XAI condition. Collectively, they will be referred to as the "AI conditions".

3.2. Hypotheses

3.2.1. Inclusion of Explanations

Based on the findings of the presented literature, I hypothesize that the inclusion of explanations will lead to:
(H1.1) higher ratings for perceived accuracy in the XAI condition than in the AI condition.
(H1.2) higher ratings for perceived comprehensibility in the XAI condition than in the AI condition.
(H1.3) higher ratings for perceived thoroughness in the XAI condition than in the AI condition.
(H1.4) higher ratings for perceived usefulness than in the XAI condition in the AI condition.
(H1.5) higher ratings for perceived cognitive effort in the XAI condition than in the AI condition.

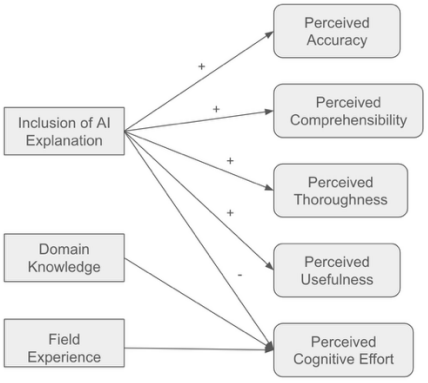

Figure 1. Inclusion of Explanations Hypotheses

3.2.2. Associations between the variables

Additionally, I hypothesize that the correlations between the variables will be:
(H2.1) positive between perceived accuracy and perceived comprehensibility.
(H2.2) positive between perceived accuracy and perceived thoroughness.
(H2.3) positive between perceived accuracy and perceived usefulness.
(H2.4) positive between perceived comprehensibility and perceived thoroughness.
(H2.5) positive between perceived comprehensibility and perceived usefulness.
(H2.6) positive between perceived thoroughness, and perceived usefulness.

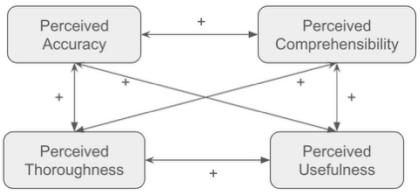

Figure 2.1. Associations between the variables Hypotheses h2.1-H2.6

(H2.7) negative between perceived cognitive effort and perceived accuracy, perceived comprehensibility, perceived thoroughness, and perceived usefulness.
(H2.8) negative between perceived cognitive effort and perceived accuracy, perceived comprehensibility, perceived thoroughness, and perceived usefulness.
(H2.9) negative between perceived cognitive effort and perceived accuracy, perceived comprehensibility, perceived thoroughness, and perceived usefulness.
(H2.10) negative between perceived cognitive effort and perceived accuracy, perceived comprehensibility, perceived thoroughness, and perceived usefulness.

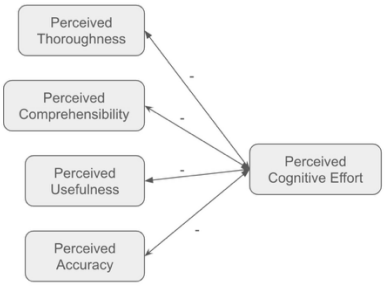

Figure 2.2. Associations between the variables Hypotheses H2.7-H2.10

3.2.3. Treatment plans

Lastly, I hypothesize that:
(H3) EMS worker-created treatment plans will be more precise than AI-generated treatment plans.

3.3. Participants

The participants were certified EMTs and paramedics currently active in the Israeli national emergency medical services MDA (Magen David Adom). Eligibility criteria included being at least 18 years of age, having a minimum of six months of field experience, and having completed either the basic EMT training course or a paramedicine degree or course. N = 8 participants were recruited via convenience sampling through the author's personal and professional network to participate in the study on a voluntary basis. Their years of experience ranged from three to eight years in EMS ($mean_{AI}$ =5.67, $median_{AI}$ =6; $mean_{XAI}$=4, $median_{XAI}$=3).

3.4. Measures

In this study, the "inclusion of an AI explanation" is the independent variable and the EMS participants' perceptions of accuracy, comprehensibility, thoroughness, and usefulness of the AI suggestions are the dependent variables. The perceived cognitive load experienced while coming up with the treatment plan is another dependent variable, however, we also need to acknowledge that domain knowledge and previous field experiences also can aid in easing the cognitive burden when thinking of a treatment plan [33, 34, 35]. When we encounter the same scenarios again and again, our reactions to them become automatic [36]. In this study, domain knowledge and field experience are not examined in particular and will be left to future

research. Additionally, they were asked to provide how many years they have volunteered and to what degree they use AI (in a multiple choice question with answers ranging from "I use AI very rarely" to "I use AI very frequently" with an option to choose "I do not use AI at all"). Table 3 presents our definitions of the dependent variables.

Table 1. Definitions of the dependent variables

Variable	Definition
Accuracy	The extent to which the AI's suggestion is perceived as correct and clinically appropriate for the scenario
Comprehensibility	The degree to which the AI's suggestion is clear and easy to understand
Throughness	The extent to which the AI's suggestion covers all key steps or details necessary for appropriate treatment
Usefulness	The perceived practical value of the AI's recommendation in supporting clinical decision-making
Cognitive Effort	The amount of mental effort participants feel is required to interpret and evaluate the AI's suggestion or think of a treatment plan for the scenario

3.5. Procedure

3.5.1. Pilot Tests

As stated, three rounds of pilot tests were conducted on four certified EMTs. To conduct effective pilot tests, cognitive interviews were conducted, from which the data was analyzed to revise the survey [23, 24]. The first pilot test was conducted on one EMT on the "control" version of the questionnaire and helped to identify inconsistencies in the scenarios, which were then adjusted in all questionnaires to make them more realistic with the help of the participant post-pilot. The second pilot test was conducted on one EMT on the "AI Suggestion + Explanation" questionnaire and brought unclear phrasing of the questions to light, which were then fixed to be clearer and more precise regarding the author's intentions. The "AI Suggestion + No Explanation" questionnaire was not piloted on due to it being identical to the "AI Suggestion + Explanation" version sans the explanations. The third and final pilot round was conducted on two EMTs and revealed unclear phrasing in the answers of the agreement or disagreement with the AI suggestion question, thus were rewritten to be clearer. These pilot tests helped to improve the internal validity of the study by ensuring that the questionnaire items measured each variable validly and reliably.

3.5.2. Participant Recruitment

After determining the questionnaire's validity, participants were recruited from the author's personal and professional circle and were contacted individually via Whatsapp. After obtaining their agreement to participate, the questionnaire was distributed to them. The participants were randomly assigned to one of the three experimental conditions (control = 2, AI suggestion + explanation = 3, AI suggestion + no explanation = 3). Overall, there were N=7 EMTs and N=1 paramedic.

3.5.3. Experimental Task Description

Three questionnaires were created for this study (one per condition) which contained six emergency scenarios (each one had its own section). This amount was sufficient to gather enough data of how the participants perceived the suggestions and how the explanations influenced their decisions, without causing the participants too much mental burden [31, 32].
A consent form was presented to them at the beginning of the questionnaire to inform them of their participant rights, what the purpose of the study was, and how data privacy was included in this study. They were then asked to provide the following information: years of experience, status (EMT/Paramedic), and how frequently they use AI. After obtaining their consent, they started the questionnaire.
In the AI conditions, each scenario was accompanied by an AI-generated suggestion for a treatment plan (and explanation if in the XAI condition). Swaroop et al. (2024) conducted a study examining the rate of over-reliance on AI suggestions under time pressure across four conditions (no AI, presenting the suggestion before they have time to make their own decision, presenting the AI after giving them time to make their own decision, and a mixed condition). They found that when time pressure wasn't involved, accuracy, response time, and overreliance on the AI suggestion were similar. Given that the time pressure factor is not included in this study, it was decided to present the AI suggestion (and explanation) before the participant has had the time to decide what the best treatment plan is, as it makes little difference.

