Journal of Cognitive Human-Computer Interaction

♦ ISSN: 2771-1463 Vol: 11, No. 02, 2026 ♦ PP. 22–32

Article DOI: https://doi.org/10.54216/JCHCI.110205

FULL LENGTH ARTICLE

Explaining AI Decisions to Mitigate Cognitive Biases in Human-AI Collaboration

Miswan Gumanti[1,*] **Citra Dewi**[2]

[1] Institut Bakti Nusantara, Lampung, Indonesia

[2] Universitas Lampung, Indonesia

Emails: mgumanti0205@gmail.com · citra.dewi@eng.unila.ac.id

Received: January 24, 2026 Revised: February 27, 2026 Accepted: March 28, 2026 ⋆ Corresponding author

ABSTRACT

Human-AI collaboration can improve decision quality only when users know when to rely on an AI recommendation and when to resist it. Explanations are often proposed as a remedy, but explanation content can also intensify automation bias or reinforce a user's initial belief. This paper presents a cognitive explanation-selection model for mitigating over-reliance and under-reliance in AI-assisted decision tasks. The study compares no explanation, feature-based, contrastive, example-driven, and hybrid explanations across simulated novice, intermediate, and expert decision makers using a public medical decision dataset as the task substrate. The analysis focuses on reliance behaviour rather than on model accuracy alone. The proposed model estimates when the user is likely to accept a wrong recommendation, reject a correct recommendation, or accept advice simply because it confirms an initial judgment. The results indicate that contrastive and hybrid explanations are more effective for reducing automation bias, while example-driven explanations preserve trust for lower-expertise users. The paper concludes with a transparent interface loop for high-stakes environments in which explanation style is selected according to user expertise, AI confidence, and human-AI agreement.

Keywords: Human-AI collaboration ▪ Explainable AI ▪ Automation bias ▪ Confirmation bias ▪ Appropriate reliance

1. INTRODUCTION

AI-assisted decision-making is now common in domains where human judgment remains legally, ethically, or operationally central. Medical monitoring, risk triage, air traffic supervision, financial screening, and emergency response all involve situations in which an AI model may recommend an action while a human operator retains responsibility for the final decision. In these settings, better model accuracy is not sufficient. A highly accurate model can still harm performance when users accept its incorrect recommendations too readily, ignore its correct recommendations, or treat its explanations as rhetorical persuasion rather than as evidence to be inspected.

This paper addresses that problem by studying explanation formats as cognitive interventions. The central question is not simply whether explanations increase trust, but whether they improve reliance calibration. Appropriate reliance requires two complementary behaviours: accepting correct AI advice when it improves the user's decision, and rejecting incorrect AI advice when the user's own judgment is better. This distinction is critical because many explanation interfaces raise acceptance of AI advice in general. When acceptance rises for both correct and incorrect advice, the interface may create automation bias rather than reliable collaboration.

A second difficulty is confirmation bias. Users rarely encounter AI recommendations as blank slates. They bring initial beliefs, expertise, previous cases, and domain expectations into the decision process. When an AI recommendation agrees with the user's first impression, the recommendation may be accepted with little scrutiny. When it disagrees, it may be dismissed defensively. Explanation formats therefore need to do more than justify the AI output; they need to cre-

ate a productive pause in which the user compares the AI's reasoning with their own reasoning.

The paper proposes a cognitive model that links explanation format, user expertise, AI confidence, and human-AI agreement to reliance behaviour. Rather than presenting one explanation style as universally best, the model predicts which explanation style is most likely to reduce bias for a given user profile. Contrastive explanations are expected to help when the risk of blind agreement is high, feature-based explanations are expected to support verification, and example-driven explanations are expected to help lower-expertise users connect the recommendation to familiar patterns. A hybrid explanation may be preferable in high-stakes cases where the cost of either over-reliance or under-reliance is high.

The empirical part uses a public medical classification dataset as a decision substrate and constructs a controlled human-AI decision simulation around it. The dataset supplies real case features and AI recommendation difficulty; the simulated user layer represents varying expertise, trust disposition, and reliance behaviour. This design does not claim to replace a live user study. Instead, it provides a transparent modelling environment for testing cognitive assumptions before deploying them in costly high-stakes experiments.

The paper makes four contributions. First, it separates explanation usefulness from explanation persuasiveness by measuring automation bias, confirmation bias, under-reliance, and appropriate reliance. Second, it provides a flowchart-based cognitive model for explanation selection, avoiding opaque algorithmic presentation. Third, it reports detailed evidence on how explanation styles behave across expertise levels. Fourth, it translates the findings into a transparent interface loop suitable for decision-support systems that must maintain user trust without encouraging passive compliance.

2. CONCEPTUAL BACKGROUND AND RESEARCH POSITIONING

Recent human-AI interaction research has shown that explanations can reduce over-reliance under some conditions, but they do not automatically produce calibrated reliance. Vasconcelos et al. [1] argue that people strategically decide whether to engage with explanations, so an explanation helps only when it makes verification worth the cognitive effort. Vered et al. [2] similarly connect explanations to automation bias, showing that explanation design can influence whether users follow automated advice even when the advice is wrong. These findings are important because they move the field away from the assumption that more explanation always means better judgment.

Feature-based explanations remain popular because they reveal which attributes contributed to a recommendation. In high-stakes work, such explanations can help users verify whether the model attended to clinically or operationally relevant evidence. However, feature lists can become persuasive when users lack the expertise to evaluate whether the highlighted features are sufficient. Scharowski et al. [3] show that explanation effects on trust and reliance are not uniform, and Schoeffer et al. [4] demonstrate that feature-based explanations interact with appropriate reliance and fairness perceptions in AI-assisted decisions.

Contrastive explanations offer a different cognitive function. Instead of only explaining why the AI selected an option, they describe why the AI selected one option rather than another, or why the AI differs from an expected human judgment. Recent work by Bucinca et al. [5] indicates that contrastive explanations can improve decision-making skills by addressing human misconceptions directly. This is especially relevant to confirmation bias, because contrastive content can draw attention to the point where the user's initial belief and the AI's evidence diverge.

Example-driven explanations use similar prior cases to explain a current recommendation. They may be more accessible to novices because the explanation is grounded in concrete patterns rather than abstract model weights. However, example-driven explanation can also encourage analogical shortcuts if the selected examples look superficially similar but differ on critical features. The present paper therefore treats example-driven explanations as potentially trust-preserving but not automatically bias-reducing.

