Understanding XAI Requirements: A Comparative Study of Repetitive and Unique Decision Contexts

Kay Krachenfels
kkrachenfels@ucsb.edu

Jasmine Lesner
jlesner@ucsb.edu

Misha Sra
sra@cs.ucsb.edu

Abstract

This paper examines how explanation requirements vary between repetitive and unique AI decision contexts through an empirical study of two XAI prototypes. We analyze user interactions with an e-commerce moderation system and a communication monitoring assistant, finding that standardized visual explanations benefit routine tasks while adaptive approaches suit context-specific decisions. Our results suggest design patterns for balancing transparency with usability across different usage scenarios. While our small-scale study (n=8) and prototype-based methodology limit generalizability, particularly regarding real-world implementation challenges and long-term user behavior, our findings provide valuable initial insights into context-dependent explanation design. Further research is needed to validate these patterns at scale and address open questions about optimal confidence communication and security-transparency tradeoffs.

CCS Concepts

• **Human-centered computing** → **User studies**; *Interaction design*; *Empirical studies in HCI*; Interactive systems and tools; • **Security and privacy** → Usability in security and privacy.

Keywords

Explainable AI, XAI, Human-Computer Interaction, HCI, User Interface Design, Content Moderation, Communication Monitoring, Interface Evaluation, AI Transparency, User Trust, Decision Support Systems, Cognitive Load Theory, Human-AI Collaboration, User Experience Design

1 Introduction

As AI systems become integral to daily life, ensuring effective human-AI collaboration requires addressing two key challenges in AI transparency: delivering consistent explanations for repetitive tasks and providing contextual explanations for unique situations. This research addresses three fundamental questions:

(1) How do explanation requirements differ between repetitive and unique decision contexts in AI systems?
(2) What design patterns most effectively support user understanding and trust across these different usage contexts?
(3) How can AI systems balance transparency with cognitive load while maintaining user engagement?

This study investigates these questions through two complementary applications of XAI: a listing moderation system for e-commerce platforms and a communication monitoring assistant for memory augmentation and misinformation detection. These applications represent **two distinct scenarios**: (1) repetitive decision-making requiring consistent explanations, and (2) one-time events needing personalized, contextual explanations.

Our **first XAI prototype**[1] addresses the challenge of listing moderation on e-commerce platforms. Platforms must ensure marketplace integrity by identifying and removing prohibited items while providing actionable feedback to sellers. Success depends on standardized yet detailed explanations that guide compliance. Effective systems must flag problematic listings accurately and communicate their reasoning clearly to foster seller understanding and compliance.

Our **second XAI prototype**[2] combines memory augmentation with misinformation detection in an AI assistant that monitors communications — such as conversations, emails, and documents — for factual inconsistencies or errors. While memory aids and fact-checking tools are well-studied separately, their integration introduces unique opportunities. Memory systems enhance recall but may propagate misunderstandings, while fact-checking tools verify information but often lack personal context. Our integrated approach addresses these limitations, helping users detect misunderstandings, errors, or dishonesty in communication. Cognitive Load Theory underscores the challenges of recalling details in information-rich environments, where human working memory struggles [7]. The assistant must offer timely, relevant interventions while minimizing cognitive overload.

This work examines how recurring and one-time decision patterns shape explanation strategies. Although both scenarios highlight explainability's role in fostering user trust and effective collaboration, they differ in requirements. E-commerce moderation calls for standardized, scalable explanations to ensure consistency across cases. In contrast, communication monitoring requires highly contextual explanations that combine personal history with real-time fact-checking. By exploring these contrasting cases, we aim to advance XAI frameworks that adapt to varied usage patterns while addressing shared challenges: balancing transparency with usability and maintaining user engagement through clear, actionable explanations.

2 Related Work

Research integrating memory augmentation with misinformation detection is scarce. However, prior work in explainable AI, content moderation, memory assistance, and misinformation detection provides foundational insights. We build on these domains to design interfaces that make AI decision-making transparent and actionable across various contexts.

Explainable AI: Effective explanations must balance clarity and utility [17]. While model-agnostic methods offer broad applicability [19], the optimal approach depends on context and decision frequency. Recent advances, such as Chen et al.'s XplainLLM dataset

[1] E-commerce Moderator [12] https://xai.ackop.com/moderator.html
[2] Communication Monitor [9] https://xai.ackop.com/monitor.html

and framework [6], leverage knowledge graphs and reasoning elements to generate grounded explanations. Inspired by such developments, we tailor explanation strategies to specific usage patterns in our dual-prototype approach.

Engaging users with AI explanations remains a challenge. Dual process theory highlights the tension between fast, intuitive 'System 1' thinking and deliberate, analytical 'System 2' thinking [11]. Structured interventions can activate System 2 processes, improving understanding [13]. Our UI designs combine intuitive workflows with prompts for deeper analysis, using visual cues to encourage reflection on AI outputs.

Content Moderation: While explainability research in e-commerce is limited, studies in social media moderation offer some insights:

- 42% of Reddit users were unaware of post removals until surveyed, but explanations improved perceived fairness [10].
- Moderation messages with AI explanations enhanced fairness perceptions, though human explanations were preferred [5].
- Moderator tools should emphasize rapid reclassification of false AI flags [4].

Content moderation highlights the value of structured explanations for systematic feedback [1] and the need for dynamic, trust-preserving explanations in real-time decisions [14]. Frameworks categorizing explainability by user needs [2, 8] inform our e-commerce moderation prototype, which uses standardized visual explanations for repetitive tasks. Meanwhile, our communication monitor incorporates dynamic explanations for contextual decisions.

Memory Augmentation and Misinformation Detection: Advances in large language models support conversational memory recall, aiding decision-making for users under cognitive load, including older adults [16, 24]. Building on these capabilities, our communication monitoring prototype addresses a critical gap: verifying information accuracy across conversations and documents.

While automated misinformation detection systems show promise, research has highlighted risks of human over-reliance on these tools, suggesting the need for transparent and educational approaches that empower rather than replace human judgment [18].

By synthesizing memory augmentation and misinformation detection, we design an interface that integrates recall assistance with accuracy validation. This integration allows our communication monitoring prototype to move beyond isolated memory support or fact-checking. It offers a unified solution to maintain information integrity across diverse digital interactions, addressing the growing complexity of modern communication.

3 E-Commerce Moderation Prototype

For regulatory compliance, fraud prevention and liability reasons online marketplaces require their listings follow (often a long list of) specific rules. According to these rules a listing may be flagged for one or more different reasons. This can leave sellers frustrated struggling to understand what happened and what actions they can take. Our UI prototype demonstrates explanations for three such reasons:

(1) The item is prohibited from being sold on the site
(2) Mismatch between listing text and listing images

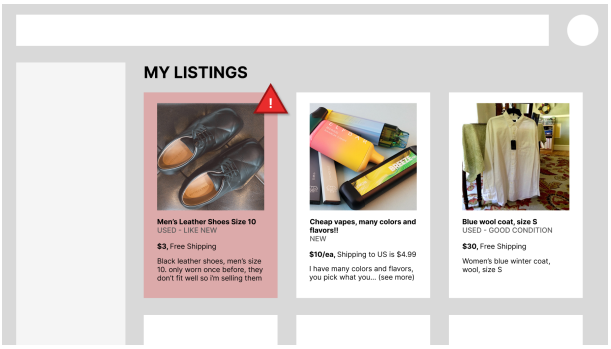

Figure 1: E-Commerce Seller UI shows their listings as a grid of cards which sellers are able to scroll up and down. Listings that have been flagged are colored red and have a red triangle which is an alert action button.

