



Balancing Trust and Deliberation in Human-AI Decision Support: The Effects of Explainable AI and Cognitive Forcing Functions

Oliver Henderson

School of Business, University of Melbourne, Russell, Australian

*Corresponding Author

ARTICLE INFO

Keywords:

Artificial Intelligence;
Human-AI Collaboration;
Explainable AI; Trust
Calibration; Cognitive Load;
Healthcare Decision-
Making

Received: Feb, 23, 2026

Accepted: Apr, 13, 2026

Published: Jun, 25, 2026

ABSTRACT

Reliance on AI systems for decision support is expanding into domains where mistakes carry real consequences, which makes it important that users can evaluate AI recommendations alongside their own judgement instead of accepting them automatically. Prior work has found that human-AI teams sometimes underperform AI alone, a pattern usually attributed to automation bias: individuals may still rely on AI outputs despite the possibility that those outputs are inaccurate. Many current approaches attempt to address this issue through text-based explainable AI (XAI) methods that clarify the logic behind AI outputs. Everyday decision-making leans heavily on fast, intuitive, bias-prone processing, users may skim past these explanations or fail to engage with them in any depth. One potential solution is the cognitive forcing function (CFF), a design approach that encourages users to pause automatic responses and engage in more careful evaluation. This study investigates how six decision-support mechanisms influence engagement, trust, and human-AI collaborative performance within a diabetes meal-planning context. In a controlled study with 108 participants (18 per condition), we tested two explanation types, text and visual, alongside four CFFs: AI confidence levels, human feedback, AI-generated questions, and performance visualization. AI confidence levels, text explanations, and performance visualization each significantly improved decision accuracy, and reasoning cues such as text and confidence levels also improved trust. Human feedback and AI-driven questions prompted deeper reflection but appear to have come at a cost to performance, plausibly because of the added cognitive effort and scrutiny involved, which in turn dampened trust. Visual explanations, despite their simplicity, did not meaningfully move trust. These results argue for a contextual, balanced pairing of CFF and XAI design that accounts for interactivity, decision frequency, and task complexity.

1. INTRODUCTION

Effective chronic disease management, particularly diabetes care, depends on making informed and continuous decisions. This burden has traditionally been borne by patients and providers, and has required continuous attention, which can be challenging day after day. AI systems can now help alleviate that burden, and generative AI in particular has been found to be useful for diabetic

people in meal planning to maintain a healthy blood sugar level (Ellahham, 2020). How well this collaboration works, though, hinges on whether users can strike the balance between trusting the AI's suggestions and applying their own judgment. The problem of helping users to decide when to trust the system and when to challenge the system's output is not uniquely found in

healthcare, but is a problem that interface designers in high-stakes domains, such as disaster response, have to address (Al-Sadi et al., 2022). User-centred design has long prioritised the user's decision process in the system, and this is a key element of solving it well, as demonstrated in various application areas (Al-Sa'di, 2022).

2. LITERATURE REVIEW

Trust represents a fundamental element in human-AI interaction, and particularly important in healthcare settings (Zuchowski et al., 2024). Over-trust or under-trust a system: Over-trust is when users take the AI's recommendations without critical evaluation, including incorrect ones, which can result in harmful recommendations (Zhang et al., 2020); under-trust is when users reject the AI's suggestions, even when they are correct, thereby missing out on health benefits that the AI could have offered. The balance is crucial for the safe and effective application of AI (Al-Sa'di & Al-Samarraie, 2022) and in healthcare, it is particularly vital because the consequences of a wrong decision can be immediate.

While several recent studies have studied trust-calibration strategies (Ma et al., 2023; Zhang et al., 2020), the issues of overtrust and undertrust persist, and human-AI teams still fail to outperform the human or the AI alone consistently (Bućinca et al., 2021). Despite the increasing popularity and focus on explainable AI (XAI) (Ali et al., 2023), explanations alone have not proven to be an effective solution to trust issues; the more complex the task or explanation, the more likely the overreliance is to persist. (Logg et al., 2019). Dual-process theory provides a useful framework for this, as it distinguishes between fast, intuitive thinking (System 1) and slower, deliberative thinking (System 2) (Kahneman, 2011). Users appear to weigh the cognitive cost of scrutinizing an AI suggestion (System 2) against the perceived payoff, and will default to uncritical acceptance (System 1) once that cost feels too high (Vasconcelos et al., 2023). The opposite failure mode also exists: excessive reliance on intuition, or conversely overly critical System 2 analysis, can each cause users to discount genuinely useful AI input. Cognitive forcing functions (CFFs) have recently been explored as a way to interrupt automatic reasoning at the point of decision and nudge users toward System 2 processing (Lambe et

al., 2016).