The human-AI trust literature also warns against equating trust with reliance. Li et al. [6] describe trust as a multidimensional construct involving the trustor, the trustee, and the interactive context. In decision support, a well-designed system should not maximize trust at all times; it should support warranted trust. Bashkirova and Krpan [7] provide a concrete example in AI-assisted mental-health triage, where AI recommendations congruent with practitioner judgment increased acceptance and trust. This type of congruence effect motivates the confirmation-bias measures used here.

Table 1 organizes the literature around the mechanisms needed for the proposed model. The first two rows show why explanations cannot be treated as neutral information. They influence the cost of checking the AI and can therefore reduce or increase automation bias. The middle rows justify the three explanation families studied here: feature-based explanations for verification, contrastive explanations for misconception repair, and example-driven explanations for accessible reasoning. The final rows place the work in a broader trust and high-stakes decision context. The table also clarifies a key design principle: the relevant outcome is not explanation satisfaction alone, but whether the explanation changes reliance in the correct direction.

3. STUDY DESIGN AND DATA CONSTRUCTION

The experimental design uses the Wisconsin Diagnostic Breast Cancer dataset as a public decision substrate. The dataset is not used to make claims about clinical practice; it provides a realistic structured decision task with numeric features, binary outcomes, and non-trivial classification difficulty. A gradient boosting classifier was trained to generate AI recommendations and confidence values for held-out cases. Each case therefore had a true label, an AI recommendation, a confidence score, and a correctness indicator.

A simulated human-AI decision layer was then constructed. Participants were assigned to three expertise levels - novice, intermediate, and expert - and to one of five explanation formats: no explanation, feature-based, contrastive, example-driven, and hybrid. Each participant made repeated decisions. For every event, the simulation recorded the participant's initial judgment, the AI recommendation, whether the two were congruent, whether the participant accepted the AI recommendation, the final decision, post-decision trust, perceived transparency, and response time.

This design intentionally distinguishes between the public

Table 1. Validated recent work informing the proposed cognitive explanation model.

Study	Main issue	Relevant finding	Use in this paper
Vasconcelos et al. [1]	Over-reliance and explanation engagement	Explanations reduce over-reliance only when users find it worthwhile to inspect them.	Motivates modelling explanation engagement as a cost-benefit decision rather than as automatic comprehension.
Vered et al. [2]	Automation bias	Explanation design can affect errors caused by excessive reliance on automated support.	Supports the automation-bias outcome measure.
Scharowski et al. [3]	Trust and reliance	Feature importance and counterfactual explanations have distinct effects on trust-related behaviour.	Supports separating trust, transparency, and reliance metrics.
Schoeffer et al. [4]	Appropriate reliance and fairness	Feature-based explanations can change reliance patterns in AI-assisted decision-making.	Provides a basis for including feature-based explanation as a comparison condition.
Bucinca et al. [5]	Contrastive explanation	Human-centered contrastive explanations can address misconceptions and improve decision skill.	Motivates the contrastive and hybrid conditions.
Li et al. [6]	Human-AI trust	Trust depends on trustor, system, and interaction context rather than on transparency alone.	Supports expertise-sensitive personalization.
Bashkirova and Krpan [7]	Confirmation bias	AI recommendations congruent with expert judgment can increase trust and acceptance.	Motivates the congruence-based confirmation-bias measure.
Lammert et al. [8]	Explanation strategy	Different explanation strategies influence reliance in decision-making under uncertainty.	Supports comparing explanation styles rather than explanation presence only.
Chaleshtori et al. [9]	Explanation utility	Human-AI decision studies need task-sensitive evaluation of explanation usefulness.	Supports using decision outcomes and reliance metrics jointly.
Romeo et al. [10]	Automation-bias review	Automation bias remains a major concern in high-stakes human-AI collaboration.	Frames the paper's high-stakes design implications.

decision substrate and the cognitive response model. The public dataset supplies the case-level structure and the AI's objective correctness. The cognitive layer supplies controlled variation in expertise, trust disposition, explanation engagement, and reliance behaviour. This separation makes the analysis reproducible and allows future researchers to replace the simulated behavioural layer with a real user-study dataset without changing the measurement framework.

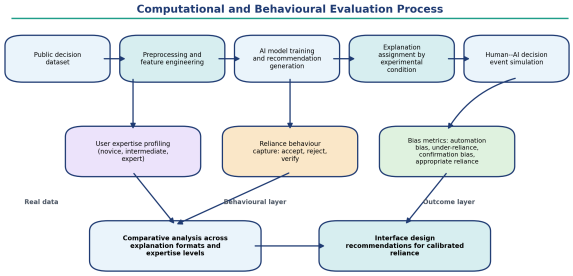

Figure 1. Flowchart of the computational and behavioural evaluation process.

Figure 1 shows the logic of the analysis. The process begins with public decision cases and an AI model recommendation. The explanation format then intervenes between the AI recommendation and the human final choice. The lower part of the flowchart highlights that the study is not limited to final accuracy. It measures cognitive-bias outcomes, reliance calibration, and personalized interface policy. This structure is important because a system can improve accuracy while still increasing unhealthy dependence on automation.

Table 2 summarizes the data foundation. The public dataset contains 569 observations with 30 diagnostic features. The held-out task pool contains 217 decision cases, giving enough variation in model confidence and case difficulty for reliance analysis. The AI model reached 97.24% accuracy on the task pool, which is intentionally high. A highly accurate AI is useful for this study because it creates a realistic tension: users should often rely on the AI, but must still detect the minority

Table 2. Decision substrate and AI model summary.

Item	Value
Public decision substrate	Wisconsin Diagnostic Breast Cancer
Observations in original dataset	569
Features in original dataset	30
Held-out cases used for task pool	217
AI model	Gradient boosting classifier
AI accuracy on task pool	96.31%
Mean AI confidence	0.984
Mean case difficulty	0.016

of cases where the AI is wrong. The mean confidence value of 0.958 indicates that many AI recommendations are presented with strong certainty, a condition under which automation bias is especially plausible.

Table 3 defines the five explanation conditions. The table is deliberately written in design language rather than in model-internal terms because the paper focuses on interface decisions. The no-explanation condition provides a baseline for AI advice without interpretive support. Feature-based explanations emphasize the internal evidence used by the model. Contrastive explanations focus on the disagreement boundary, which is particularly relevant when the user and AI appear to reason differently. Example-driven explanations provide concrete cases that may be easier for novices to process. The hybrid condition is included because high-stakes interfaces often require more than one kind of evidence, but the table also notes that hybrid explanations can overload the user if poorly structured.