(3) The item price deviates too much from the norm

Each flagged listing follows a similar review process, in which a user views explanations for the listing being flagged and then is offered actions to take. Listings that are flagged are colored red, and a red triangle alert button is placed at the corner of the image, easily noticeable, but non-intrusive. For each flagged listing the UI flow (Figure 2 and Figure 3) follows three steps:

Step 1: Notify The user sees the alert icon/button in the corner of the listing, and the listing is colored red.

Step 2: Explain The interface informs the user of why the item was flagged. Contextual visual elements are used:

- *Price mismatch/deviation:* The popup displays a graph showing the average prices of similar items. The mismatch in price will be highlighted on the listing.
- *Image-text mismatch:* The mismatched portion of the image is highlighted (via segmentation) and the mismatched portion of the text is highlighted.
- *Prohibited item:* The mismatched portion(s) of the image and/or text are highlighted and/or segmented.

Step 3: Adjust or Appeal The user now has the opportunity to take action. In each case, the user has the option to click a button labeled "unflag and post," which allows them to appeal the flagging decision. The context specific courses of action they can take are:

- *Price mismatch/deviation:* adjust the list price of their item.
- *Image-text mismatch:* change the listing text or image to remove the mismatch.
- *Prohibited item:* view the list of prohibited items. The prohibited item detected will be highlighted on the list.

3.1 Promotion of Analytical Thinking

The e-commerce moderation design encourages analytical thinking through the following elements:

(1) **Progressive Information Disclosure**: The three-step flow (notify → explain → act) presents details incrementally, prompting users to process information step-by-step.
(2) **Visual Evidence Presentation**:

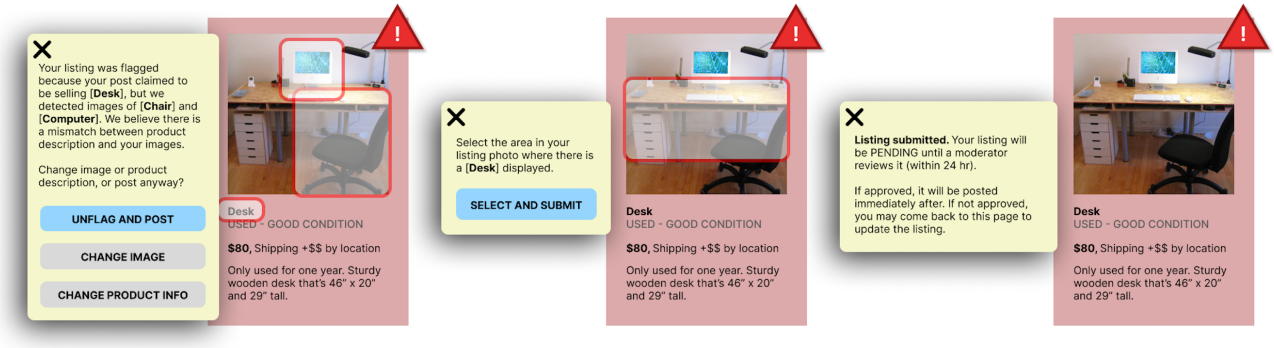

Figure 2: UI flow for notify → explain → adjust or appeal flow, where the model has incorrectly flagged a listing. In this case, the user is *notified* by the alert icon, and clicks on it to see that their listing has been (incorrectly) flagged as an image-text mismatch. The pop-ups *explain* what the mismatch is, highlighting/segmenting parts of the image and text that the AI thinks do not align. The user *appeals* by clicking "unflag and post", following prompts to submit the petition.

Figure 3: Two other e-commerce moderation examples. In both cases the model has correctly identified listings to be flagged. The user has two options for each; they can petition to unflag, or adjust price (shoes) or view all prohibited items (e-cigarettes).

- Price graphs show market distributions, guiding sellers to analyze pricing strategies.
- Image segmentation highlights mismatched elements, encouraging careful comparison.
- Prohibited item indicators, paired with a comprehensive list, clarify policy context.

(3) **Interactive Decision Points**: Users actively engage through:
- Explicit choices between adjustment and appeal options.
- Evidence review before decision-making.
- A structured appeal process requiring justification.

(4) **Contextual Learning**: When viewing prohibited items, the system highlights specific violations alongside the full policy, helping users understand rules through concrete examples.

Together, these elements transform moderation into an interactive learning experience, fostering understanding of marketplace policies and listing best practices.

4 Communication Monitoring Prototype

AI-powered communication monitoring identifies potential issues like ambiguities, discrepancies, inconsistencies, and mismatched facts across spoken and written communication. While some memory augmentation tools require specialized devices [16, 24], our UI design (Figures 4 and 5) is tailored for smartphones.

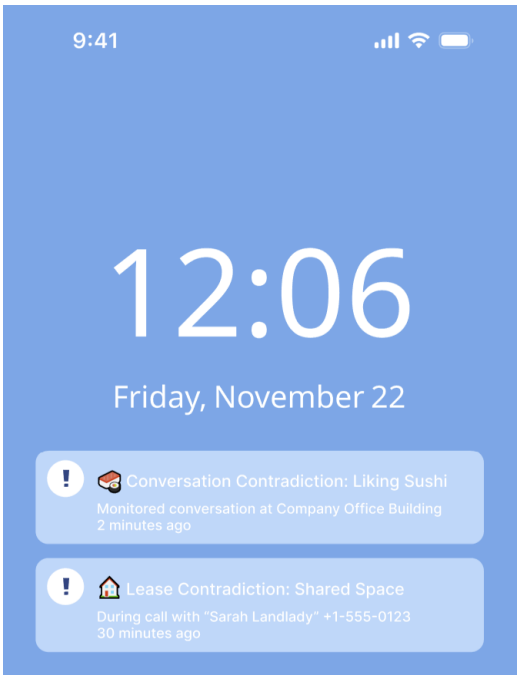

Figure 4: Users receive notifications on their smartphones, leveraging familiar mobile alerts to communicate detected inconsistencies non-intrusively.

4.1 Nav-bar Sections

The footer nav-bar has five main sections:

(1) **Documents** are used by the assistant during analysis. The UI supports uploading documents, connecting storage accounts (e.g., Google Drive), searching, tracking recent changes, and ensuring important data is available for contextual alerts.

(2) **Notifications** are color-coded by severity:
- *Red* for high-risk alerts (e.g., medical or financial issues).
- *Orange* for moderate risks (e.g., schedule conflicts).
- *Yellow* for low-risk issues.

(3) **Recording** gives users control over what is recorded and analyzed. This feature accommodates those who prefer manual monitoring or want to temporarily disable it.