They were then asked to what degree they agree with the suggestion: “full agreement”, “partial agreement”, “no agreement”, or “irrelevant suggestion”. If they did not agree fully, they were asked to provide an answer why. Then, they were asked to rate the suggestion across the five aforementioned dimensions (“accuracy”, "comprehensibility", “thoroughness”, “usefulness”, and “cognitive effort”) on 5-point likert scales. 5-point Likert scales were chosen because they allow us to quantify “behaviors” (and in our case, perceptions) in a scientifically valid way [25]. The last section included two optional open-ended questions: one asking if they had any thoughts or comments about the suggestions (and explanations) that the AI provided, and the other asking if they had general comments about future implementation of AI in EMS. The questions that appeared for both AI conditions are presented in table 1.

In the control condition, they were presented with each scenario, and were asked to provide a treatment plan for the scenario in an open-ended response format and then rate the “cognitive effort” they felt while thinking of the treatment plan on a 5-point likert scale. The inclusion of the “cognitive effort” question in the control condition as well would allow us to compare it across all conditions and see how the inclusion of suggestions and explanations impact cognitive effort. The last section included an optional open-ended question asking if they had general comments about future implementation of AI in EMS. The questions that appeared for the control condition are presented in table 2.

After completing the study, they were thanked for their time.

Table 2. The questions for each scenario in the AI Conditions

Question	Type	Possible Answers
Do you agree with the treatment proposed by the artificial intelligence tool?	Multiple Choice	1. I completely agree with the artificial intelligence tool’s recommendation (the proposed treatment is complete and correct, with no need for additional steps). 2. I partially agree with the artificial intelligence tool’s recommendation (some details/steps in the treatment are missing). 3. I do not agree with the artificial intelligence tool’s recommendation (the proposed treatment is completely incorrect). 4. I will not take the artificial intelligence tool’s recommendation into account.
If you selected partial agreement, disagreement, or chose not to follow the AI’s recommendation, why did you make that choice? What would you have done instead or in addition to the suggestion? (If you selected full agreement, please write “Not applicable.”)	Open-Ended	
How accurate did you think the artificial intelligence tool’s recommendation was?	5-point Likert Scale	Scale of 1 (to a very small extent) to 5 (to a very large extent)
How comprehensible did you think the artificial intelligence tool’s recommendation was?	5-point Likert Scale	Scale of 1 (to a very small extent) to 5 (to a very large extent)
How thorough did you think the artificial intelligence tool’s recommendation was (i.e., did it include all the important details/steps in the course of treatment)?	5-point Likert Scale	Scale of 1 (to a very small extent) to 5 (to a very large extent)
How useful did you think the artificial intelligence tool’s recommendation was?	5-point Likert Scale	Scale of 1 (to a very small extent) to 5 (to a very large extent)
How difficult did you find it to think of the appropriate treatment for this case?	5-point Likert Scale	Scale of 1 (to a very small extent) to 5 (to a very large extent)

Table 3. The questions asked per scenario in the control Condition

Question	Type	Possible Answers
How would you respond to this scenario? Please describe the steps you would take to treat the patient.	Open-Ended	
How difficult did you find it to think of the appropriate treatment for this case?	5-point Likert Scale	Scale of 1 (to a very small extent) to 5 (to a very large extent)

3.6. Data Analysis

Both quantitative and qualitative analyses were conducted in this study. The quantitative data consisted of Likert-type items for rating the suggestion across several variables and multiple-choice responses regarding agreement with the AI suggestion. The qualitative data included open-ended responses regarding treatment plans in the control condition and explanations regarding non-agreement with the AI suggestion. Additional information was collected regarding their years of experience in MADA and how frequently they use AI before the experiment began. All analyses were conducted using nonparametric statistical tests due to the ordinal nature of the Likert-type data and small sample size [26, 27]. A significance level of alpha = 0.1 was adopted for all tests to accommodate for the small sample size, and effect sizes and descriptive statistics were reported to provide context for statistical findings.

3.6.1. Suggestion Ratings

As mentioned, each AI suggestion for handling the emergency scenario was rated across five dimensions: accuracy, comprehensibility, thoroughness, usefulness, and cognitive effort. The analyses performed on this data include: descriptive statistics (median, mode, and frequencies of answers were found best for likert-type items [26]); a Chi-square test between the two AI conditions per variable; Mann-Whitney U test between the two AI conditions per variable; and a Kruskal-Wallis test between all conditions for the cognitive effort variable [26, 27]. The Mann-Whitney U test was also conducted for the cognitive effort variable pairwise after a significant result came from the Kruskal-Wallis test. Furthermore, Kendall's Tau-b tests were conducted between the variables per AI condition to test for associations between them.

3.6.2. Level of Agreement with AI Suggestions

For the analysis of the multiple choice question on level of agreement with the AI suggestion on how to handle the emergency scenario ("full agreement", "partial agreement", "no agreement", "irrelevant suggestion"), the data was treated as ordinal. Frequencies were calculated for all the ratings between the AI conditions. A chi-square test was conducted for each variable between the conditions. Kendall's Tau-b tests were conducted between the level of agreement and the dependent variables to investigate any associations between them.

3.6.3. AI Suggestion Non-agreement Explanations

If participants were to choose any option besides "full agreement" when asked about to what degree they agree with the AI suggestion for the treatment, they were asked to provide a further open-response explanation for this. These answers were then analyzed thematically by finding initial codes and then categorizing them into larger sub-categories and categories [28, 29, 30]. Then, chi-square tests were performed to find any associations between the themes found in the answers and the dependent variables.

3.6.4. Treatment Plans

There were two types of treatment plans in this study: AI-generated treatment plan suggestions per scenario and EMT-created treatment plans. These treatment plans were then analyzed thematically [30]. First, I went over the data several times to familiarize myself with it, then the treatment plans were analyzed and basic codes were extracted from them [28, 29]. This happened in two iterations for both types of treatment plans. After the codes were extracted, they were sorted into sub-categories, and from there into overall categories. These sub-categories and categories were revised several times. Then, descriptive statistics were calculated for the frequency of the categories and sub-categories in the treatment plans.

3.7. Ethical Considerations

An informed consent form was presented at the start of the questionnaire explaining the goal of the research project, the task they would have to complete, their rights as participants, and the data privacy protection strategies. No identifying information was collected from the participants and they were also assured that their answers were stored in an aggregated manner in a google drive folder accessible only to the researcher. Additionally, due to the AI suggestions detailing the advised treatment plan

for the simulated patients, a disclaimer was written into the informed consent explaining that these suggestions are by no means medical advice and are only to be regarded in the context of this research.

4. RESULTS

4.1. Suggestion Ratings

The frequencies were calculated per rating per dependent variable, such that: f_1 is the frequency of a rating of one and so on. Table 3 summarizes the results from the descriptive statistics analyses and statistical tests. Table 4 further summarizes the frequencies of the ratings across conditions for each variable. Figure 3 shows the percentage of each rating per condition and variable.

Accuracy. Participants gave the AI very high ratings (mode = 5.0, median = 5.0), while the XAI received more moderate scores (modes = 3.0/4.0, median = 3.0). In terms of rating frequencies, the ratings hovered around the 4.0-5.0 ratings in both conditions. Both the Mann–Whitney U test (U = 265.0, p = 0.0008) and the Chi-square test (X^2 = 16.033, p = 0.0029) confirmed a significant difference, with the AI condition receiving higher accuracy ratings and thereby disproving hypothesis H1.1.

Comprehensibility. The AI condition received very high ratings (mode = 5.0, median = 5.0), while the XAI condition was rated slightly lower (mode = 5.0, median = 4.0). In terms of rating frequencies, the ratings hovered around the 4.0-5.0 ratings in both conditions. Differences were statistically significant (U = 217.0, p = 0.0514; X^2 = 4.761, p = 0.0924), with the AI condition receiving higher comprehensibility ratings and thereby disproving hypothesis H1.2.