To probe this shift from System 1 to System 2, and to see whether it can correct over- or under-trust, we built six decision-support mechanisms, including a baseline text explanation. These fall into two broad families. XAI mechanisms aim to make the AI more transparent and split by presentation modality into textual, visual, and multimodal forms. CFFs aim to reduce cognitive bias and encourage deliberate processing, and we grouped ours by the kind of cognitive effort each demands:

- Attention-directing CFFs guide the user's focus and prompt critical thinking, such as AI-generated questions.
- Complexity-reduction CFFs lower cognitive load by making AI output easier to interpret, such as confidence levels that communicate model uncertainty, or performance visualizations that show task accuracy over time.
- Decision-constraint CFFs let users actively intervene in the decision process, such as human feedback mechanisms that let users validate AI outputs.

We tested these mechanisms in a human-subjects study with 108 participants, 18 per condition, using a diabetes meal-planning task. Each participant moved through three phases across 20 questions: an unassisted baseline decision (P1), a decision made with AI assistance but no explanation (P2), and a decision made with AI assistance plus one of the six support mechanisms (P3). We combined quantitative and qualitative measures, collecting both objective accuracy data and subjective ratings of satisfaction, system complexity, reliability, trust, and perceived accuracy through pre- and post-task questionnaires, alongside open-ended responses covering confidence, cognitive load, usefulness, and trust dynamics. Statistical analyses controlled for individual differences such as prior AI familiarity and educational background.

This work makes three contributions. First, it offers empirical evidence on how different decision-support mechanisms shape trust, engagement, and decision-making. Second, it adds to our understanding of trust dynamics in human-AI interaction by examining how explanation type and CFF design each influence trust calibration, extending a line of work on how AI-mediated tools such as ChatGPT reshape user experience and

expectations more broadly (Al-Sa'di & Miller, 2023). Third, it lays out a controlled experimental method for evaluating decision-support mechanisms in isolation, one that should transfer to domains beyond healthcare, including education and vocational training, where similar questions of engagement and retention already motivate interface research (Yamjal & Ahmed, 2022).

3. METHODS

The study used a custom web interface simulating interaction with a generative AI system tasked with meal-planning support for diabetes management. Each participant was assigned one experimental condition and moved through three phases.

3.1. Participants

In order to assess data 108 participants were enlisted ($N = 108$) through Prolific, randomly assigned across six conditions, 18 participants each. A power analysis (Cohen, 2013) confirmed that 18 participants per condition was sufficient to detect an effect size of $r = 0.60$ at 80% power and $\alpha = 0.05$, for both within- and between-group comparisons. Participants were drawn from a pool varied in age, gender, educational background, AI familiarity, and diabetes knowledge; all listed English as a primary language and held a Prolific approval rating of 99-100%. Compensation was \$10.15 per hour, with a \$1 bonus for decision accuracy above 75%. Participants were screened to exclude anyone with medical expertise in diabetes management, so as to approximate general users engaging with AI in a healthcare context. Informed consent was obtained from all participants under institutional ethical guidelines. Table 1 summarizes participant demographics.

Table 1. Participant demographic summary ($N = 108$).

Variable	Summary
Sample size	$N = 108$ (18 per condition)
Diabetes familiarity (1–7 scale)	$M = 4.09$, $Mdn = 4.0$, $SD = 1.86$
Age (years)	$M = 36.99$, $Mdn = 35.5$, $SD = 12.01$, range 16–67
Education	High school 53.78%; college degree 38.89%;

Variable	Summary
	master's 4.63%; PhD 3.70%
Gender	Woman 51.85%; man 45.37%; nonbinary 1.85%; other 0.93%

3.2. Task and procedure

Participants completed a simulated meal-planning task designed to mirror real diabetes-management decisions, structured in three phases:

- Phase 1 (no AI assistance).** Participants chose between two meal options for a hypothetical Type 2 diabetic person, establishing a baseline measure of unaided accuracy.
- Phase 2 (AI assistance, no explanation).** The same meal pairs were shown again, now with an AI recommendation but no supporting rationale, isolating the effect of AI assistance itself.
- Phase 3 (condition-specific mechanism).** Participants saw one of the six explanation mechanisms and could revise or affirm their earlier choice, isolating the effect of that specific mechanism.