(4) **Settings** allows users to adjust detection thresholds, manage app permissions, and configure features like internet-based fact-checking. These controls emphasize user privacy and autonomy.

(5) **Information** explains how the assistant works, highlights secure data storage practices, and outlines alert categories. Clear communication in this section builds trust.

4.2 Notification Features and Tabs

Notifications combine text with contextually relevant emojis (e.g., a house emoji for lease-related alerts). Users can sort notifications by recency, importance, or confidence level.

Detailed explanations are accessible through the Notifications view, where users can review issues via notification cards. Tapping 'Learn More' provides in-depth insights and guidance for resolving inconsistencies, as shown in Figure 7. This process moves through three stages, implemented as tabs:

(1) **Overview Tab**. This tab summarizes the detected issue, confidence level, and supporting evidence. Problematic statements are highlighted in red, and contrasting evidence from verified documents is shown in green (e.g., highlighting a lease clause permitting backyard access that contradicts the landlady's claim). This transparent design helps users quickly understand the issue.

(2) **Discuss Tab**. This tab uses an interactive chatbot interface to explain the detection process, explore potential risks, and clarify confidence levels. Users can ask questions or use suggested prompts. The chatbot provides detailed responses tailored to each scenario, as illustrated in Figure 8.

(3) **Resolve Tab**. This tab focuses on actionable solutions. It suggests AI-generated resolutions (e.g., contacting relevant parties via default apps or addressing discrepancies in healthcare or financial records), allows users to report false positives to improve system accuracy, and enables custom resolution paths. These options personalize the user experience while improving future recommendations.

4.3 Promotion of Analytical Thinking

The communication monitor design employs several elements to encourage analytical thinking and deeper user engagement:

(1) **Progressive Information Architecture**: The three-tab system (Overview → Discuss → Resolve) guides users through increasingly detailed analysis levels, promoting systematic

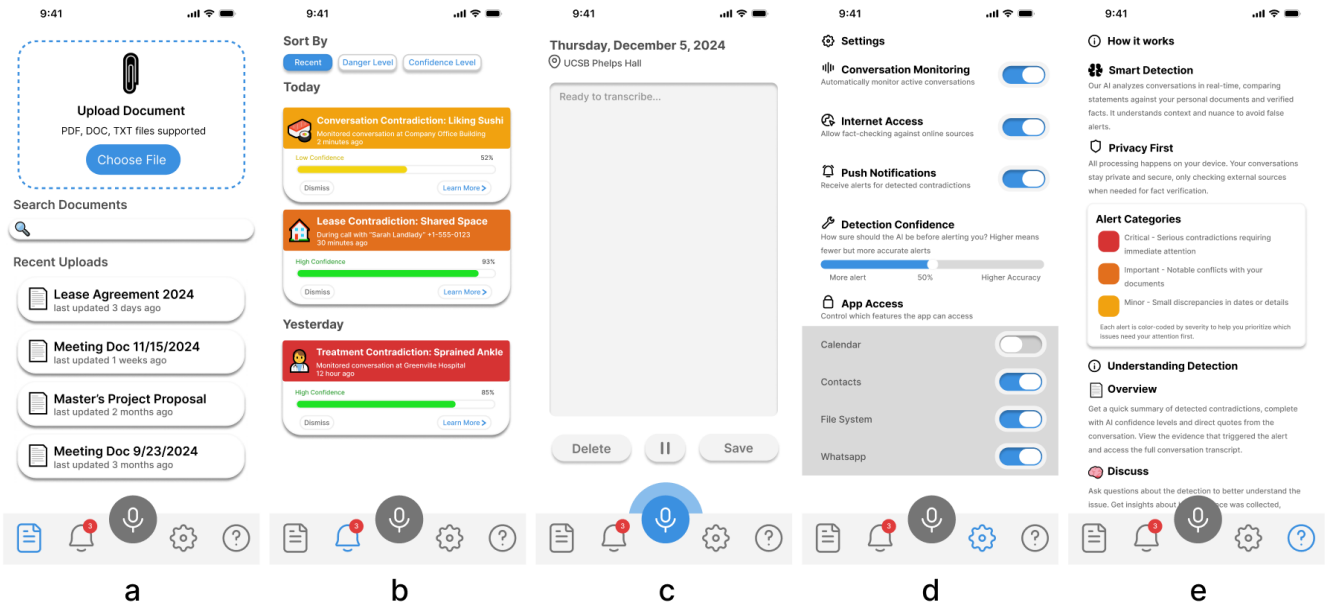

Figure 5: Communication monitoring nav-bar includes: (a) *Documents* for managing personal files, (b) *Notifications* for reviewing detected issues, (c) *Recording* for manual audio monitoring control, (d) *Settings* for customization, and (e) *Information* for user education.

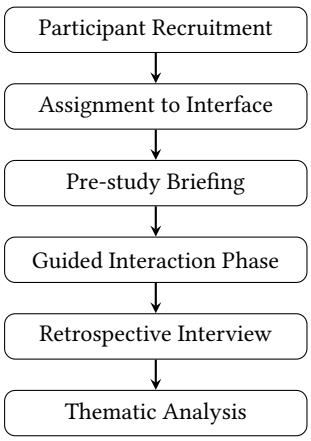

Figure 6: User Study Methodology

evaluation of detected inconsistencies. The multimodal approach enhances comprehension through Dual Coding Theory [20], using emojis selected by LLMs for context-appropriate symbolism [23].

(2) **Interactive Evidence Review**: Users actively engage with supporting documentation through color-coded highlights and linked references, encouraging critical comparison of contradictory information. Confidence scores use color indicators (green for high, yellow for low), progress bars, and concise descriptions, aligning with best practices to prevent AI overreliance [15, 22].

(3) **Guided Inquiry Interface**: The Discuss tab's chatbot uses structured prompts and follow-up questions to promote deeper analysis of detected issues, their implications, and confidence assessments. This draws on research about AI conversational interfaces promoting reflective thinking [21].

(4) **Action-Oriented Resolution**: The Resolve tab requires users to evaluate and select appropriate responses, transforming passive consumption into active decision-making. These structured interactions help manage cognitive load while maintaining engagement.

5 User Study Methodology

We conducted a qualitative study (Figure 6) with eight participants (university students, ages 18–22). This sample size was chosen as appropriate for our preliminary investigation, allowing for detailed qualitative analysis while gathering initial insights to inform future prototype iterations. The limited scale enabled in-depth interviews and thorough analysis of user interactions, with plans to expand to a larger participant pool in subsequent studies.

Participants were evenly split: four computer science (CS) majors with AI experience and four non-CS majors with limited AI exposure. This balance allowed us to assess how varying AI literacy affects understanding and trust in explainable interfaces.