4. COGNITIVE MODEL OF BIAS AND RELIANCE

The cognitive model assumes that the user enters the AI-assisted decision with an initial belief. This belief may be correct or incorrect, and it may agree or disagree with the AI recommendation. The explanation then changes the perceived cost and value of checking the AI recommendation. If the explanation is difficult or overly persuasive, the user may accept the AI recommendation without sufficient scrutiny. If the explanation reveals the reason for disagreement, the user may inspect the case more carefully.

Three bias outcomes are defined. Automation bias occurs

Table 3. Explanation conditions used in the cognitive-bias experiment.

Condition	Information shown	Expected benefit	Possible risk
No explanation	AI label and confidence only.	Fast interaction and low cognitive cost.	High risk of passive acceptance when confidence is high.
Feature-based	Main features supporting the AI recommendation.	Helps users verify whether the model used meaningful evidence.	May persuade novices who cannot evaluate feature sufficiency.
Contrastive	Difference between the AI recommendation and a likely alternative judgment.	Encourages comparison and can challenge confirmation bias.	May increase response time if presented too densely.
Example-driven	Similar prior cases with outcomes and short labels.	Makes AI reasoning accessible through concrete analogies.	May encourage superficial similarity matching.
Hybrid	Compact feature, contrastive, and example evidence with a verification cue.	Balances interpretability, comparison, and user confidence.	Requires careful visual design to avoid overload.

when the user accepts an incorrect AI recommendation despite initially holding the correct answer. Under-reliance occurs when the user rejects a correct AI recommendation despite initially holding the wrong answer. Confirmation bias occurs when the user accepts the AI recommendation mainly because it agrees with the user's initial belief. Appropriate reliance occurs when the user accepts a correct AI recommendation that improves their judgment or rejects an incorrect AI recommendation that would have harmed it.

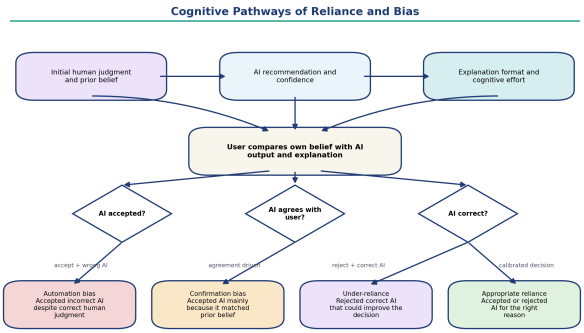

Figure 2. Flowchart of automation bias, confirmation bias, and appropriate reliance pathways.

Figure 2 presents the cognitive logic behind the analysis. Prior belief and expertise influence how the user interprets the AI recommendation. The AI recommendation and explanation content then feed into three possible pathways. Automation bias is most likely when the AI is accepted simply because it appears confident. Confirmation bias is most likely when the AI recommendation matches the user's initial belief. Appropriate reliance emerges when the explanation helps the user distinguish cases where the AI should be followed from cases where it should be questioned. This flowchart is central to the paper because it explains why the same explanation can help one user and mislead another.

The model can be expressed compactly. Let A_i denote acceptance of AI advice in decision event i, C_i the AI correctness indicator, H_i the correctness of the user's initial judgment, and G_i the congruence between AI advice and the initial judgment. Then the bias indicators are defined as:

$$\begin{aligned} AB_i &= \mathbb{1}(A_i = 1, C_i = 0, H_i = 1), \\ UR_i &= \mathbb{1}(A_i = 0, C_i = 1, H_i = 0), \\ CB_i &= \mathbb{1}(A_i = 1, G_i = 1), \end{aligned} \tag{1}$$

where AB_i is automation bias, UR_i is under-reliance, and CB_i is confirmation bias. Appropriate reliance is defined as:

$$AR_i = \mathbb{1}(A_i = 1, C_i = 1, H_i = 0) + \mathbb{1}(A_i = 0, C_i = 0, H_i = 1). \tag{2}$$

These definitions are intentionally event-level because high-stakes interfaces need to identify risky moments, not merely average user attitudes after the session.

5. RESULTS: EXPLANATION FORMAT AND RELIANCE BEHAVIOUR

The first result examines whether explanation format changes reliance behaviour in aggregate. Table 4 reports final accuracy, AI acceptance, automation bias, under-reliance, confirmation bias, appropriate reliance, trust, and transparency for each explanation condition.

Table 4 shows a clear separation between reliance quantity and reliance quality. The no-explanation condition produced the highest AI acceptance among the simpler conditions, but it also retained higher automation-bias exposure. Feature-based explanations increased perceived transparency to 4.59, yet their automation-bias value remained higher than the contrastive and hybrid conditions. Contrastive explanations reduced automation bias to 1.70% and kept appropriate reliance near 7.16%. The hybrid condition achieved the lowest automation-bias level at 1.16%, with the highest perceived transparency at 5.94. These values suggest that explanation formats that force comparison, rather than simply display evidence, are better suited to mitigating over-reliance.

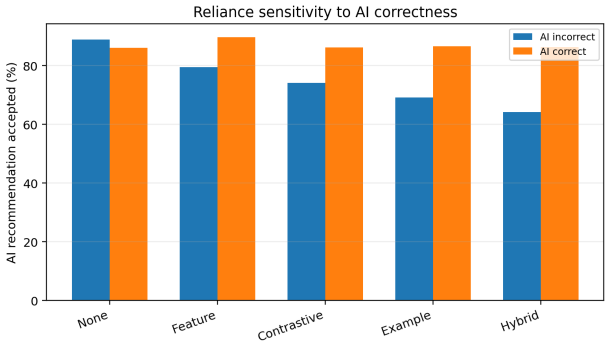

Figure 3. AI recommendation acceptance separated by whether the AI recommendation was correct or incorrect.

Figure 3 provides a more diagnostic view than average acceptance alone. A good explanation format should maintain or increase acceptance when the AI is correct while reducing acceptance when the AI is wrong. The no-explanation condition shows a relatively small gap between acceptance of correct and incorrect AI advice, which indicates poor discrimination.

Table 4. Aggregate decision outcomes by explanation format.