3.3. Experimental conditions

- **Textual explanations (C1).** A written rationale for the AI's choice (e.g., "Meal 2 contains fewer carbohydrates, which is critical for managing blood sugar levels"), testing whether a logical, written account increases transparency and critical thinking.
- **Visual explanations (C2).** Key nutritional components (carbohydrates, protein) highlighted graphically alongside a concise nutritional summary, testing whether visual framing reduces cognitive load while still building trust.
- **AI confidence levels (C3).** Confidence levels shown for both AI and user; the AI's confidence reflects model uncertainty, and the user's is estimated from their task history (see the Human confidence-level prediction subsection below). This tests whether a shared, quantifiable certainty scale helps calibrate trust.
- **Human feedback (C4).** Participants directly input confidence levels for both the

AI and themselves, testing whether active reflection on both parties' certainty improves engagement and calibration.

- **AI-driven questions (C5).** Three AI-generated questions prompt reflection on the decision and the AI's recommendation, testing whether guided reflection deepens engagement and calibrates trust.
- **Performance visualization (C6).** A chart comparing the AI's and the participant's past accuracy, testing whether historical performance data supports better trust calibration.

3.4. Measures

Participant responses were recorded at each phase, capturing:

- **Decision accuracy**, scored against a ground truth derived from American Diabetes Association guidelines (see below).
- **Pre-post questionnaire measures:** satisfaction, system complexity, reliability, trust, and perceived system accuracy, all on 7-point scales.
- **Post-task measures:** confidence and understanding, consistency and cognitive load, ease and usefulness, trust and comfort, and trust dynamics, each captured through condition-general and condition-specific items, with open-text responses collected alongside the Likert items.

3.4.1. Explanation Information Load (EIL)

The concept of EIL is based on information theory (Shannon, 1948) and on previous research that applied logarithmic complexity metrics to the evaluation of interfaces (Sarsam et al., 2021). In each condition, the EIL includes a mechanism-specific information term and a common image-information term (which is not used to compare across conditions to isolate mechanism-specific load):

- **Textual explanations:** $EIL = \log_2(\text{text length}) + \text{image information}$.

Visual explanations: $EIL = (n \times [\log_2(n) + I(n)]) + I(n)$ where n is the number of segments, $I(n)$ is the length of the label, and $I(n)$ is the information in the image.

- **AI confidence levels:** $EIL = \log_2(\text{shown number of confidence values}) + \text{image information}$.
- **Human feedback:** $EIL = \log_2(\text{number of inputs} \times \text{number of confidence values}) + \text{image}$

information.

AI-generated questions: $EIL = \log_2(\text{question length}) + \text{image information}$.

- **Performance visualization:** $EIL = \log_2(\text{number of chart components}) + \text{image information}$.

3.4.2. Human confidence-level prediction

Under AI confidence-level condition, a generative model (GPT-4o-2024-05-13) estimated each participant's confidence in making the right decision based on their previous 20 decisions in Phase-2 (the nutritional profile of both meal options and the ground-truth answer, along with the nutritional profile of the pair of meals in Phase-2) and the nutritional profile of the pair of meals in Phase-2. This is a model that was asked to return the probability of this happening as a 0-100% estimate, given the participant's past decision-making history and the ground truth of the current task.

3.4.3. Ground-truth calculation

The ground truth for each meal pair was obtained from the American Diabetes Association nutritional guidelines (Colberg et al., 2016; Evert et al., 2019). The percentages of total calories for protein, fat, and carbohydrates were computed for each meal, and balance criteria were applied: a meal was considered "balanced" (indicator $B = 1$) if a protein percentage was between 20% and 32% and fat percentage was between 20% and 35%, and "unbalanced" ($B = 0$) otherwise. In the unbalanced meal condition, the balanced meal was the ground truth meal, and in the equally balanced meal condition, the meal with the lower carbs percentage was chosen, because of direct effect of carbs on post meal glucose.