Using a between-subjects design, participants evaluated one interface: two CS and two non-CS users tested the e-commerce

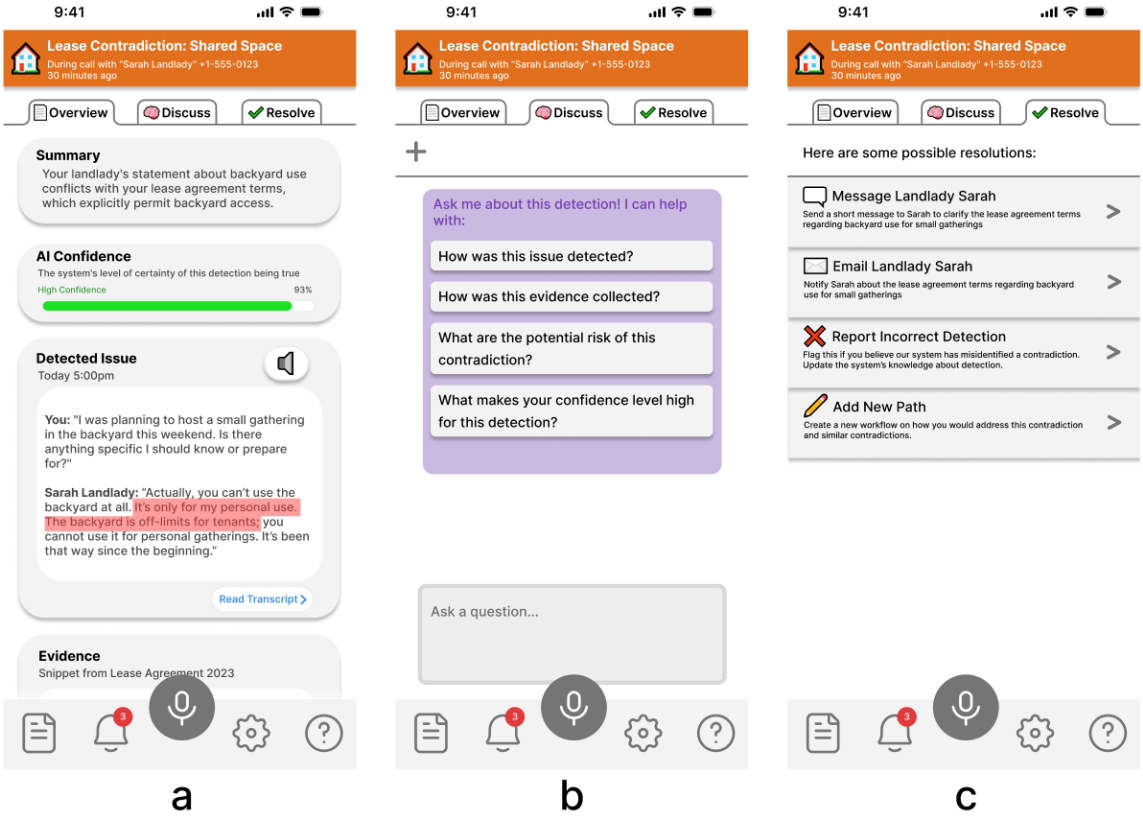

Figure 7: A lease agreement contradiction detected with 93% confidence, showing tabs for (a) Overview (detection details), (b) Discuss (chatbot interaction), and (c) Resolve (proposed solutions).

moderation prototype, while the same split applied to the communication monitoring prototype.

E-commerce Moderation Tasks:

(1) Review a listing flagged for image-text mismatch: Figure 2 showing desk photo with incorrect description.
(2) Evaluate a price mismatch case: top of Figure 3 showing shoes priced significantly above market.
(3) Assess a prohibited item flag: bottom of Figure 3 showing e-cigarette listing.

Communication Monitor Tasks:

(1) Analyze a lease agreement contradiction: Figures 7 and 8 showing backyard access dispute.
(2) Review a medical information discrepancy: Figures 10 and 11 showing conflicting treatment advice.
(3) Evaluate a social inconsistency: Figures 12 and 13 showing conflicting food preferences.

A retrospective interview collected feedback on participants' experiences, understanding of AI explanations, and improvement suggestions. Participants also rated explanation effectiveness on Likert scales, focusing on understanding the AI's decision-making and trust in its explanations (Appendices A, B).

We focused our analysis on how the explanations affected comprehension and trust metrics, with particular attention to differences between CS and non-CS participants' responses.

6 User Study Results

Table 1 shows how users rated our prototypes on "Understanding" and "Trust". For the e-commerce prototype, while understanding was uniformly high (5/5) across both groups, CS participants showed more skepticism in their trust ratings (3-4/5) compared to non-CS participants (4-5/5) suggesting that technical expertise can lead to more critical evaluation of AI systems. In the communication monitoring prototype, CS participants demonstrated higher understanding (4-5/5) but lower trust (2-3/5) compared to non-CS participants' more varied understanding (3-5/5) and trust (3-5/5) ratings. CS participants cited specific technical concerns about privacy and verification mechanisms, while non-CS participants focused more on practical utility.

6.1 E-Commerce Moderation

The key themes from user feedback are shown in Table 2.

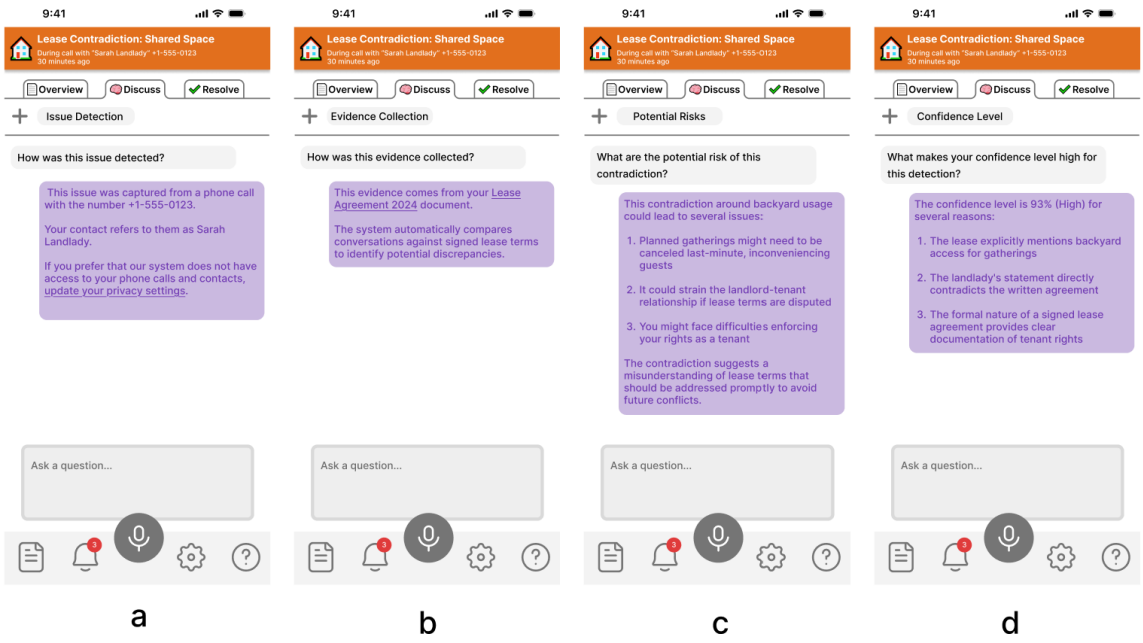

Figure 8: Key stages of lease inconsistency analysis: issue identification from phone call, examination of lease agreement evidence, assessment of potential tenant risks, and explanation of high confidence (93%) based on clear lease terms.