Thoroughness. The AI condition received higher ratings (mode = 5.0, median = 4.5) compared to XAI condition (modes = 3.0/4.0, median = 3.0). In terms of rating frequencies, the ratings hovered around the 4.0-5.0 ratings in both conditions. Both Mann–Whitney (U = 273.5, p = 0.0002) and Chi-square (X^2 = 15.5, p = 0.0037) results indicating a significant difference. Hypothesis H1.3 was therefore disproved.

Usefulness. Ratings for AI (mode = 5.0, median = 3.5) were more varied compared to XAI (mode = 3.0, median = 3.0). The discrepancy between these two measures in the AI condition indicates that the ratings are distributed between two “peaks” in the data. We can see this in the rating frequencies, which peaked at the 1.0 and 5.0 ratings in both conditions. In the XAI condition however, the ratings hovered around the 3.0 rating. While the Mann–Whitney U test showed no significant difference (U = 167.0, p = 0.8834), the Chi-square test indicated a significant association (X^2 = 15.336, p = 0.0041) between the two conditions. The AI condition received higher usefulness ratings and thereby disproving hypothesis H1.4.

Cognitive effort. The AI condition received consistently low ratings (mode = 1.0, median = 1.0), indicating that participants perceived minimal effort was required to process AI suggestions. In comparison, the XAI condition was rated slightly higher (mode = 2.0, median = 2.0), suggesting that participants felt suggestions demanded a bit more effort to assess with the addition of explanations. The control condition showed similarly low effort (mode=1.0, median = 1.5) meaning the treatment plans came to them fairly effortlessly. In terms of rating frequencies, the ratings hovered around the 1.0-2.0 ratings across all conditions. A Kruskal-Wallis test was conducted, and found H=9.621, p=0.0081. This indicates that there were significant differences between the groups, and to explore these differences further pairwise Mann-Whitney U tests were conducted. The tests revealed: between the control and AI conditions U=64.5 and p=0.0009; between the AI and XAI conditions U=125.0 and p=0.4612; and between the control and XAI conditions U=64.5 and p=0.0009. Therefore, we can see that there is only a significant difference in the distribution of values between the control and AI conditions. The Chi-square test found X^2=17.552, p=0.0074, indicating a positive correlation between all conditions. Further pairwise tests between the conditions revealed: between the control and AI conditions U=9.513 and p=0.0231; between the AI and XAI conditions U=12.192 and p=0.0067; and between the control and XAI conditions U=6.527 and p=0.0885. Therefore, we can see that there are statistically significant positive correlations between all conditions. Hypothesis H1.5 was therefore proven true.

Table 3. Results for descriptive statistics and statistical tests for all dependent variables (accuracy, comprehensibility, thoroughness, usefulness, and cognitive effort).

Measure	Accuracy	Comprehensibility	Thoroughness	Usefulness	Cognitive Effort
Mode (AI)	5.0	5.0	5.0	5.0	1.0
Mode (XAI)	3.0/4.0	5.0	3.0/4.0	3.0	2.0
Mode (Control)	-	-	-	-	1.0

Median (AI)	5.0	5.0	4.5	3.5	1.0
Median (XAI)	3.0	4.0	3.0	3.0	2.0
Median (Control)	-	-	-	-	1.5
Mann-Whitney U (AI and XAI)	U = 265.0 p = 0.0008	U = 217.0 p = 0.0514	U = 273.5 p = 0.0002	U = 167.0 p = 0.8834	See results
Chi-Square (AI and XAI)	X^2 = 16.033 p = 0.0029	X^2 = 4.761 p = 0.0924	X^2 = 15.5 p = 0.0037	X^2 = 15.336 p = 0.0041	See results

Table 4. Suggestion ratings frequencies

Variable/Rating	Condition	1	2	3	4	5
Accuracy	AI	0	2	1	5	10
	XAI	0	1	3	7	7
Comprehensibility	AI	0	0	1	4	13
	XAI	0	0	6	4	8
Thoroughness	AI	0	1	1	7	9
	XAI	0	1	3	7	7
Usefulness	AI	6	0	3	2	7
	XAI	0	3	8	0	2
Cognitive Effort	AI	2	11	4	1	0
	XAI	6	1	4	1	0
	Control	12	5	1	0	0

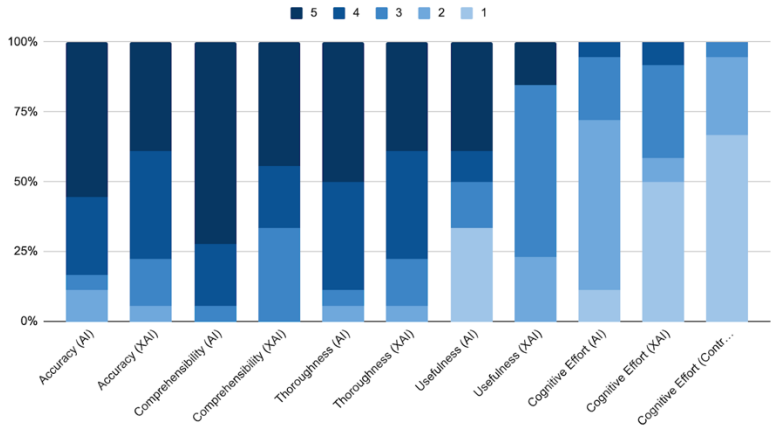

Figure 3. Suggestion ratings frequencies by percentage

4.1.1. Kendall's Tau-b between the Variables

Kendall's Tau-b tests were conducted between the variables to identify any associations between them. They were conducted within the AI and XAI conditions. Results are shown in table 5. From these results, we can see more associations between the variables in the XAI condition than in the AI condition.

Table 5. Kendall's Tau-b test results between variables

Condition	Variable 1	Variable 2	τ	p	Association
AI	Accuracy	Comprehensibility	0.721	0.0013	Strong positive association
	Accuracy	Thoroughness	0.5803	0.0084	Strong positive association

	Accuracy	Usefulness	0.3725	0.0804	Moderate positive association
	Accuracy	Cognitive Effort	-0.266	0.2363	Statistically insignificant
	Comprehensibility	Thoroughness	0.258	0.2567	Statistically insignificant
	Comprehensibility	Usefulness	0.26	0.2365	Statistically insignificant
	Comprehensibility	Cognitive Effort	-0.192	0.4082	Statistically insignificant
	Thoroughness	Usefulness	0.499	0.0205	Statistically insignificant
	Thoroughness	Cognitive Effort	-0.139	0.5392	Statistically insignificant
	Usefulness	Cognitive Effort	0.268	0.2229	Statistically insignificant
XAI	Accuracy	Comprehensibility	0.358	0.096	Moderate positive association
	Accuracy	Thoroughness	0.907	2.0616e-05	Strong positive association
	Accuracy	Usefulness	0.648	0.0021	Strong positive association
	Accuracy	Cognitive Effort	-0.393	0.0684	Moderate negative association
	Comprehensibility	Thoroughness	0.245	0.2548	Statistically insignificant
	Comprehensibility	Usefulness	0.269	0.2059	Statistically insignificant
	Comprehensibility	Cognitive Effort	-0.071	0.7416	Statistically insignificant
	Thoroughness	Usefulness	0.648	0.0021	Strong positive association
	Thoroughness	Cognitive Effort	-0.474	0.028	Moderate negative association
	Usefulness	Cognitive Effort	-0.547	0.0104	Strong negative association

In the AI condition, several significant associations were found between the variables. Accuracy was strongly and positively associated with both comprehensibility ($\tau = 0.721$, $p = 0.0013$) and thoroughness ($\tau = 0.5803$, $p = 0.0084$), indicating that when participants rated the AI as more accurate, they also tended to evaluate it as more comprehensible and more thorough. A moderate positive association was also observed between accuracy and usefulness ($\tau = 0.3725$, $p = 0.0804$), suggesting that higher accuracy ratings were moderately associated with higher usefulness ratings. All other relationships, including those involving cognitive effort, were statistically insignificant.