3.4.4. Statistical analysis

All analyses were conducted in Python 3.10.0 and SciPy, and significance was set at $p < 0.05$ for all analyses; data and code will be made available upon publication. AI suggestions (H1) were tested using McNemar's test (McNemar, 1947) which is suitable for paired nominal data. Within-condition comparisons across the three phases (H2, H3) used the Wilcoxon signed-rank test (Wilcoxon, 1992), chosen after Shapiro-Wilk tests (Shapiro & Wilk, 1965) showed non-normal or borderline-normal distributions for most conditions (C1: $p = 0.004$; C2: $p = 0.001$; C3: $p < 0.0001$; C6: $p = 0.045$; C4: $p = 0.051$; C5: $p = 0.061$). Comparisons were made between conditions (H3) using a non-parametric

analogue of one-way ANOVA, the Kruskal-Wallis test (Kruskal & Wallis, 1952) for independent groups that were not normally distributed. Such a paired comparison approach, where the discrete responses of users evolve in parallel over two measurement time points, has been helpful in identifying correlated behavioural events in other large-scale user studies (Sarsam et al., 2021).

4. DATA ANALYSIS AND RESULTS

The analysis was based on three hypotheses:

H1. Even if the AI gives little reason, users are inclined to believe the suggestions, leading to overreliance.

H2. Explanation mechanisms can help users question the AI's suggestions and provide more critical evaluation.

H3. Explanation mechanisms that have a balanced Explanation Information Load (EIL) – that is, they convey interpretable explanations without overwhelming the user – best support effective collaboration by improving decision-making and calibrating trust.

Users over-rely on AI suggestions, even when incorrect (H1)

The McNemar's test was employed to compare the decisions made by users before and after viewing the AI suggestions, both when the AI was correct and when it was incorrect (Table 2). A significant number of users followed the AI's suggestions even when they were wrong ($p < 0.001$), confirming previous findings of automation bias (Goddard et al., 2012). If the suggestions from AI were accurate, a similarly high result ($p < 0.001$) revealed that users were highly motivated to accept the AI suggestions.

Overall, these findings suggest that users are more likely to adopt AI recommendations without deep critical analysis, a behavior that aligns with relying on the AI as an authoritative source when performing a complex task, thus supporting H1. The trust paid off when the AI was correct, but when it was incorrect, there was little to no correction, indicating a genuine risk of overreliance, and with obvious implications in high-stakes environments such as healthcare.

Table 2 (see below) reports the McNemar test results.

Table 2. McNemar test results for user decisions relative to AI suggestions, split by whether the AI's suggestion was itself correct or incorrect.

Condition	Correct → Correct	Correct → Incorrect	Incorrect → Correct	Incorrect → Incorrect	<i>p</i>
AI suggestion incorrect	109 (22.4%)	94 (19.3%)	19 (3.9%)	264 (54.3%)	< 0.001
AI suggestion correct	1,131 (67.6%)	62 (3.7%)	200 (11.9%)	281 (16.8%)	< 0.001

Explanation mechanisms can help people engage (H2)

We hypothesized that decision-support mechanisms would increase engagement (H2) between Phase 1 and Phase 2, and Phase 1 and Phase 3, using the Wilcoxon signed-rank test (Wilcoxon, 1992), while comparing pre- and post-questionnaire scores on satisfaction, system complexity, reliability, trust, and perceived accuracy (see Methods).

The baseline condition (textual explanations) was also found to significantly improve engagement ($p = 0.049$) and perceived reliability ($p = 0.048$), suggesting that providing explanations in text format increases the likelihood of users understanding and following the AI's reasoning and developing some trust in the system, while also increasing the complexity of the system ($p = 0.008$). There was no significant engagement gain ($p = 0.776$) for the visual explanations (C2), which indicates that a simple visual representation of the AI's reasoning without further explanation might not inspire users to engage with the explanation in greater detail; users may have taken the output at face value. AI confidence levels (C3) resulted in a significant increase in engagement ($p = 0.001$): It seems there is a strong drive to compare the AI's confidence and one's own, making the user more than just a recipient of the AI. Reliability and trust did not change significantly. Human feedback (C4) also significantly increased engagement ($p = 0.006$) likely through introspection, but without a corresponding increase in reliability or trust, suggesting that engagement and trust can be independent. For the AI-driven questions (C5),

there was a close to significant effect ($p = 0.058$): some users were nudged toward more reflective thinking by the AI questions, requiring a shift from System 1 to System 2, but the effect was not quite significant, indicating potential for improvement. Performance visualization (C6) also contributed to a significant increase in engagement ($p = 0.009$); this may be due to the fact that users have concrete evidence with which to evaluate the performance of AI and humans over a period of time.