Table 1: System Understanding and Trust Ratings (Likert 1-5 Scale) Across CS and Non-CS Major Participants

Prototype	Participant (CS/non-CS)	System Understanding	Trust in System
E-Commerce Moderation Prototype	P1 (non-CS)	5	4
	P2 (non-CS)	5	5
	P3 (CS)	5	3
	P4 (CS)	5	4
Communication Monitoring Prototype	P5 (non-CS)	3	5
	P6 (non-CS)	5	3
	P7 (CS)	4	3
	P8 (CS)	5	2

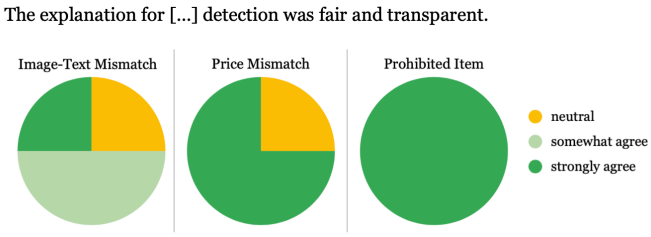

Figure 9: Perception of fairness and transparency of e-commerce moderation explanations.

Understanding and Trust: As shown in Table 1 participants have relatively high trust in the e-commerce moderation explanations, responding with at least a 3/5 or above. Additionally, participants indicated a strong understanding and rated this a 5/5.

Fairness and Transparency: Results are displayed in Figure 9. In general, participants felt at least neutral about the statement that each type of scenario explanation was neutrally fair and transparent (we had no responses for somewhat disagree or strongly agree here).

In the image-text mismatch scenarios, participants generally agreed with explanations being fair and transparent. In the price mismatch scenario, both P1 and P2 strongly agreed with this statement. P4 was neutral, but P3 participant strongly agreed and commented they *liked the graph and distribution of the price in the shoes listing.* All participants strongly agreed that the prohibited item explanation was fair and transparent, with P4 saying that they *"liked the warning about prohibited items"* and P3 saying that they *"liked the e-cigarette pop-up that highlighted which item it was on the list that was prohibited."*

Frustration: For each scenario, participants were asked to rate agreement/disagreement with the phrase "The pop-ups for [X] detection were frustrating." where [X] was one of "image detection mismatches", "price detection mismatches" or "prohibited item detection". During surveys, we realized that this question may not have been worded clearly as participants interpreted this differently. However, we have a few interesting remarks. P4, who indicated neutral agreement towards price mismatch fairness and transparency, also gave strong agreement towards this scenario's pop-ups being frustrating, commenting: *"price mismatch warning didn't feel necessary."* For image-text mismatch, P1 said they "somewhat agreed" with the pop-ups being frustrating, and remarked that they would have liked to *"see alternatives for what the computer thought the detected image was or see other detected items in the image."*

Table 2: Key Themes from E-Commerce Moderation Feedback

Theme	Representative Participant Comments
Visual Explanations	"Liked the graph showing price distribution" (P3) "Highlighting prohibited items on the list was helpful" (P3)
Actionable Feedback	"Clear options to adjust or appeal" (P2) "Easy to understand what needed to be fixed" (P1)
Areas for Improvement	"Would like to see what items AI detected in images" (P1) "Price mismatch warnings felt unnecessary" (P4)

6.2 Communication Monitoring

The key themes from user feedback are shown in Table 3.

Trust Dynamics and Verification: Participants' trust in the system varied depending on the type of inconsistency it flagged. Factual inconsistencies were trusted more than subjective interpretations. As P5 stated, *"If it is based on anything factual, I would trust it very highly. If it comes to interpreting what someone means...I would give it a three."* Users also emphasized the importance of transparent confidence metrics, preferring simplified indicators over precise percentages. P6 asked, *"What makes it 92% not 100%?"* while P7 expressed the need for verifiable confidence, saying, *"I'd probably want some stats to verify...the goodness of the measure."*

Interestingly, all participants did not realize that the highlights on evidence were not reflective of the AI's attention mechanisms but were designed to help users quickly spot inconsistencies. Despite this, the highlights were universally appreciated for their utility in guiding the user's attention. As P7 noted, *"The red highlights showed me exactly what to focus on...it helped me see the contradiction clearly."* This misunderstanding did not appear to undermine user trust in the system.

Privacy and Control Preferences: Participants strongly preferred opt-in controls and manual recording options, valuing agency over when and how the system monitors conversations. P8 noted, *"I probably wouldn't let it quit conversation monitor...I'd probably do the recording myself."* There was also concern about data handling, with users wanting greater transparency regarding who can access their data and how it is used. P6 summarized this sentiment, saying, *"I don't think I would just carry this around and be like, sure, listen to all of my conversations."*

Interface and User Experience: The system's use of color coding for risk levels was well-received, as it simplified comprehension, though text-heavy explanations were criticized for being overwhelming. P7 remarked, *"The green to the red to green is nice...there's a lot of text here and that really simplifies it."* Initial difficulties understanding the help sections improved after hands-on interaction, with P6 observing, *"While I didn't understand the explanations in the info tab, once I was exposed to the notifications...I understood what it meant."*

Use Case Applications: Users identified value in professional and academic settings, particularly for verifying accuracy in presentations and documents. P6 suggested, *"If I was doing a presentation and I uploaded all of my references...I'd want to know if anything I'm saying is contradicted in a reference."* However, caution was expressed about the system challenging domain expertise. As P7 explained, *"This approaches like this territory where you're challenging someone who has a few years of experience in this field."*

7 Discussion

We analyzed how explanation designs in AI systems enhance user understanding, trust, and satisfaction, informing future XAI development. Our findings align with and extend prior work on explainable AI interfaces [1, 17], while addressing our core research questions about explanation requirements across contexts, effective design patterns, and balancing transparency with cognitive load.

For e-commerce moderation, our first research question revealed that repetitive tasks benefit from structured explanations that effectively clarify flagged items and outline corrective actions. Visual segmentation, particularly for prohibited items, improved comprehension, supporting research on the value of visual explanations in AI systems [19]. Graphs highlighting price deviations demonstrated the effectiveness of multimodal elements for nuanced decisions, consistent with findings on dual coding theory [20]. However, participants expressed frustration with explanations lacking sufficient context for price mismatches, echoing challenges identified in social media moderation research [5, 10]. This highlights the need for user-friendly, data-grounded rationales.

For communication monitoring, addressing our second research question revealed that design patterns must prioritize simplicity and clarity in unique decision contexts. Participants preferred intuitive confidence metrics over precise percentages, aligning with research on effective confidence visualization [15, 22]. Highlighted inconsistencies effectively guided users, though some misinterpreted them as representing AI attention mechanisms. This finding suggests opportunities to leverage 'System 2' analytical thinking through structured visual cues [11, 13].

Regarding our third research question on cognitive load, our results demonstrate that effective explanations must balance detail with accessibility. The highlights significantly aided comprehension despite misinterpretation, emphasizing the importance of designs that communicate complex processes without overwhelming users. This validates our approach of progressive disclosure and visual anchoring as methods to manage cognitive load while maintaining engagement.