By contrast, the XAI condition revealed a broader pattern of significant associations between the variables. Accuracy was moderately and positively related to comprehensibility ($\tau = 0.358$, $p = 0.096$), and strongly and significantly related to usefulness ($\tau = 0.648$, $p = 0.0021$). Thoroughness also showed a strong positive association with usefulness ($\tau = 0.648$, $p = 0.0021$). Importantly, several significant negative associations were found with cognitive effort. Specifically, accuracy and cognitive effort displayed a moderate negative association ($\tau = -0.393$, $p = 0.0684$), thoroughness was moderately and negatively related to cognitive effort ($\tau = -0.474$, $p = 0.028$), and usefulness showed a strong negative association with cognitive effort ($\tau = -0.547$, $p = 0.0104$). These results indicate that participants who rated the explanations as more accurate, thorough, or useful also felt they required less effort to process them. Other relationships were not statistically significant.

Therefore, hypotheses H2.1, H2.2, and H2.3 were proven right across both conditions. Hypotheses H2.4-H2.10 need to be rejected, even though there were stronger associations in the XAI condition.

4.2. Level of Agreement with Suggestion

Since the suggestion level of agreement question was multiple choice, the answers were codified to allow for analysis such that: 1 = irrelevant suggestion; 2 = no agreement; 3 = partial agreement; and 4 = full agreement. Table 6 presents the results for the descriptive statistics and statistical tests.

As we can see from the frequencies, mode and median, the “irrelevant suggestion” and “no agreement” options were not selected at all, whereas the “partial agreement” was chosen the most. The Chi-square test between the conditions showed insignificant results with $X^2=0.125$, $p=0.7236$. This indicates that there is no difference in the values distribution between the two conditions.

Table 6. Descriptive statistics and statistical tests' results for level of agreement

Measure	Level of Agreement
Mode (AI)	3.0
Mode (XAI)	3.0
Median (AI)	3.0
Median (XAI)	3.0
Frequencies (AI)	f_1=0, f_2=0, f_3=13, f_4=5
Frequencies (XAI)	f_1=0, f_2=0, f_3=11, f_4=7
Chi-Square (AI and XAI)	$X^2 = 0.125$
	p = 0.7236

4.2.1. Associations with Variables

Kendall's Tau-b tests were conducted across both conditions between the variables and the level of agreement with the AI suggestion. Results are presented in table 7.

Table 7. Results from Kendall's Tau-b between "level of agreement" and the dependent variables

Condition	Variable	τ	p	Association
AI	Accuracy	0.314	0.1704	Statistically insignificant
	Comprehensibility	0.119	0.6148	Statistically insignificant
	Thoroughness	0.569	0.0142	Strong positive association
	Usefulness	0.536	0.0169	Strong positive association
	Cognitive Effort	-0.183	0.4385	Statistically insignificant
XAI	Accuracy	0.844	0.0002	Strong positive association
	Comprehensibility	0.514	0.0252	Strong positive association
	Thoroughness	0.844	0.0002	Strong positive association
	Usefulness	0.616	0.0061	Strong positive association
	Cognitive Effort	-0.406	0.0776	Moderate negative association

As we can see, the AI condition only has two significant associations between the level of agreement and thoroughness and between level of agreement and usefulness (both strong positive associations). In the XAI condition, on the other hand, we can see significant associations between the level of agreement and all variables (all of them being strong positive associations and cognitive effort being a moderate negative association).

4.3. Explanations for Non-agreement

After analyzing the explanations for non-agreement, we found six themes across both conditions, summarized in table 8.

Table 8. Themes from explanations for non-agreement with AI suggestion for treatment

Theme	Definition	Number of Occurrences (AI)	Number of Occurrences (XAI)
no alternative actions suggested	No alternative actions provided in case variables change or unexpected factors arise in the scenario	3	2
lacking or unclear information on how to act	Omitted or unclear information regarding actions that EMS personnel would need to know or perform in this situation	9	6
incorrect information about how to act	Presented incorrect information or suggested inappropriate actions	2	0
inaccurate information about how to act	Presented inaccurate information or actions that would need further refinement	4	5
unsure of suggestion	Reflected low confidence in the suggestion itself	1	0
made assumptions	Made non-obvious and potentially mistaken assumptions regarding the scenario	0	1
Total		19	14

4.3.1. Thematic Analysis Results

We can see that the overall number of issues presented in the explanations are lower in the XAI condition than the AI condition. In almost all themes (except for "inaccurate information about how to act" and "made assumptions"), the AI condition has more occurrences than the XAI condition. We can also see that for both conditions, the most-occurring theme was the "lacking or unclear information on how to act" which indicates that the AI suggestion doesn't provide all the information that EMTs deem necessary to treat patients. "Inaccurate information about how to act" was also prominent in both conditions, meaning the AI suggestion included unnecessary or misplaced information or actions for the treatment.

4.3.2. Correlation with Suggestion Ratings

Correlations between the suggestion ratings and the explanations for non-agreement were investigated via chi-square test. The results indicate statistically insignificant correlations for the AI condition and largely statistically insignificant correlations for the XAI condition except for three correlations that were found. These correlations are between "lacking or unclear information on how to act" and each of the three variables: comprehensibility (X^2=8.0, p=0.0915); thoroughness (X^2=8.0, p=0.0915); and usefulness (X^2=6.0, p=0.0497). Therefore, we can conclude that participants who felt the explanations were lacking or unclear were more likely to also rate the XAI suggestions as less comprehensible, useful, and thorough.

4.4. Treatment Plans

The results of the thematic analysis are presented in tables. Table 9 presents the identified categories and their definitions and table 10 presents the frequency in which these themes appeared in the treatment plans. The full codebook can be found in Appendix A3.

Table 9. Definitions of the categories and sub-categories

Category	Sub-categories	Definition
Safety	-	Actions to protect both the patient and EMS personnel, including clearing the area, preventing delays in evacuation, and ensuring safe working conditions.
Assessments	Primary Survey Assessment, Secondary Survey Assessment, Neurological Assessment, Physical Survey Assessment	Evaluations of the patient's condition, ranging from primary surveys (ABC checks, consciousness) to secondary surveys (history, pre-existing conditions), plus neurological and physical surveys to detect underlying issues or injuries.
Patient Symptoms	-	Observable or reported medical signs and conditions experienced by the patient, such as fainting, vomiting, hypoglycemia, confusion, or convulsions, which inform EMS assessment and treatment.
EMS Interventions	Positioning and Stabilization, Resuscitation and Life Support, Protective Measures, Medicinal and Nutritional Administration, Breathing Support, Care Approach, Situational Criteria for Certain Interventions	Direct medical actions to stabilize or treat the patient, including positioning and stabilization, resuscitation, protective measures, administration of substances, breathing support, protocol-based care, and conditional instructions.

Communication	Patient, Bystanders, EMS Team	Information exchange with patients, bystanders, or EMS teams, including assessment questions, requests for assistance, and reporting patient status to dispatch or hospitals.
Tests/Measurements	Vital Signs Tests, Other Standard Tests	Objective physiological checks, including vital signs (pulse, oxygen saturation, blood pressure) and standard tests (blood glucose, temperature, respiratory rate).
Evacuation	-	Procedures for removing and transporting patients, including transfer to emergency vehicles, general evacuation, and helicopter dispatch.