Explanation information load and interpretable reasoning jointly shape collaboration (H3)

We computed an Explanation Information Load (EIL) score for each condition to capture how much a mechanism demands of users cognitively (see Methods for the full derivation). Table 3 summarizes the results.

Table 3. Explanation Information Load (EIL) across the six conditions.

Condition	Mechanism	EIL
C1	Textual explanations	1.520
C2	Visual explanations	1.749
C3	AI confidence levels	0.602
C4	Human feedback	1.204
C5	AI-driven questions	1.965
C6	Performance visualization	0.602

Consistent with H2, we expected a moderate EIL to support better collaboration outcomes (H3), and the pattern across conditions bears this out, if unevenly. Visual explanations (C2) and AI confidence levels (C3) both produced significant gains in user accuracy ($p = 0.018$ and $p = 0.004$, respectively). C2's moderate EIL (1.749) appears to give users enough detail to engage critically without becoming overwhelmed, while C3's low EIL (0.602) lets users read AI certainty at a glance, functioning as an efficient cognitive shortcut. Both mechanisms deliver reasoning users can actually interpret, pairing a manageable load with substance that supports informed decisions.

Mechanisms lacking clear interpretive content told a different story. Performance visualization (C6), also at an EIL of 0.602, showed only borderline significance for accuracy ($p = 0.092$): users were exposed to historical performance data but had to draw their own conclusions without explicit interpretive scaffolding, consistent with cognitive-psychology findings that dense information without interpretive support can produce strain rather than insight (Arnold et al., 2023). Visual explanations considered on their own terms ($EIL = 1.520$ by an alternate accounting), human feedback (C4, $EIL = 1.204$), and AI-driven questions (C5, $EIL = 1.965$) showed no significant accuracy gains ($p = 0.673$, $p = 0.674$, and $p = 0.302$, respectively). Textual explanations (C1) offer intuitive but limited content, possibly too thin to support well-calibrated decisions; human feedback demands self-assessment without giving users AI-side reasoning cues to anchor it; and AI-driven questions, with the highest EIL of any condition, likely tip into information overload given how much reflection they require.

The overall picture supports H3: collaboration benefits most from mechanisms that pair a manageable EIL with genuinely interpretable reasoning, while mechanisms that either lack interpretive content or demand heavy cognitive investment tend to underperform, regardless of intent.

Condition-level effects on satisfaction, complexity, reliability, trust, and perceived accuracy
Detailed within-condition statistics (before versus after interacting with each mechanism) appear in Table 4; the pattern across conditions is summarized below.

Table 4. Within-condition significance (p -values) for pre- versus post-task ratings of satisfaction, system complexity, reliability, trust, and perceived accuracy. n.s. = not significant ($p > 0.05$).

Condition	Satisfaction p	Complexity p	Reliability p	Trust p	Perceived accuracy p
C1 — Textual explanations	n.s.	0.008	0.048	0.368 (n.s.)	n.s.
C2 —	0.050	n.s.	0.005	0.4	margi

Condition	Satisfaction p	Complexity p	Reliability p	Trust p	Perceived accuracy p
Visual explanations				21 (n.s.)	nal
C3 — AI confidence levels	n.s.	n.s.	n.s.	n.s.	0.039
C4 — Human feedback	n.s.	n.s.	n.s.	0.417 (n.s.)	n.s.
C5 — AI-driven questions	n.s.	0.008	n.s.	0.032	n.s.
C6 — Performance visualization	n.s.	n.s.	n.s.	0.099 (n.s.)	n.s.

Textual explanations (C1) significantly improved reliability ($p = 0.048$) but also significantly increased system complexity ($p = 0.008$), and trust dipped slightly, though not significantly ($p = 0.368$). Users appreciated the added clarity — one noted that "the reasoning for its decisions helped me pick which meals were better" — but the extra processing required to read and weigh the explanation appears to have offset some of that benefit, an illustration of a broader design tension: explanations can raise perceived reliability while simultaneously raising cognitive load enough to blunt trust.

Visual explanations (C2) significantly improved satisfaction ($p = 0.050$) and reliability ($p = 0.005$) with no noticeable increase in complexity, but did not increase trust ($p = 0.421$) and only slightly increased perceived accuracy. The difference in reliability and trust indicates that a system may be reliable but not necessarily trusted to make decisions; trust is more than just a demonstration

of reliability; it requires some transparency about how the reliability is achieved.