7.1 Implications for Practitioners

Our findings suggest several practical recommendations for XAI interface design:

Table 3: Key Themes from Communication Monitoring Feedback

Theme	Representative Participant Comments
Trust Factors	"Trust high for factual checks, lower for interpretations" (P5) "Need more explanation of confidence scores" (P6)
Privacy Concerns	"Want control over what gets recorded" (P8) "Unclear who has access to conversation data" (P6)
Interface Design	"Color coding simplified understanding" (P7) "Text explanations could be overwhelming" (P7)
Use Cases	"Useful for verifying presentation accuracy" (P6) "Concerns about challenging expert knowledge" (P7)

1. **Progressive Disclosure**: Layer explanations to prevent cognitive overload [1]. Start with high-level summaries and allow users to drill deeper as needed.

2. **Visual Anchoring**: Use consistent visual elements (highlighting, color-coding) to draw attention to key information. This aligns with research showing improved comprehension through multimodal presentation [20].

3. **Confidence Communication**: Present certainty levels through simple visual indicators rather than just precise percentages, following best practices in uncertainty visualization [15].

4. **Interactive Controls**: Provide mechanisms for users to adjust explanation depth and verify AI decisions, supporting findings on the importance of user agency in XAI systems [14].

These guidelines particularly benefit developers implementing explainable AI in production systems, where balancing transparency with usability is crucial for adoption.

7.2 Key Lessons

Our user study reinforces three key lessons for explainable AI design:

(1) **Context-Aware Explanations**: Effective explanations align with task frequency. Structured feedback works well for repetitive tasks, while personalized explanations suit unique, context-specific decisions.
(2) **User Agency**: Participants valued mechanisms to act on explanations, such as "adjust" or "appeal" options in the e-commerce system and interactive resolution tabs in the communication monitor. Allowing users to control their engagement with explanations fosters trust and autonomy.
(3) **Transparency and Usability**: Transparent systems build trust but must avoid overwhelming users with excessive detail. Combining confidence scoring with visual aids, like color-coded indicators, effectively communicates certainty. Adapting explanation depth based on user engagement balances clarity with simplicity.

7.3 Ethical Considerations

Our prototypes aim to adhere to four key principles:

Privacy: In e-commerce, it's crucial to anonymize sensitive data to protect sellers' competitive information. For communication monitoring, privacy is safeguarded through strong speaker verification, clear consent mechanisms, and transparency about recorded data. Systems must also tackle risks like altered memories from selective reinforcement or falsified recordings. Tools such as audio replays can aid this effort, but generative AI adds challenges by enabling convincing fake recordings.

Bias and Fairness: E-commerce moderation must avoid reinforcing marketplace biases by auditing explanation patterns across seller demographics. Communication monitoring should account for cultural differences to ensure fair flagging and scoring. Systems must also address risks of altered recollections from selective reinforcement or faked recordings. Features like audio replays can help, but generative AI complicates this by enabling realistic but falsified recordings.

Limitations of Explanations: Systems must communicate the limits of their explanations for example by indicating AI model confidence. In e-commerce moderation, users should be cautioned against over-relying on explanations, which cannot fully represent the underlying moderation processes.

Balanced Adaptability: Systems should adapt to user behavior, such as frequent dismissal of explanations, by adjusting explanation depth. Effective explanations must also balance transparency with system security and user privacy. Continuous evaluation of explanations' impact on user understanding and decisions is essential.

7.4 Limitations and Future Work

This study faces four key limitations that present opportunities for future research:

Prototype Fidelity: Our low-fidelity prototypes simulating AI responses prevented evaluation under real conditions with varying accuracy, processing speeds, and edge cases. Future work should implement production-grade AI models to validate findings and identify real-world challenges.

Implementation Challenges: Real-world deployment introduces integration hurdles with existing platforms, scalability requirements, and privacy compliance needs across jurisdictions. Research is needed on balancing explanation quality with performance and regulatory constraints.

Confidence Communication: Current percentage-based confidence scores lack contextual meaning. Future work should explore personalized calibration techniques for contextualizing scores.

Security-Transparency Tradeoffs: Over-disclosure of detection methods risks exploitation by malicious actors seeking to evade detection. Research must determine optimal transparency levels that build trust while maintaining system integrity. This extends work on adversarial robustness in explainable systems [3] by examining explanation-specific vulnerabilities.

8 Conclusion

Our study of explainable AI interfaces across e-commerce moderation and communication monitoring revealed distinct patterns in how explanation design impacts user trust and understanding. The findings advance our understanding of context-dependent XAI design in three key areas:

Task-Specific Explanation Patterns. For repetitive e-commerce tasks, structured visual explanations with clear action paths proved effective, with users reporting high understanding (5/5) across both technical and non-technical backgrounds. In contrast, communication monitoring required more nuanced, contextual explanations, leading to varied understanding (3-5/5) but highlighting the importance of progressive disclosure for complex decisions.

Trust Dynamics. Our results revealed that trust formation differs by context and user expertise. E-commerce users showed high trust in standardized explanations for routine decisions (3-5/5), while communication monitoring users exhibited more variance (2-5/5), particularly around privacy and verification. This aligns with our discussion findings on the importance of transparent confidence metrics and user control in sensitive contexts.

Design Implications. The synthesis of our findings suggests a framework for context-aware XAI:

- Repetitive tasks benefit from standardized, action-oriented explanations with consistent visual elements
- Unique decisions require adaptive explanations with progressive disclosure and strong privacy controls
- Both contexts need clear confidence communication and user agency in verification

While our small-scale study provides valuable initial insights, several critical questions warrant future investigation at scale:

- How do these explanation patterns perform under real-world conditions with varying accuracy and edge cases?
- What are optimal approaches for contextualizing confidence scores across different usage scenarios?
- How can explanation designs balance transparency with security against adversarial exploitation?

By identifying these context-dependent patterns in XAI design, our work provides a foundation for developing more effective, human-centered explainable AI systems. Future research can build on these findings to create interfaces that maintain user trust and understanding while addressing the unique challenges of different application domains.

Author Contributions

K. K. developed the e-commerce moderation prototype. J. L. developed the communication monitoring prototype. M. S. guided and supervised the work.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[3] Hubert Baniecki and Przemyslaw Biecek. 2024. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion* 107 (2024), 102303. https://doi.org/10.1016/j.inffus.2024.102303 arXiv:2306.06123 [cs.CR]

[4] Enrico Bunde. 2021. AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators – A Design Science Approach. *Proceedings of the 54th Hawaii International Conference on System Sciences* 3 (2021), 1264–1273. https://hdl.handle.net/10125/70766

[5] Erik Calleberg. 2021. Making Content Moderation Less Frustrating. https://www.diva-portal.org/smash/get/diva2:1576614/FULLTEXT01.pdf

[6] Zichen Chen, Jianda Chen, Mitali Gaidhani, Ambuj Singh, and Misha Sra. 2023. XplainLLM: A QA Explanation Dataset for Understanding LLM Decision-Making. *arXiv preprint arXiv:2311.08614* (2023). https://arxiv.org/abs/2311.08614

[7] Nelson Cowan. 2010. The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science* 19, 1 (2010), 51–57.