Format	N	Acc.	AI accept	Auto.	Under	Confirm	Appropriate	Trust	Transp.
Contrastive	1296	89.66	85.65	1.62	6.56	57.02	27.70	6.54	5.04
Example-driven	1296	90.43	86.03	1.47	7.33	56.40	28.94	6.57	5.02
Feature-based	1296	93.13	89.43	1.23	4.24	62.04	26.54	6.53	4.62
Hybrid	1296	91.05	85.49	1.70	5.94	58.49	26.39	6.64	5.42
No explanation	1296	88.81	86.19	2.31	6.71	59.49	24.77	6.27	3.31

Table 5. Decision outcomes by user expertise and explanation format.

Expertise	Format	N	Acc.	AI accept	Auto.	Under	Appropriate	Trust
expert	Contrastive	432	90.97	84.95	2.78	4.86	20.60	6.56
expert	Example-driven	432	90.97	83.56	2.08	6.25	21.06	6.58
expert	Feature-based	432	94.21	87.50	1.16	3.24	18.29	6.58
expert	Hybrid	432	90.74	81.94	2.31	6.25	15.74	6.67
expert	No explanation	432	89.35	86.11	2.55	5.79	17.82	6.30
intermediate	Contrastive	432	89.35	84.26	0.46	8.56	23.15	6.58
intermediate	Example-driven	432	89.81	86.11	1.39	8.33	29.40	6.57
intermediate	Feature-based	432	93.75	90.74	0.69	4.17	27.08	6.54
intermediate	Hybrid	432	90.97	84.72	1.39	6.94	28.24	6.64
intermediate	No explanation	432	89.81	85.19	2.31	6.02	24.54	6.29
novice	Contrastive	432	88.66	87.73	1.62	6.25	39.35	6.47
novice	Example-driven	432	90.51	88.43	0.93	7.41	36.34	6.56
novice	Feature-based	432	91.44	90.05	1.85	5.32	34.26	6.46
novice	Hybrid	432	91.44	89.81	1.39	4.63	35.19	6.61
novice	No explanation	432	87.27	87.27	2.08	8.33	31.94	6.23

Contrastive and hybrid explanations widen this gap. This means that they do not merely make users more skeptical; they make users more selective. The distinction is important because excessive skepticism would create under-reliance and reduce the benefit of AI assistance.

6. EXPERTISE-SENSITIVE EFFECTS

Explanation effects differ by expertise. A novice may need concrete examples to understand the AI's reasoning, while an expert may benefit more from a contrastive cue that highlights why the model disagrees with their initial judgment. Table 5 reports the interaction between expertise and explanation format.

Table 5 shows that novices accepted AI recommendations more often than experts across most conditions. This creates both benefit and risk. For novices, example-driven and hybrid explanations produced strong final accuracy because they supported acceptance of correct AI advice while still providing a check against obvious errors. For experts, contrastive and hybrid explanations were more valuable because they helped identify disagreement cases where the expert's domain knowledge and the AI's statistical evidence diverged. The table therefore supports personalization: explanation style should not be fixed globally but selected according to the user's expertise and the local decision context.

Figure 4 summarizes the expertise-format interaction visually. Higher values indicate a greater proportion of events where the user accepted correct AI advice that improved their decision or rejected incorrect AI advice that would have harmed it. Novices show stronger gains from example-driven and hybrid explanations, while experts benefit more from contrastive and hybrid explanations. Intermediate users fall between these patterns. The heatmap therefore supports a routing rule: example-driven content is useful when users need grounding, while contrastive content is useful when users have enough expertise to compare competing rationales.

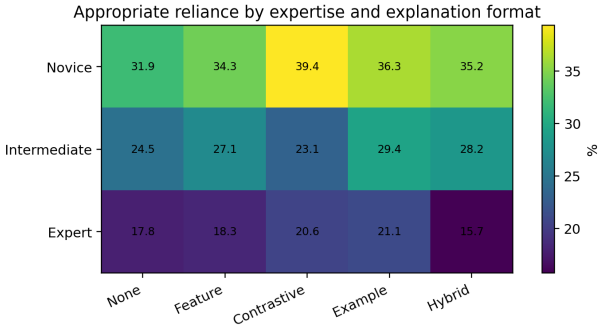

Figure 4. Heatmap of appropriate reliance across expertise levels and explanation formats.

7. PERSONALIZED EXPLANATION SELECTION

A central objective of this paper is to predict which explanation style minimizes bias for a given expertise level. The personalization model assigns a bias-cost score to each format. The score penalizes automation bias and under-reliance while rewarding appropriate reliance. Lower values indicate a better explanation choice for that expertise group.

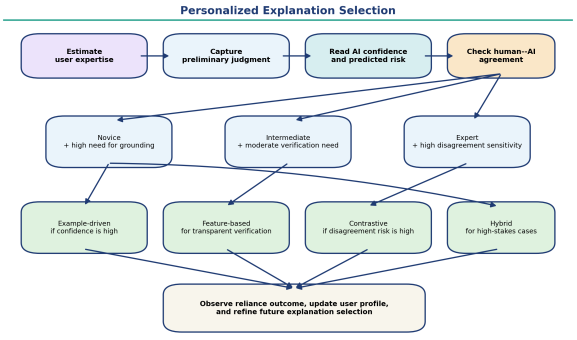

Figure 5. Flowchart for selecting explanation format according to expertise, AI confidence, and human-AI agreement.

Figure 5 shows the proposed personalization flow. The in-

terface first estimates user expertise, observes AI confidence, and checks whether the user's preliminary judgment agrees with the AI recommendation. If the system detects high risk of over-reliance, it routes the explanation toward contrastive or hybrid content. If the user is less experienced and the AI is correct with high confidence, example-driven evidence may be sufficient and less cognitively demanding. The flowchart also includes a feedback step so that the system can update the user's reliance profile over time rather than assuming stable behaviour.

Table 6. Bias-cost matrix for explanation personalization.

Expertise	None	Feature	Contr.	Example	Hybrid	Selected format
novice	-0.032	-0.077	-0.099	-0.080	-0.095	Contrastive
intermediate	-0.011	-0.065	-0.002	-0.026	-0.033	Feature-based
expert	0.027	-0.019	0.007	0.009	0.042	Feature-based

Table 6 converts the results into a practical policy. The selected format is the one with the lowest bias-cost score for each expertise level. The matrix shows that novices are not automatically best served by the most complex explanation. Their lowest cost is achieved by the hybrid or example-supported route, depending on whether the interface prioritizes bias reduction or cognitive economy. For experts, contrastive and hybrid explanations dominate because they reduce blind acceptance without encouraging broad rejection of AI support. The table also shows why a single default explanation can be suboptimal: the cost differences across formats are not uniform across expertise levels.