Table 10. Summary of categories and sub-categories from coded data (codebook included in appendix A3)

Theme	**Sub-category**	**Sub-category Frequency**		**Theme Frequency**	
		Control	AI Suggestion	Control	AI Suggestion
Safety	-	-	-	15	1
Assessments	Primary Survey Assessment	28	4	56	9
	Secondary Survey Assessment	10	0		
	Neurological Assessment	11	5		
	Physical Survey Assessment	7	0		
Patient Symptoms	-	-	-	10	3
EMS Interventions	Positioning and Stabilization	8	7	46	36
	Resuscitation and Life Support	3	3		
	Protective Measures	3	2		
	Medicinal and Nutritional Administration	5	5		
	Breathing Support	9	10		
	Care Approach	5	7		
	Situational Criteria for Certain Interventions	13	2		
Communication	Patient	5	0	20	2
	Bystanders	13	0		
	EMS Team	2	2		
Tests/Measurements	Vital Signs Tests	29	8	40	17
	Other Standard Tests	11	9		
Evacuation	-	-	-	2	2
Total	-	-	-	189	70

Across both types of treatment plans, N=75 codes were extracted. N=64 distinct codes were extracted from the control condition questionnaire, from the answers of N=2 participants which gives us 6 scenarios x 2 plans per scenario = 12 plans. N=36 distinct codes were extracted from the AI-generated treatment plans. Overall, N=39 unique codes were extracted from the control condition questionnaire, N=11 unique codes were extracted from the AI-generated suggestions, and N=25 unique codes were found in both. Overall, N=189 occurrences of the codes were present in the control condition questionnaire (mean=15.75 codes per answer) and N=70 occurrences of the codes in the AI-generated suggestions (mean=11.667 per suggestion). This is a deficit of 119 codes overall, and a deficit of 4.083 codes per suggestion on average. Therefore, we can conclude that hypothesis H3 was proven correct.

4.5. Comments and Thoughts on AI in EMS

The participants expressed skepticism about the AI suggestions' usefulness and necessity. Some participants questioned the role of AI in EMS, noting: "In truth, I don't see the need for this... why do people study medicine for years just to get help from AI." Another felt that the explanations provided little added value for experienced personnel: "They also aren't very useful as I already know what needs to be done for the treatment."

One emphasized the need for clearer and more structured guidance within the system: "There is a need to explain in a more accurate manner and maybe write what to put emphasis on at each step before moving on to the next step." They stressed that the suggestions and explanations should be simplified and accessible, particularly for less experienced staff: "The explanations need to be as simple as possible and not include too many technical terms—it can confuse novices like the teenage volunteers."

Regarding AI in EMS more broadly, participants expressed doubts about its role in critical decision-making. As one participant put it: "For decision-making, I don't think [it will be implemented]. There are some things that a person needs to know how to do alone, especially if it's regarding saving a human life. We can't know what will go wrong with AI. Knowledge isn't the only way—there is also a need for experience, and AI cannot decide what will happen with a human life."

5. DISCUSSION

The purpose of this study was to examine how accompanying AI suggestions for patient treatment in various emergency scenarios with explanations would influence perceptions of accuracy, comprehensibility, thoroughness, usefulness, and cognitive effort.

We found that the AI condition generally rated better across the variables. Perhaps surprisingly, we saw stronger associations between these variables in the XAI condition than in the AI condition. We also saw that the most frequently chosen option for level of agreement was "partial agreement" with "no agreement" and "irrelevant suggestion" never having been chosen in both AI conditions. This indicates that the suggestions for treatment were adequate, yet lacking in some aspects. A further analysis revealed that the most lacking aspects were lacking or unclear information on how to act and inaccurate information about how to act. Additionally, we also saw stronger associations between "level of agreement" and our dependent variables in the XAI condition than in the AI condition.

Control-condition participants' treatment plans contained substantially more codes compared to AI-generated suggestions. This shortfall reflects the issues raised in participants' non-agreement explanations, where AI suggestions were frequently described as lacking detail and omitting essential information needed for effective patient care or even giving out inaccurate suggestions to handle the scenario. On the other hand, the statistical measures of the cognitive effort variable across the AI and control conditions are nearly the same, therefore one might wonder if including more codes to potentially increase the perceived accuracy would have on the perceived cognitive effort.

Finally, participants were generally skeptical about the necessity of AI in EMS, stressing that trained professionals already know the required procedures. While some found the explanations confusing or redundant for experienced personnel, there was a strong emphasis on the need for simple, structured, and accurate guidance in the suggestions. Broader concerns also emerged regarding AI's role in life-and-death decision-making, with participants expressing greater trust in human expertise and experience.

Technology designers for EMS should recognize that explanations accompanying AI suggestions do not automatically enhance perceptions of accuracy, usefulness, or clarity, and may even increase perceived cognitive effort if presented with unnecessary complexity. Explanations must therefore be designed with brevity and precision, avoiding technical jargon and redundant details

that EMS personnel might find confusing or unnecessary. At the same time, explanations should be tailored to the user's level of expertise: novice EMS staff may benefit from more structured and step-by-step guidance, while seasoned paramedics and EMTs require only concise cues that support rapid decision-making under pressure. Above all, AI support should aim to reduce cognitive burden rather than add to it, by providing actionable, context-relevant information without overwhelming the user with excess text.

To summarize the results of this study, we found that the explanations did not positively affect perceptions of accuracy, comprehensibility, thoroughness, usefulness, and cognitive effort regarding the suggestions. Stronger associations between them could indicate that it helped participants think more critically about the suggestion and make more consistent and structured judgments. However, less issues were pointed out in the explanations to non-agreement in the XAI condition than in the AI condition. Therefore, we cannot confidently conclude whether XAI improves perceptions of AI suggestions. This is inline with findings from Rosenbacke et al. (2024), Morandini et al. (2023) and Laxar et al. (2023) who also found mixed results regarding the impact of XAI on trust. However, the literature regarding the true effectiveness of explainability is still underdeveloped [19].

5.1. Limitations and Future Work

This study has several limitations. The first being the sample size; at N=8 participants the statistical inferences are weaker than with larger sample sizes. While the statistical tests were chosen with the sample size in mind so that the results are completely inaccurate, it is still a fundamental limitation. A smaller sample size also increases the chances of type II errors. By setting $\alpha = 0.1$, we increased the risk of Type I error, which was a deliberate tradeoff to reduce the risk of Type II errors given our small sample size [38].

The scenarios themselves might not be representative of real-world emergency scenarios as they are typically highly dynamic and contain a lot of unpredictable variables. This is despite verifying the scenarios through two certified EMTs and co-creating them with one. In future work more realistic cases can be used as the scenarios to provide a more realistic scenario, thus resulting in more accurate treatment plans. Additionally, it would be interesting to conduct a similar study on paramedics as they tend to deal with more uncertainty as they have more legibility to administer more complex treatments and make more complex decisions than EMTs [9]. Thus, more complex scenarios can also be presented for them to solve.

Finally, this specific method of decision-support - providing real-time suggestions on how to handle emergency scenarios - might not be desirable by EMTs and paramedics in general. Other methods need to be tested under different circumstances to understand where exactly they need support and what their pain points are. As the comments at the end indicated when asked about the suggestions, explanations and the future of AI in EMS, they didn't see the need for this kind of "help" from AI.

Future work could introduce time pressure into the experiment and examine how the EMS participants' perceptions of accuracy, comprehensibility, usefulness, and cognitive load change in accordance. This could present more realistic circumstances for the participants as time pressure is a big factor in their work. Building on the work of Swaroop et al. (2024), we could also examine the effects of presenting the AI suggestions with the explanations in three conditions: before the participant has had the chance to make an independent decision, after the participant has had the chance to make an independent decision, and a mix of both of these conditions. This can inform us whether they tend to rely on the AI suggestions provided or not provided explanations.