[8] Google. 2024. Guidebook for Pair Programming. https://pair.withgoogle.com/guidebook/. Accessed: 2024-11-30.

[9] Jasmine Lesner. 2024. Communication Monitoring Prototype. Online. https://xai.ackop.com/monitor.html

[10] Shagun Javier, Darren Scott Apling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 192 (2019). https://doi.org/10.1145/3359294

[11] Daniel Kahneman. 2002. Maps of bounded rationality: A perspective on intuitive judgement and choice. (2002).

[12] Kay Krachenfels. 2024. E-commerce Moderator Prototype. Online. https://xai.ackop.com/moderator.html

[13] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety* 25, 10 (2016), 808–820.

[14] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1257–1268.

[15] Mélanie Lubrano, Yaëlle Bellahsen-Harrar, Rutger Fick, Cécile Badoual, and Thomas Walter. 2023. Simple and Efficient Confidence Score for Grading Whole Slide Images. arXiv:2303.04604 [eess.IV] https://arxiv.org/abs/2303.04604

[16] Natasha Maniar, Samantha Chan, and Pattie Maes. 2024. MemPal: Wearable Memory Assistant for Aging Population. https://www.media.mit.edu/projects/mempal/overview/. Accessed: 2024-11-29.

[17] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[18] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st annual ACM symposium on user interface software and technology*. 189–199.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[20] Mark Sadoski and Allan Paivio. 2004. A dual coding theoretical model of reading. *Theoretical models and processes of reading* 5 (2004), 1329–1362.

[21] Marco Antonio Rodrigues Vasconcelos and Renato P. dos Santos. 2023. Enhancing STEM learning with ChatGPT and Bing Chat as objects to think with: A case study. *Eurasia Journal of Mathematics, Science and Technology Education* 19, 7 (July 2023), em2296. https://doi.org/10.29333/ejmste/13313

[22] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. ACM. https://doi.org/10.1145/3351095.3372852

[23] Yuhang Zhou, Paiheng Xu, Xiyao Wang, Xuan Lu, Ge Gao, and Wei Ai. 2024. Emojis Decoded: Leveraging ChatGPT for Enhanced Understanding in Social Media Communications. arXiv:2402.01681 [cs.CL] https://arxiv.org/abs/2402.01681

[24] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 450, 18 pages. https://doi.org/10.1145/

3613904.3642450

A E-Commerce Moderation User Survey

Each question or statement was rated on a Likert scale from 1-5, where 1 indicated strong disagreement, 3 was neutral, and 5 was strong agreement. Participants were asked two general questions after reviewing all scenarios and the full interface, as well as two questions for each type of listing flag scenario.

The two general questions were:

(1) "I would trust this system."
(2) "I understand this system."

For each of the three scenarios, participants were asked two additional questions about frustration and about fairness and transparency:

(1) "The pop-ups for [X] were frustrating."
(2) "The explanations for [X] were fair and transparent."

where [X] was one of "image detection mismatches", "price detection mismatches" or "prohibited item detection".

Finally, participants provided open-ended feedback on what aspects of the interface they liked, disliked, or would change.

B Communication Monitoring User Survey

As before, each question or statement was rated on a Likert scale. Participants were asked five general questions after reviewing all scenarios and the full interface, along with an initial comprehension question for each inconsistency scenario.

For each inconsistency scenario, participants were first asked:

(1) "In your own words, explain what inconsistency the system detected."

After reviewing all scenarios, participants rated their agreement with the following statements:

(1) "The system's inconsistency detection is reliable."
(2) "I trust this system's ability to detect inconsistencies."
(3) "I would be comfortable using this system in my daily life."
(4) "I understand the system's explanations."

Finally, participants provided open-ended feedback on what aspects of the explanations influenced their trust or distrust in the system.

C Additional Scenarios

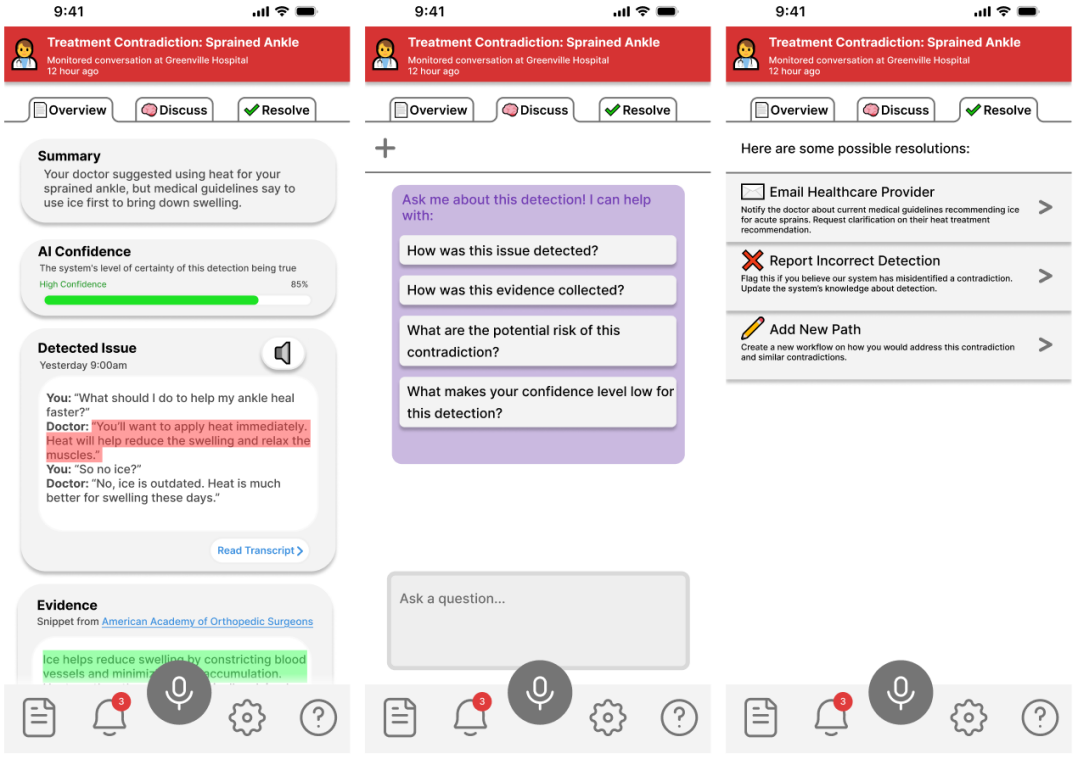

Figure 10: A medical alert for conflicting sprained ankle treatments, where a doctor recommended heat therapy contrary to guidelines advocating ice therapy.