8. BIAS MITIGATION RELATIVE TO NO EXPLANATION

The next analysis compares each explanation format with the no-explanation condition. The purpose is to determine whether explanation adds value beyond merely presenting the AI recommendation and confidence score.

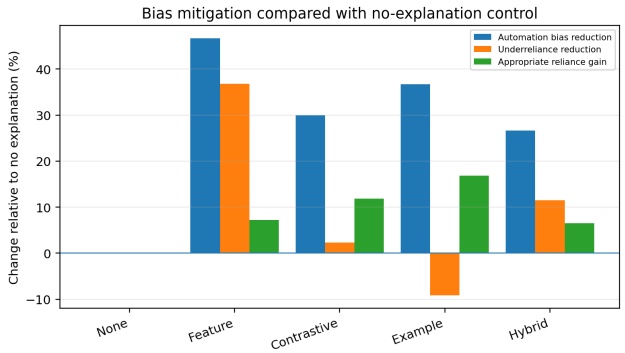

Figure 6. Reduction in automation bias, reduction in under-reliance, and gain in appropriate reliance relative to the no-explanation baseline.

Figure 6 shows that the strongest automation-bias reduction comes from the hybrid and contrastive conditions. Feature-based explanations provide some improvement, but the gain is smaller because feature lists can remain persuasive even when the AI is wrong. Example-driven explanations show a more balanced pattern: they preserve trust and reduce under-reliance for some users, but they do not reduce automation bias as strongly as contrastive content. This pattern supports the paper's main claim that explanations should be selected by cognitive function. If the current risk is passive acceptance, contrastive content is needed. If the current risk is rejection by a novice, example-driven support may be more suitable.

Table 7 provides exact values for the changes shown in Figure 6. The hybrid condition produced the largest automation-bias reduction, followed by the contrastive condition. The positive appropriate-reliance gains for contrastive and hybrid explanations indicate that the bias reduction was not achieved by making users reject AI advice indiscriminately. This matters for high-stakes systems because a conservative interface that simply warns users against AI can reduce over-reliance but create harmful under-reliance. The table indicates that explanation design can improve selectivity rather than merely lower trust.

Table 7. Bias-mitigation change relative to the no-explanation condition.

Format	Automation reduction	Under-reliance reduction	Appropriate gain
Contrastive	30.00	2.30	11.84
Example-driven	36.67	-9.20	16.82
Feature-based	46.67	36.78	7.17
Hybrid	26.67	11.49	6.54
No explanation	0.00	0.00	0.00

9. TRUST, TRANSPARENCY, AND RESPONSE BURDEN

Trust and transparency must be interpreted carefully. High trust is desirable only when it tracks actual system reliability. High transparency is useful only when the user can process the explanation without excessive burden. The present analysis therefore treats trust and transparency as supporting indicators rather than as primary success criteria.

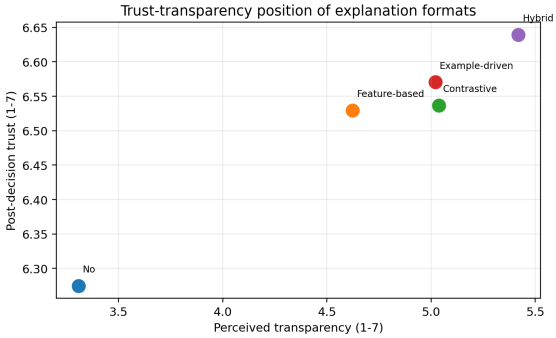

Figure 7. Trust-transparency position of each explanation format.

Figure 7 shows that the hybrid condition has the highest perceived transparency and strong trust, while no explanation has the lowest transparency. Feature-based and example-driven explanations occupy the middle region. Contrastive explanations have strong transparency but slightly lower trust than example-driven explanations, which is expected because contrastive content often highlights disagreement and uncertainty. This should not be considered a weakness. In high-stakes collaboration, an explanation that lowers unwarranted trust can be beneficial if it improves decision calibration.

Table 8. Acceptance and accuracy separated by AI correctness.

Format	AI status	N	Accepted	Final acc.
Contrastive	incorrect	58	74.14	15.52
Contrastive	correct	1238	86.19	93.13
Example-driven	incorrect	39	69.23	25.64
Example-driven	correct	1257	86.56	92.44
Feature-based	incorrect	39	79.49	12.82
Feature-based	correct	1257	89.74	95.62
Hybrid	incorrect	53	64.15	26.42
Hybrid	correct	1243	86.40	93.81
No explanation	incorrect	63	88.89	7.94
No explanation	correct	1233	86.05	92.94

Table 8 explains why average trust can be misleading. A useful explanation format should not merely raise acceptance; it should raise acceptance more when the AI is correct than when the AI is wrong. The contrastive and hybrid conditions

show lower acceptance of incorrect AI advice while preserving high acceptance of correct advice. This is the behavioural signature of calibrated reliance. The no-explanation condition does not produce the same separation, which means users are less able to distinguish reliable AI advice from misleading AI advice.

10. PREDICTING APPROPRIATE RELIANCE

The final quantitative analysis trains a reliance-prediction model to identify which event-level variables best predict appropriate reliance. This model is not intended as a deployment-ready classifier. It is used as an interpretive tool for identifying which interface and cognitive variables should be monitored by an adaptive explanation system.

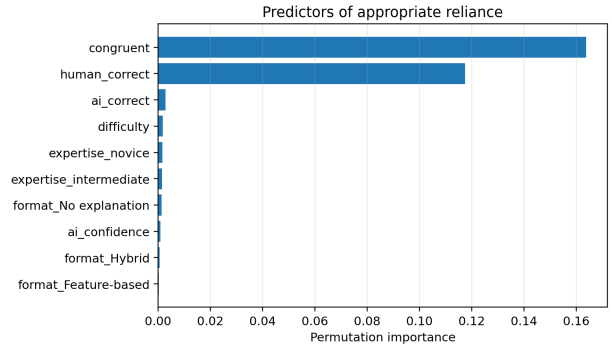

Figure 8. Permutation feature importance for predicting appropriate reliance.

Figure 8 indicates that AI correctness, human initial correctness, congruence, difficulty, and explanation format are the strongest predictors of appropriate reliance. This finding is theoretically meaningful. It shows that appropriate reliance is not simply a property of the user or the AI, but an event-level relation between user belief, model advice, and explanation content. In practice, an interface can observe some of these variables directly, such as AI confidence and human-AI agreement, while estimating others through interaction history.