6. CONCLUSION

This study set out to examine whether explanations accompanying AI-generated treatment suggestions improve EMS personnel's perceptions of accuracy, comprehensibility, thoroughness, usefulness, and cognitive effort. Contrary to expectations, the AI-only condition received higher ratings across most measures, while explanations did not yield significant improvements. However, the XAI condition showed stronger associations between variables and fewer issues in non-agreement responses, suggesting that explanations may support more structured reasoning even if they do not enhance overall perceptions. Treatment plan analysis confirmed that human-created responses were more comprehensive than AI-generated ones, highlighting current limitations of AI in capturing the detail and nuance of EMS practice. Qualitative comments further emphasized skepticism toward AI's role in life-and-death decision-making and stressed the importance of tailoring explanations to users' expertise levels.

Overall, the findings suggest that explanations must be designed with precision, brevity, and contextual relevance to avoid adding cognitive effort or confusion. While explainability remains a crucial element for trust in AI, it is not a guarantee of acceptance or perceived value. Future work should investigate larger and more diverse EMS samples, more realistic and complex scenarios, and the impact of contextual factors such as time pressure. Designing AI systems that truly support EMS professionals will require balancing transparency with efficiency, ensuring that decision-support tools enhance rather than hinder critical care delivery.

7. REFERENCES

[1] Ahmed Ali Alghamdi, Saeed Ali Alkatheri, Abdulrahman Jameel Aljohani, Fahad Ali Madkhali, Ayman Jamaan Alomary, Attieh Yahya Alzahrani, Sharaf Ahmed Muhammad Al Munimi, Sultan Ateq Abdullmain Albeshri, Waleed Ghoneim Mohammed Al-Jahdali, and Saber Obaid Alotaibi. 2024. Artificial intelligence in prehospital emergency care: A literature review. Neuropsychopharmacologia Hungarica, 22(3), 384–393. https://doi.org/10.1556/2066.2024.00034K.

[2] Badawi and A. J. Nashwan, "The future of prehospital emergency care: Embracing AI applications in ambulance services," Int. Emerg. Nurs., vol. 72, p. 101385, 2024, doi: 10.1016/j.ienj.2023.101385.

[3] Han Shi Jocelyn Chew and Palakorn Achananuparp. 2022. Perceptions and needs of artificial intelligence in health care to increase adoption: Scoping review. Journal of Medical Internet Research, 24(1), e32939. https://doi.org/10.2196/32939

[4] Abirami Kirubarajan, Ahmed Taher, Shawn Khan, and Sameer Masood. 2020. Artificial intelligence in emergency medicine: A scoping review. In JACEP Open, 1(6), 1691–1702. https://doi.org/10.1002/emp2.12277

[5] Lucy Shinners, Sandra Grace, Stuart Smith, Alexandre Stephens, and Christina Aggar. 2022. Exploring healthcare professionals' perceptions of artificial intelligence: Piloting the Shinners Artificial Intelligence Perception tool. In Digital Health, 8, 1–8. https://doi.org/10.1177/20552076221078110Stai et al., "Public perceptions of artificial intelligence and robotics in medicine," J. Endourol., vol. 34, no. 10, pp. 1041–1048, 2020, doi: 10.1089/end.2020.0137.

[6] Bethany Stai, Nick Heller, Sean McSweeney, Jack Rickman, Paul Blake, Ranveer Vasdev, Zach Edgerton, Resha Tejpaul, Matt Peterson, Joel Rosenberg, Arveen Kalapara, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. 2020. Public perceptions of artificial intelligence and robotics in medicine. In Journal of Endourology, 34(10), 1041–1048. https://doi.org/10.1089/end.2020.0137

[7] David Wallstén, Gregory Axton, Eunji Lee, Anna Bakidou, Bengt Arne Sjöqvist, and Stefan Candefjord. 2023. Design for integrating explainable AI for dynamic risk prediction in prehospital IT systems. In Artificial Intelligence, Social Computing and Wearable Technologies, 113, 268–278. https://doi.org/10.54941/ahfe1004199T.

[8] Tiago Araujo, Ana Duboc, and Raian Ali. 2020. In AI we trust: Perceptions about automated decision-making by artificial intelligence. In AI & Society, 35, 611–623. https://doi.org/10.1007/s00146-019-00931-w.

[9] Hana Harenčárová. 2017. Managing uncertainty in paramedics' decision making. In Journal of Cognitive Engineering and Decision Making, 11(1), 42–62. https://doi.org/10.1177/1555343416674814

[10] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2024. Accuracy-time tradeoffs in AI-assisted decision making under time pressure. In Proceedings of the 29th International Conference on Intelligent User Interfaces. ACM, New York, NY, 138–154. https://doi.org/10.1145/3626420.3628544

[11] David B. Barr, Mary P. LaVenture, Susan J. Zahner, Linda L. Bersick, and Daniel M. Sosin. 2010. Gap assessment in the emergency response community: Findings and recommendations. In Public Health Reports, 125(5), 756–763. https://doi.org/10.1177/003335491012500516

[12] Mojtaba Bijani, Zahra Khaleghi, Abbas Ebadi, Mostafa Bijani, and Hassan Moosavi. 2021. Major challenges and barriers in clinical decision-making as perceived by emergency medical services personnel: A qualitative study. BMC Emergency Medicine, 21(1), 126. https://doi.org/10.1186/s12873-021-00501-9

[13] Sharon Lawn, Louise Roberts, Eileen Willis, Leah Couzner, Leila Mohammadi, and Elizabeth Goble. 2020. The effects of emergency medical service work on the psychological, physical, and social well-being of ambulance personnel: A systematic review of qualitative research. BMC Psychiatry, 20(1), 348. https://doi.org/10.1186/s12888-020-02752-4

[14] Martin Sedlár. 2020. Cognitive skills of emergency medical services crew members: A literature review. BMC Emergency Medicine, 20(1), 44. https://doi.org/10.1186/s12873-020-00330-1

[15] Martin Sedlár and Zuzana Kaššaiová. 2022. Markers of cognitive skills important for team leaders in emergency medical services: A qualitative interview study. BMC Emergency Medicine, 22(1), 80. https://doi.org/10.1186/s12873-022-00629-1

[16] Anna Poranen, Anne Kouvonen, and Hilla Nordquist. 2022. Perceived human factors from the perspective of paramedics – a qualitative interview study. BMC Emergency Medicine, 22(1), 178. https://doi.org/10.1186/s12873-022-00738-x

[17] SwissCognitive. 2023. Life on the front lines: How AI is reimagining emergency medical services. SwissCognitive – The Global AI Hub. June 27, 2023. https://swisscognitive.ch/2023/06/27/life-on-the-front-lines-how-ai-is-reimaging-emergency-medical-services/

[18] Magen David Adom. 2023. MDA to launch new AI-based system for emergency medical services. Magen David Adom News. March 23, 2023. https://www.mdais.org/news/23032301

[19] Hajo Hildt. 2025. What is the role of explainability in medical artificial intelligence? A case-based approach. AI and Ethics, 5, 1–12. https://doi.org/10.1007/s43681-025-00471-9

[20] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312

[21] Yohei Okada, Yilin Ning, and Marcus Eng Hock Ong. 2023. Explainable artificial intelligence in emergency medicine: An overview. Clinical and Experimental Emergency Medicine, 10(4), 354–362. https://doi.org/10.15441/ceem.23.145

[22] Brian Y. Lim, Qian Yang, Ashraf Abdul, and Danding Wang. 2019. Why these explanations? Selecting intelligibility types for explanation goals. In Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019, 7 pages. https://doi.org/10.1145/1234567890

[23] Don A. Dillman, Jolene D. Smyth, and Leah Melani Christian. 2016. Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method. John Wiley & Sons, Hoboken, NJ.