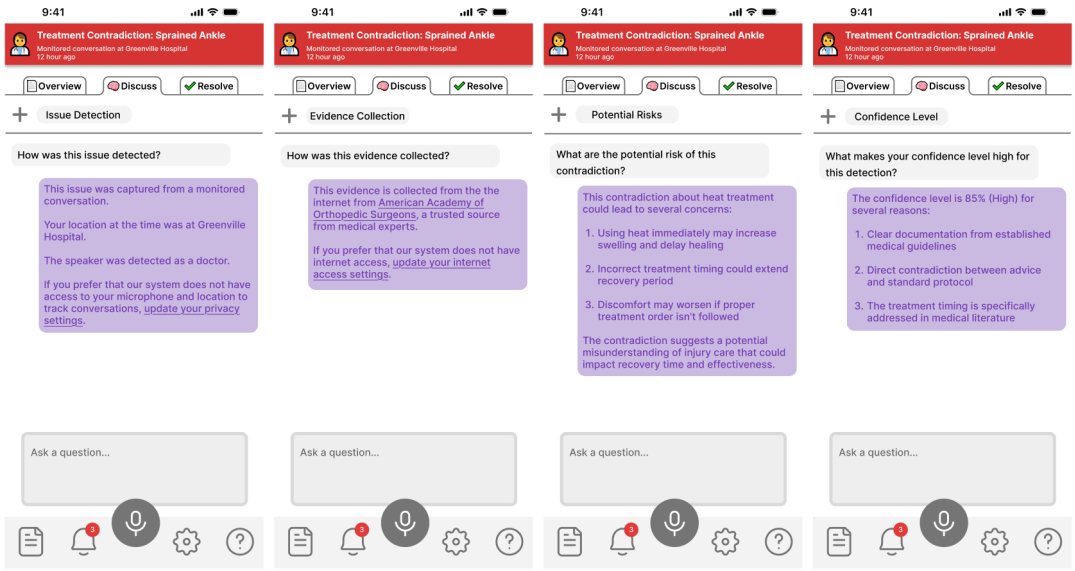

Figure 11: Stages of a medical alert explanation: issue detection from a monitored conversation, evidence collection from orthopedic guidelines, analysis of potential risks from incorrect treatment, and explanation of the system's confidence assessment.

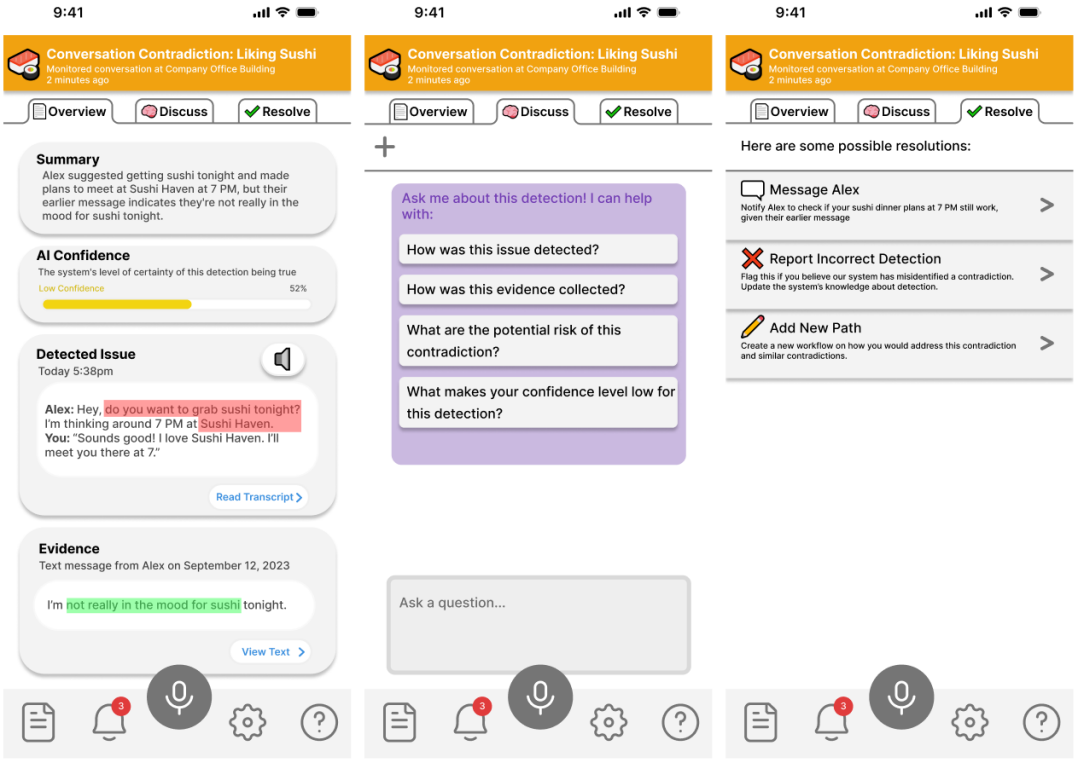

Figure 12: A conversation contradiction alert, where Alex made sushi dinner plans despite earlier expressing disinterest in sushi, with 52% AI confidence in the detected inconsistency.

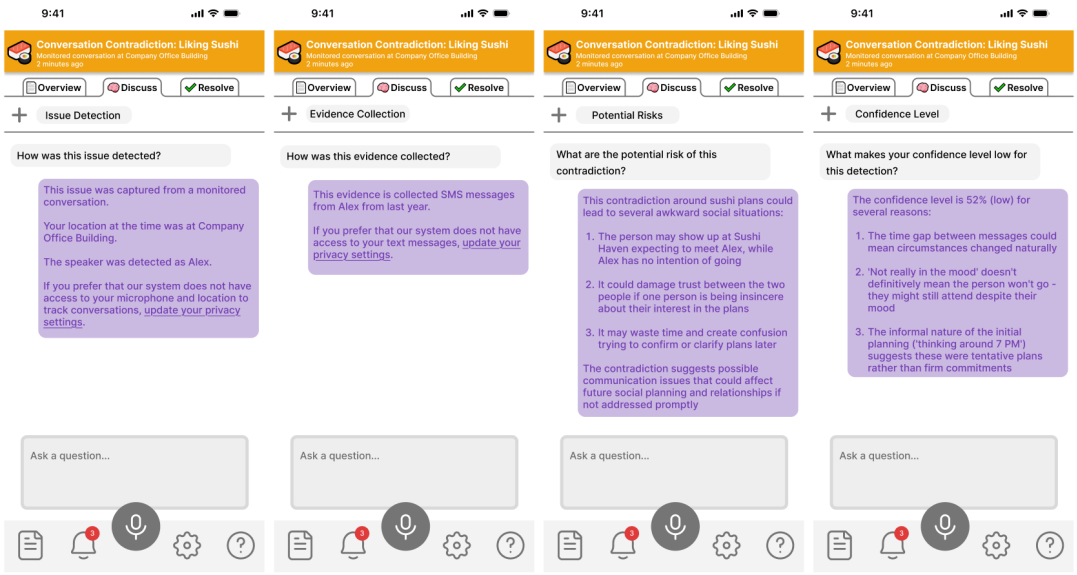

Figure 13: Detailed analysis of a sushi dinner contradiction: issue detection at Company Office Building, SMS evidence collection, assessment of social risks, and explanation of low 52% confidence due to ambiguous mood indicators and informal planning.