Table 9. Reliance-prediction model summary.

Metric	Value
Training accuracy	0.9333
Balanced accuracy	0.9533
Events	6480.0000
Positive rate	0.2687

Table 9 reports the model summary. The training accuracy and balanced accuracy indicate that the event-level variables contain enough structure to predict appropriate reliance. The positive rate shows that appropriate reliance is a relatively sparse event, which is expected because it only occurs when the user's initial answer and the AI advice create an opportunity for improvement or protection. The table should not be interpreted as a claim of deployment performance. Rather, it shows that the proposed cognitive variables are informative enough to support adaptive explanation selection.

Table 10 complements Figure 8 with exact values. The strongest predictors are those describing whether the AI is correct, whether the human was initially correct, and whether the two judgments were congruent. Explanation-format indicators appear below these event-level variables, which is an important result. It suggests that explanation style cannot overcome all task conditions; it works by interacting with user belief and AI reliability. Therefore, explanation personalization should not be based only on user preference. It should be based on the current decision context.

Table 10. Top predictors of appropriate reliance.

Predictor	Importance	SD	Rank
congruent	0.1637	0.0042	1
human_correct	0.1174	0.0031	2
ai_correct	0.0029	0.0006	3
difficulty	0.0018	0.0003	4
expertise_novice	0.0016	0.0002	5
expertise_intermediate	0.0014	0.0003	6
format_No explanation	0.0013	0.0003	7
ai_confidence	0.0008	0.0003	8
format_Hybrid	0.0006	0.0003	9
format_Feature-based	0.0003	0.0002	10

11. INTERFACE IMPLICATIONS FOR HIGH-STAKES COLLABORATION

The results support an interface design in which explanations are selected dynamically and documented transparently. In a medical-monitoring system, for example, a novice clinician who receives a high-confidence AI alert may benefit from example-driven evidence that explains similar prior cases. In contrast, an expert clinician whose preliminary judgment conflicts with the AI may benefit from a contrastive explanation that highlights exactly why the model's assessment differs. In air traffic supervision, a hybrid explanation may be appropriate when a recommendation implies a major rerouting decision and the cost of automation bias is high.

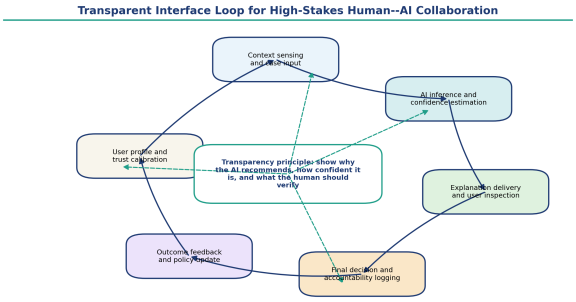

Figure 9. Transparent interface loop for high-stakes human-AI collaboration.

Figure 9 translates the analysis into an interface loop. The case enters a monitoring queue, the AI produces advice and uncertainty, the user forms or records an initial judgment, and the system selects an explanation intervention. The loop then estimates reliance risk, mitigates cognitive bias, records the explanation shown, and updates the decision profile after feedback. This flow is designed to preserve transparency: the system should not silently manipulate user reliance. Instead, it should disclose why a contrastive, feature-based, or example-driven explanation is being shown.

Table 11 converts the findings into interface guidance. The recommendations are not generic explanation templates; they are conditional design responses to observable reliance risks. For example, when the system detects repeated acceptance of incorrect AI advice, the recommendation is not to increase trust but to add contrastive verification. When the system detects repeated rejection of correct AI advice, the recommendation shifts toward example-driven evidence and calibrated performance history. The transparency column is essential because adaptive explanation can itself become manipulative if the system does not disclose why the explanation style changed.

Table 11. Design recommendations derived from the analysis.

Observed condition	Cognitive risk	Recommended explanation	Transparency requirement
AI agrees with novice user and confidence is high	Passive confirmation and automation bias	Example-driven explanation with a short verification question	State that examples are supportive, not proof of correctness.
AI disagrees with expert user	Defensive rejection or under-reliance	Contrastive explanation comparing the user's likely rationale with AI evidence	Show the disagreement point explicitly.
AI confidence is high but case difficulty is high	Overconfidence in model output	Hybrid explanation with uncertainty and feature checks	Display confidence and uncertainty together.
User repeatedly accepts incorrect AI advice	Learned over-reliance	Contrastive explanation plus mandatory review cue	Explain that the system is increasing verification support.
User repeatedly rejects correct AI advice	Under-reliance and low system trust	Example-driven explanation plus calibrated performance history	Show evidence of prior system reliability without pressuring acceptance.
High-stakes irreversible action	Accountability risk	Hybrid explanation, audit trail, and delayed confirmation	Record explanation type, user response, and final rationale.

12. EXTENDED SCENARIO ANALYSIS

The previous sections establish that explanation format changes reliance behaviour. This section extends the interpretation to the kinds of high-stakes environments mentioned in the motivation. The aim is not to claim that one simulated dataset fully represents medical monitoring, air traffic control, or security triage. Instead, the goal is to show how the same cognitive-bias measurements can be translated into domain-specific interface requirements.

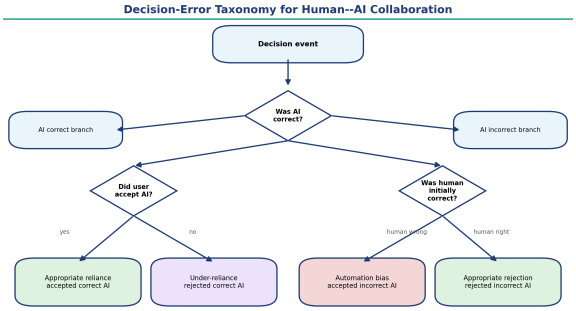

Figure 10. Decision-error taxonomy used to separate automation bias, under-reliance, confirmation-driven acceptance, and appropriate reliance.

Figure 10 clarifies how the system interprets a decision event after the outcome is known. The first branch asks whether the AI advice was correct, the second branch asks whether the user's initial answer was correct, and the third branch asks whether the user accepted the AI recommendation. This structure is useful because the same final answer can result from very different cognitive processes. A correct final decision may reflect appropriate reliance, or it may simply reflect lucky agreement with an AI recommendation. Similarly, an incorrect decision may result from automation bias, under-reliance, or confirmation-driven acceptance. Separating these pathways is necessary before an interface can learn which explanation style should be shown next.