[24] Jason M. Etchegaray and Wayne G. Fischer. 2011. Understanding evidence-based research methods: Pilot testing surveys. In HERD: Health Environments Research & Design Journal, 4(4), 143–147. https://doi.org/10.1177/193758671100400411

[25] Ankur Joshi, Saket Kale, Satish Chandel, and Dinesh Kumar Pal. 2015. Likert scale: Explored and explained. In British Journal of Applied Science & Technology, 7(4), 396–403. https://doi.org/10.9734/BJAST/2015/14975

[26] Harry N. Boone Jr. and Deborah A. Boone. 2012. Analyzing Likert data. In Journal of Extension, 50(2), Article 48. https://doi.org/10.34068/joe.50.02.48

[27] Spencer E. Harpe. 2015. How to analyze Likert and other rating scale data. In Currents in Pharmacy Teaching and Learning, 7(6), 836–850. https://doi.org/10.1016/j.cptl.2015.08.001

[28] Victoria Elliott. 2018. Thinking about the coding process in qualitative data analysis. In The Qualitative Report, 23(11), 2850–2861. https://nsuworks.nova.edu/tqr/vol23/iss11/14

[29] Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three approaches to qualitative content analysis. In Qualitative Health Research, 15(9), 1277–1288. https://doi.org/10.1177/1049732305276687

[30] Mohammed Ibrahim Alhojailan. 2012. Thematic analysis: A critical review of its process and evaluation. In West East Journal of Social Sciences, 1(1), 39–47.

[31] Hamed Taherdoost. 2022. Designing a questionnaire for a research paper: A comprehensive guide to design and develop an effective questionnaire. In Asian Journal of Managerial Science, 11(1), 8–16. https://doi.org/10.51983/ajms-2022.11.1.3087

[32] Wai-Ching Leung. 2001. How to design a questionnaire. In Student BMJ, 9, 187–190. https://www.studentbmj.com/back_issues/0601/education/187.html

[33] Janet L. Kolodner. 1982. The role of experience in development of expertise. In Proceedings of the National Conference on Artificial Intelligence (AAAI-82), Pittsburgh, PA, 295–297. AAAI Press.

[34] Jan Breckwoldt, Sebastian Klemstein, Bergit Brunne, Luise Schnitzer, Hans-Richard Arntz, and Hans-Christian Mochmann. 2012. Expertise in prehospital endotracheal intubation by emergency medicine physicians—Comparing 'proficient performers' and 'experts'. In Resuscitation, 83(4), 434–439. https://doi.org/10.1016/j.resuscitation.2011.10.011

[35] Patricia A. Alexander, Julianne L. Schallert, and Vicky C. Hare. 1989. Coming to terms: How researchers in learning and literacy talk about knowledge. In Review of Educational Research, 59(3), 315–342. https://doi.org/10.3102/00346543059003315

[36] Gordon D. Logan. 1985. Skill and automaticity: Relations, implications, and future directions. In Canadian Journal of Psychology, 39(2), 367–386. https://doi.org/10.1037/h0080066

[37] Nicolas Laxar, Philipp Burckhardt, Lorenz Harer, and Ulrich Sax. 2023. The influence of explainable vs non-explainable clinical decision support systems on rapid triage decisions: A simulation study. In Frontiers in Artificial Intelligence, 6, 1173848. https://doi.org/10.3389/frai.2023.1173848

[38] Anthony K. Akobeng. 2016. Understanding type I and type II errors, statistical power and sample size. In Acta Paediatrica, 105(6), 605–609. https://doi.org/10.1111/apa.13384

[39] Sofia Morandini, Federico Fraboni, Gabriele Puzzo, Davide Giusino, Lucia Volpi, Hannah Brendel, Enzo Balatti, Marco De Angelis, Andrea De Cesarei, and Luca Pietrantoni. 2023. Examining the nexus between explainability of AI systems and users' trust: A preliminary scoping review. In Proceedings of the 1st World Conference on eXplainable Artificial Intelligence (xAI 2023), Lisbon, Portugal, July 26–28. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-3445/paper7.pdf

[40] Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. 2024. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. In JMIR AI, 3, e53207. https://doi.org/10.2196/53207

[41] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. In International Journal of Human–Computer Studies, 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

8. APPENDIX

8.1. Appendix A1: Scenarios, AI Suggestions, and AI Explanations in the questionnaire (translated to English from Hebrew)

	Scenario	Suggestion	Explanation
1	A 17-year-old girl participated in her school graduation ceremony on a hot summer day. While standing on the stage, she suddenly appeared very pale, then collapsed backward and lost consciousness for a few seconds. Witnesses at the scene report that she did not jump or trip on anything, but simply "fell all at once." She is now lying on the ground, breathing, but appears confused. After about a minute, she opens her eyes.	The patient should be laid on her back, any tight clothing should be loosened, and airway, breathing, and pulse should be checked. Oxygen should be provided if necessary. Her legs should be elevated to improve venous return, and vital signs should be monitored. After initial recovery, medical supervision should continue, and other potential causes of fainting (such as dehydration, heat, or cardiac issues) should be ruled out.	The key findings were a brief loss of consciousness while standing in the heat, pallor, preserved breathing and pulse, and rapid recovery with temporary confusion—indications consistent with syncope. When consciousness returns with normal breathing and pulse, the patient should be stabilized, venous return improved by elevating the legs, oxygen provided if needed, and vital signs monitored. This approach balances immediate supportive care in the field with the need for continued medical evaluation to rule out life-threatening causes.
2	During a morning shift, an ambulance team receives a call about a man, approximately 30 years old, who collapsed at a café in the city center due to seizures. Upon arrival at the scene, the team finds the patient lying on the floor, unresponsive to verbal stimuli. One of the bystanders reports that the seizures	Dangerous objects should be removed from the area around the patient, the airway should be secured, and oxygen provided if necessary. During the seizure, the patient should not be forcibly restrained; instead, protect the head and monitor vital signs. Once the seizure ends,	The key findings included evidence of an active seizure that began approximately two minutes earlier, lack of response to verbal stimuli, and no known history of similar events—indications pointing to a first-time seizure. The recommended actions