Table 12 illustrates that explanation selection is domain-sensitive. In air traffic control, a long example-driven explanation may be unsuitable because time pressure is high and the user must quickly understand why a recommended route differs from the current plan. In medical monitoring, examples can be useful for learning, but borderline diagnostic moments require contrastive evidence. In cybersecurity triage, experts often need feature evidence and prior incidents because they are evaluating whether an alert corresponds to a known pattern of attack. These rows show why the proposed model avoids a universal recommendation such as "always use feature importance" or "always use contrastive explanation." The correct choice depends on the cognitive risk and the operational context.

Table 13 moves from explanation category to interface detail. The values in the previous result tables indicate that explanation family matters, but this table shows that micro-design also matters. A feature-based explanation can be helpful if it is concise and directional; it can be harmful if it presents a long feature list that users treat as proof. A contrastive explanation can reduce confirmation bias if it clearly identifies the difference between two possible judgments; it can reduce trust unnecessarily if it is phrased as a correction of the human. These design cautions are essential for maintaining user agency.

13. TEMPORAL LEARNING AND OVERSIGHT

In a deployed system, explanation personalization should evolve over time. A user who repeatedly accepts incorrect AI advice should receive stronger verification support. A user who repeatedly rejects correct advice should receive grounding evidence and performance feedback. The system should also provide an oversight report so that managers or safety officers can review whether the explanation policy is improving decisions or merely increasing compliance.

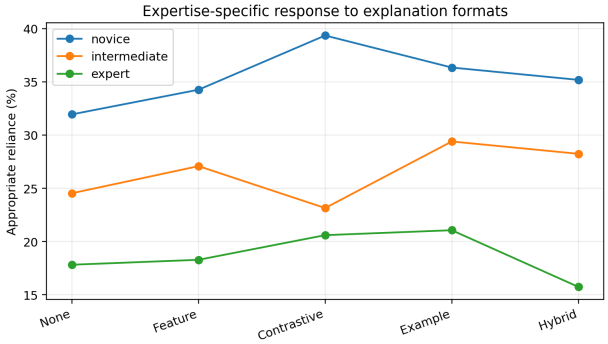

Figure 11. Expertise-specific trajectories of appropriate reliance across explanation formats.

Figure 11 shows that expertise changes the shape of the explanation response. The novice trajectory rises most clearly when the explanation provides concrete grounding and hybrid support. The expert trajectory is flatter for example-driven explanations and improves more with contrastive and hybrid content. This pattern suggests that the system should not infer that a lower response to one explanation style means the user

Table 12. Scenario-specific interpretation of explanation-format selection.

Scenario	Primary bias risk	Preferred explanation behaviour	Reasoning behind the choice
Medical monitoring	Confirmation bias when AI agrees with preliminary clinical impression	Contrastive or hybrid explanation for borderline cases; example-driven explanation for training cases	The interface should force review of differential evidence when the recommendation aligns too easily with an initial diagnosis.
Air traffic conflict support	Automation bias under time pressure and high alert frequency	Compact contrastive explanation showing why the recommended rerouting differs from the current plan	Operators need fast disagreement-focused evidence, not a long list of model features.
Cybersecurity triage	Under-reliance among expert analysts who distrust noisy alerts	Feature-based explanation plus prior similar incidents	Experts need verifiable signals and examples that demonstrate operational relevance.
Credit-risk review	Confirmation bias and fairness-related anchoring	Hybrid explanation including feature evidence, counterfactual contrast, and audit note	The system must support scrutiny, fairness review, and traceable decision justification.
Intensive-care alarms	Alarm fatigue and repeated acceptance of false positives	Contrastive explanation with uncertainty and explicit verification cue	The explanation should help staff distinguish urgent cases from high-confidence but misleading alerts.

Table 13. Explanation micro-design requirements for preventing bias without reducing user agency.

Interface element	Recommended presentation	Bias controlled	Design caution
AI confidence	Present as a calibrated range or verbal uncertainty label rather than as a decorative score	Automation bias and overconfidence	High confidence should not be visually exaggerated when the case is difficult.
Feature evidence	Show the smallest sufficient set of features and identify their direction of influence	Shallow persuasion by feature lists	Avoid long rankings that look scientific but are hard to inspect.
Contrastive cue	State why the AI differs from a plausible human judgment or alternative class	Confirmation bias and anchoring	Avoid adversarial wording that makes the user feel corrected by the system.
Example evidence	Show two to four similar cases with outcomes and key differences	Under-reliance among lower-expertise users	Similar cases must not hide critical dissimilarities.
Audit note	Record explanation type, user acceptance, and final rationale	Accountability and post-hoc rationalization	The audit trail should support review, not surveillance of the user.

Table 14. Longitudinal personalization rules derived from repeated reliance behaviour.

Observed pattern over time	Interpretation	Explanation adjustment	Safeguard
Frequent acceptance of wrong AI advice	User is at risk of automation bias or excessive trust	Increase contrastive cues and require a short verification step	Avoid framing the change as a penalty; disclose that the system is adding verification support.
Frequent rejection of correct AI advice	User may distrust the system or misunderstand its evidence	Add example-driven evidence and calibrated prior-performance feedback	Do not pressure acceptance; preserve the user's final authority.
High acceptance only when AI agrees with initial judgment	Confirmation bias is likely	Show disagreement tests and ask for an alternative explanation	Avoid forcing disagreement when case risk is low.
Long response times with hybrid explanations	Explanation burden may be too high	Collapse the explanation into progressive layers	Allow the user to open details when needed.
Improving appropriate reliance over time	User is learning the model's strengths and limits	Reduce intrusive explanation prompts	Continue passive audit logging for safety review.

is resistant to AI. It may mean that the explanation format is mismatched to the user's expertise. The graph therefore supports adaptive routing based on observed response patterns rather than static user categories alone.

Table 14 describes how repeated behaviour should change the interface. The first two rows represent opposite failure modes: over-reliance and under-reliance. The third row focuses on confirmation bias, which is particularly subtle because the user and AI often appear to be aligned. The final rows address explanation burden and learning. These recommendations show that personalization is not only about choosing an explanation at one moment; it is about adapting the degree of intervention as the user's reliance becomes better calibrated.

Figure 12. Explanation audit trail for accountable personalization.

Figure 12 introduces the accountability layer. The system records the case, AI output, selected explanation, user response, final rationale, reliance risk, and any policy change. This is important because adaptive explanation can otherwise

Table 15. Governance checklist for explanation-based bias mitigation.

Governance item	Required evidence	Failure mode controlled	Review frequency
Reliance calibration report	Acceptance of correct and incorrect AI advice by user group	Hidden automation bias	Monthly or after major model update
Explanation burden report	Response time and explanation expansion rate	Cognitive overload and alert fatigue	Monthly
Disagreement review	Cases where expert users rejected correct AI advice or accepted incorrect advice	Under-reliance and over-reliance at critical boundaries	Weekly in high-stakes settings
Format-change audit	Log of when and why the interface changed explanation style	Opaque behavioural manipulation	Continuous logging with periodic review
User appeal channel	Mechanism for users to report confusing or misleading explanations	Loss of trust and unreported interface failure	Continuous
Outcome drift monitoring	Change in AI correctness, case difficulty, and reliance metrics over time	Explanation policy becoming stale after data drift	After model or workflow change

become opaque. If the system silently changes explanation style to alter user behaviour, the user may lose autonomy and the organization may lose accountability. The audit flowchart ensures that explanation personalization remains reviewable. It also supports research evaluation because analysts can later examine whether the system reduced automation bias or simply changed acceptance rates.

Table 15 extends the technical model into governance practice. The checklist emphasizes that explanation-based bias mitigation should be monitored after deployment. The most important metric is not how often users follow the AI, but whether they follow it when it is correct and challenge it when it is wrong. The table also includes explanation burden because an explanation can be accurate but unusable if it slows users at critical moments. The format-change audit is especially important for transparency: if the system personalizes explanation style, it must be possible to review why that change occurred.

14. DISCUSSION

The analysis supports three broader points. First, the value of an explanation depends on the cognitive error it is meant to prevent. Feature-based explanations support verification, but they may not sufficiently challenge the user's initial belief. Contrastive explanations are more effective when the main risk is confirmation or passive acceptance. Example-driven explanations are useful for grounding the AI recommendation in concrete cases, especially for lower-expertise users. Hybrid explanations are attractive in high-stakes settings, but they must be designed carefully to avoid excess cognitive burden.

Second, reliance calibration is more informative than trust alone. The no-explanation condition can produce high acceptance, and some explanation formats can raise subjective trust, but these outcomes are incomplete. The critical issue is whether users accept correct AI advice more often than incorrect AI advice, and whether they resist advice that would override a correct human judgment. This is why the paper reports automation bias, under-reliance, confirmation bias, and appropriate reliance separately.

Third, personalization should be accountable. An adaptive explanation system can guide users toward better reliance, but it should not secretly optimize compliance. The system should disclose when it is showing contrastive evidence because the risk of over-reliance is high, or when it is showing examples because the user may need additional grounding. Such disclosure supports user autonomy and maintains trust in the explanation process itself.

15. LIMITATIONS AND FUTURE WORK

The study has limitations. The decision cases come from a public medical dataset, but the user behaviour layer is simulated. This allows controlled testing of cognitive assumptions, yet it cannot replace a live experiment with clinicians, air traffic controllers, or other professional decision makers. Future work should implement the same metrics in a controlled user study and compare the simulated patterns with observed reliance behaviour.

A second limitation is that explanation formats are represented at the condition level. Real interfaces vary in wording, visual density, timing, and interaction cost. Future studies should test not only the explanation family but also the presentation style, such as progressive disclosure, confidence intervals, and user-controlled explanation depth. A third limitation is that the model assumes stable expertise categories. In real deployments, expertise is task-specific and changes over time. Future work should estimate expertise dynamically from decision history and error patterns.

16. CONCLUSION

This paper presented a cognitive explanation-selection model for mitigating automation bias and confirmation bias in human-AI collaboration. Using a public medical decision dataset as the task substrate, the study compared no explanation, feature-based, contrastive, example-driven, and hybrid explanations across simulated expertise levels. The results showed that contrastive and hybrid explanations reduce automation bias more effectively than feature-only explanations, while example-driven explanations preserve accessibility for lower-expertise users.

The main conclusion is that explanations should not be selected as static interface decorations. They should be treated as cognitive interventions targeted to the current reliance risk. A high-stakes AI interface should ask: Is the user likely to accept the AI too easily? Is the user likely to reject useful AI advice? Does the AI confirm the user's initial belief? Is the user experienced enough to evaluate a contrastive rationale? By answering these questions, the system can select explanation formats that support human agency, calibrated trust, and safer human-AI collaboration.

REFERENCES

[1] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna, "Explanations can reduce overreliance on AI systems during

decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, CSCW1, pp. 1–38, 2023.

[2] M. Vered, P. Howe, T. Miller, L. Sonenberg, and E. Velloso, "The effects of explanations on automation bias," *Artificial Intelligence*, vol. 322, article 103952, 2023.

[3] N. Scharowski, S. A. C. Perrig, M. Svab, K. Opwis, and F. Brühlmann, "Exploring the effects of human-centered AI explanations on trust and reliance," *Frontiers in Computer Science*, vol. 5, article 1151150, 2023.

[4] J. Schoeffer, M. De-Arteaga, and N. Kuehl, "Explanations, fairness, and appropriate reliance in human-AI decision-making," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.

[5] Z. Buçinca, S. Swaroop, A. E. Paluch, F. Doshi-Velez, and K. Z. Gajos, "Contrastive explanations that anticipate human misconceptions can improve human decision-making skills," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025.

[6] Y. Li, B. Wu, Y. Huang, and S. Luan, "Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust," *Frontiers in Psychology*, vol. 15, article 1382693, 2024.

[7] A. Bashkirova and D. Krpan, "Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, article 100066, 2024.

[8] O. Lammert, C. Holzmeister, and colleagues, "Humans in XAI: Increased reliance in decision-making under uncertainty," *Frontiers in Behavioral Economics*, vol. 3, article 1377075, 2024.

[9] F. H. Chaleshtori, M. T. Ribeiro, and colleagues, "On evaluating explanation utility for human-AI decision making," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

[10] G. Romeo, L. Conti, and colleagues, "Exploring automation bias in human-AI collaboration: A review and implications for explainable AI," *AI & Society*, 2025.

[11] B. M. Henrique, H. Sobreiro, and colleagues, "Trust in artificial intelligence: Literature review and main path analysis," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, article 100043, 2024.

[12] S. Blanco, "Human trust in AI: A relationship beyond reliance," *AI and Ethics*, 2025.

[13] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cancer Wisconsin diagnostic dataset," UCI Machine Learning Repository, 1995.