	began about two minutes ago and that this is the first time he has seen the man in such a condition.	place the patient in the recovery position, assess their level of consciousness and breathing, and prepare to administer medication or arrange urgent transport according to medical guidelines.	were selected to prevent further injury (removing nearby objects, protecting the head), maintain a clear airway, and support breathing with oxygen. Placing the patient in the recovery position after the seizure is intended to reduce the risk of aspiration and enable assessment of consciousness and breathing for continued treatment.
3	An ambulance team is dispatched to an apartment following a report of a 75-year-old man found unconscious. Upon arrival, the patient is found lying on the floor, unresponsive to verbal or painful stimuli, with white foam around his mouth and deep, rapid breathing. According to his roommate, the patient has type 1 diabetes and had complained earlier that morning of nausea, abdominal pain, and extreme thirst. It is unknown whether he had eaten or drunk anything.	An open airway must be ensured, oxygen should be administered, and vital signs—including blood glucose—should be monitored. If hyperglycemia is detected and diabetic ketoacidosis (DKA) is suspected, initiate intravenous fluid administration under physician guidance and prepare for urgent hospital transport. Continuous close monitoring is required throughout evacuation due to the risk of deterioration in consciousness or respiratory status.	The key findings included unconsciousness, deep and rapid breathing, frothing at the mouth, a history of type 1 diabetes, and preceding symptoms such as nausea, abdominal pain, and extreme thirst—all pointing to suspected diabetic ketoacidosis (DKA). The recommended actions were selected to maintain airway and respiratory support, diagnose through blood glucose measurement, and initiate fluid therapy, which is the first-line treatment for DKA. The recommendation for urgent hospital transport is based on the need for advanced care and the potential for rapid deterioration in the patient's condition.
4	A 35-year-old male construction worker fell from scaffolding at a height of approximately 2 meters during his work. Upon arrival of the ambulance team, he had regained consciousness but appeared confused and was lying on his back. He complained of severe pain in the lower abdomen and lower back and reported nausea and dizziness. There were no significant external injuries aside from abrasions. His skin was cold and clammy. A witness stated that the fall was direct to the ground without any attempt to break the fall, and that the patient primarily impacted the back side of his body.	The cervical and spinal column must be stabilized, the patient should be placed on a backboard, and unnecessary movement should be avoided. Vital signs should be monitored, oxygen administered, and internal bleeding suspected due to abdominal and lower back pain accompanied by signs of shock. Intravenous fluids should be initiated according to medical guidance, and urgent hospital transport should be arranged with advance notification to the trauma center.	The key findings included a fall from height with direct impact to the back, post-recovery confusion, sharp abdominal and back pain, nausea, dizziness, and cold, clammy skin—signs indicating suspected internal bleeding and possible spinal injury. The recommended actions were chosen to prevent worsening of a potential spinal injury (cervical stabilization and backboard), support respiratory and hemodynamic status (oxygen, monitoring, fluids), and prepare for advanced care at a hospital. The decision to notify the trauma center in advance stems from the risk of rapid deterioration and the need to mobilize immediate resources for continued treatment.
5	During physical activity in the park, a 67-year-old man suddenly collapsed on the path and lost consciousness. Bystanders who witnessed the incident reported that he simply "dropped to the ground" without making a sound. Upon arrival of the ambulance team, the patient was found unconscious, not breathing, and with no detectable pulse. One of the bystanders had already begun performing chest compressions.	Basic life support should be continued immediately (chest compressions at a rate of 100–120 per minute and ventilations if possible), a monitor/defibrillator should be connected, and protocols followed (including delivering a shock if VF/VT is detected). The airway should be secured, oxygen administered, intravenous access established, and resuscitation medications given according to protocol. Prepare for urgent transport to the hospital while continuing advanced resuscitation efforts.	The key findings included unconsciousness, absence of breathing, and absence of a pulse in a patient who suddenly collapsed during physical activity—indications consistent with cardiac arrest. The recommended actions were chosen to ensure immediate life-saving treatment: continuation of basic life support, defibrillator connection for shock delivery if indicated, and support of airway and hemodynamic status through oxygen, IV access, and medications. The recommendation for urgent transport during ongoing resuscitation aims to enable advanced hospital care and reduce the risk of mortality.
6	A 16-year-old boy fell backward while jumping on a trampoline in his backyard. His head struck the ground, and he remained motionless for several seconds. Upon arrival of the emergency team, the patient was	Immediate stabilization of the cervical and spinal column is required using a rigid cervical collar and backboard, and all patient movement should be minimized. Vital signs should be monitored, oxygen	The key findings included a direct fall onto the head, post-trauma confusion, neck pain, and inability to move the legs—all strongly indicating suspected cervical spinal cord injury. The

	conscious but appeared confused, reporting severe headache, blurred vision, and nausea. He also complained of sharp neck pain and was unable to move his legs. There was no external bleeding, but witnesses noted that he landed directly on his head without using his hands to break the fall.	administered, and continuous neurological assessment performed. The patient should be urgently transported to the hospital with prior notification, due to suspected head and cervical spinal cord injury.	recommended actions were chosen to prevent further neurological damage through full immobilization, ensure respiratory and hemodynamic support, and allow for continuous neurological assessment. The decision for immediate hospital transport is based on the need for urgent imaging and potential surgical intervention.

8.2. Appendix A2: Prompts for Suggestions and Explanations

To generate the AI suggestions, the following prompt was entered:

> "Given the following medical emergency scenarios, I want you to describe to me the steps that need to be taken to treat the patient concisely (no more than 3 sentences). Provide an answer as if you are a decision-support AI tool designed to assist EMTs and paramedics during emergencies and suggest treatment plans to treat patients. Provide your suggestion in Hebrew.
>
> The scenario: [description of the emergency scenario]"

To generate the AI explanations, The following prompt was entered:

> "Given the following medical emergency scenario and the suggestion of the treatment plan that you provided, provide me with an explanation of the rationale behind the suggestion, i.e., why did you make that specific suggestion? Make it concise (no more than three sentences) and write the explanation in Hebrew.
>
> Provide a post-hoc explanation and include the following information:
>
> Inputs explanations - inform users what input values from data instances or sensors that the application is reasoning for the current case. When a user asks a why question, she may naively be asking for the Inputs state. We also consider this to be the basic form of explanation to support transparency by showing the current measured input or internal state of the application.
>
> Why explanations - inform users why the application derived its output value from the current (or previous) input values. This is typically represented as a set of triggered rules (rule trace) for rule-based systems or feature attributions (or weights of evidences) for why the inferred value was inferred over alternative values.
>
> The scenario: [description of the emergency scenario]
>
> The AI suggestion: [the AI-generated suggestion]"

8.3. Appendix A3: Treatment Plan Codebook

Category	Sub-category 1	Sub-category 2	Code	Occurrences (Control)	Occurances (Suggestion)
Safety			patient safety	6	0
			EMS Worker Safety	6	0
			Clear area	1	1
			no delaying of evacuation	2	0
Assessments	Primary Survey Assessment	ABC Protocol	Consciousness Assessment	6	0
			Breathing Assessment	10	2
			Pulse Assessment	6	1
			Emergency Case Inquiry	6	1
	Secondary Survey Assessment		Pre-existing Conditions Inquiry	5	0
			Drink/Food Intake Inquiry	1	0
			First occurrence of emergency	3	0

			injury severity	1	0
	neurological assessment	Orientation	Temporal Orientation	4	1
			Spatial Orientation	4	1
			Incontinence assessment	2	1
	physical survey assessment		Cause of unconsciousness	1	2
			Check for additional injury	5	0
			check for external bleeding	1	0
			body scan	1	0
Patient Symptoms			Fainting	1	0
			Vomiting	1	0
			regained consciousness	2	0
			hypoglycemia	1	0
			hyperglycemia	0	1
			internal bleeding	0	1
			DKA	0	1
			Lack of pulse	1	0
			confusion	1	0
			blurry vision	1	0
			convulsions	1	0
			communication ability	1	0
EMS Interventions	Positioning and stabilization		Elevate Legs	1	1
			Lateral positioning	2	0
			spine stabilization	2	2
			lay patient on their back	0	2
			fixation on backboard	3	2
	Resuscitation and Life Support		CPR medicine	0	1
			CPR	1	1
			chest compressions	1	0
			defibrillator attachment	1	1
	Protective measures		protect head	0	1
			release tight clothes	0	1
			Transfer to Shaded Area	1	0
			Cool patient's body	1	0
			keep patient alert	1	0
	Medicinal and Nutritional administration		give juice	1	1
			give bread	1	0
			glucogel administration	1	0
			medicine administration	1	1
			Fluid administration	1	3
	Breathing Support		Oxygen administration	4	6

		Open airway	5	4
	Care Approach	treat according to medical instruction	0	4
		Treatment based on measurements	4	0
		follow protocol	1	1
		supervise	0	2
	Situational Criteria for certain interventions	if patient is	5	0
		if patient can	5	0
		as needed	3	2
Communication	Patient	Communication with Patient	5	0
	Bystanders	Communication with Bystanders	5	0
		Communication with Friends/Relatives	6	0
		request help from bystanders	2	0
	EMS Team	report patient status to call center	1	0
		report patient status to hospital	0	2
		request help from EMS team	1	0
Tests/Measurements	Vital Signs Tests	Pulse test	11	4
		Oxygen Saturation Test	7	0
		Blood pressure test	11	4
	Other standard tests	Blood Glucose Test	8	1
		Body temperature measurement	1	4
		respiratory rate measurement	2	4
Evacuation		transfer to emergency vehicle	1	0
		evacuation	0	5
		helicopter dispatch	1	0
			64	36