AI confidence levels (C3) resulted in an important improvement in perceived accuracy ($p = 0.039$) and a non-significant improvement in reliability, with little change in complexity or satisfaction. The confidence levels seemed to be an effective, low-cost indicator that enabled users to adjust their trust, with a slight (but not statistically significant) drop in trust when the AI was not 100% confident, suggesting that seeing it not always confident makes users more critical of its suggestions.

Human feedback (C4) satisfaction, complexity, reliability, and perceived accuracy remained essentially the same with a slight, non-significant decrease in trust ($p = 0.417$). Some participants noted that the reflection prompt was too late to be an important factor in their decision-making: "I had already decided by the third phase, so it was not that much of a factor for me," and "I felt like the AI was not very reliable. Self-confidence does not necessarily lead to trust that is well calibrated.

AI-driven questions (C5) produced a significant, counterintuitive drop in perceived system complexity ($p = 0.008$), suggesting the reflective prompts may have focused users' attention rather than burdening it, but also a significant drop in trust ($p = 0.032$). Participants who noticed the AI's errors during reflection said as much directly: "It felt a bit unreliable when it kept saying that there was pizza on the plate despite there very clearly being no pizza on the plate," and "When I saw the AI incorrectly thought one plate had pizza, it made me want to pick the other one instead." Prompting users to think critically, in short, can expose the AI's limitations along with its reasoning, and users respond to both.

Performance visualization (C6) satisfaction and complexity were essentially unchanged and trust was non-significantly reduced ($p = 0.099$) by performance visualization (C6). While some participants found it helpful to use it in a positive way, for example, "I used it to reflect" and "the bar chart helped give me a visual 'faith' in the AI", it did not appear to change the perceived accuracy of the AI, and users still relied primarily on the AI's immediate output when determining its correctness, rather than on its track record.

Cross-condition analysis: confidence, cognitive load, usefulness, trust, and trust dynamics.

Table 5. Cross-condition comparison across five

post-task measures. *Confidence-and-understanding p-values ranged from 0.002 to 0.008 depending on the pairwise comparison; see text for condition-specific values. n.s. = not significant ($p > 0.05$).*

Measure	Overall p	Highest-scoring condition	Lowest-scoring condition
Confidence and understanding	0.007 – 0.008	Textual explanations ($M = 6.000$)	AI-driven questions
Consistency and cognitive load	0.263 (n.s.)	AI confidence levels ($M = 4.694$)	AI-driven questions ($M = 3.944$)
Ease and usefulness	0.029	Textual explanations ($M = 5.444$)	AI-driven questions ($M = 3.852$)
Trust and comfort	0.012	Textual explanations	Human feedback / AI-driven questions
Trust dynamics (pre-post shift)	0.602 (n.s.)	AI confidence levels ($M = 4.500$)	Human feedback ($M = 3.444$)

5. DISCUSSION

The comparison between all six conditions shows that there are significant differences in confidence ($p = 0.007$), understanding ($p = 0.007$ for textual explanations and $p = 0.002$ for human feedback, $p = 0.003$ for AI-driven questions, $p = 0.008$ for performance visualization). The textual explanations received the highest level of confidence ($M = 6.000$, $SD = 0.553$): even a simple explanation, even if not very transparent, can inspire confidence, plausibly because of the ease with which it can be used. Those that required more involvement, such as human feedback and AI-generated questions, performed much worse in accordance with the notion that increased

cognitive load leads to greater skepticism of AI-generated output.

There were no significant differences across the conditions in terms of consistency ($p = 0.263$) or cognitive load ($p = 0.723$), although AI confidence levels were higher than AI-driven questions when users perceived the system as more consistent ($M = 4.694$ vs. 3.944), perhaps because the transparent uncertainty estimates helped users read the system as more stable, while the conditions where users were required to engage more heavily added load without a corresponding increase in perceived consistency.

The mean score for ease and usefulness was significantly higher for textual explanations than for human feedback ($p = 0.025$) and AI-driven questions ($p = 0.009$), with textual explanations scoring the highest ($M = 5.444$, $SD = 0.882$), and AI-driven questions scoring the lowest ($M = 3.852$, $SD = 1.799$), possibly due to the extra deliberative step that AI-driven questions require before a decision is made. Here, as with the AI-driven questions, performance visualization outperformed AI-driven questions ($p = 0.017$), indicating that it can be used to guide decisions without causing as much friction.

Finally, trust and comfort varied greatly ($p = 0.012$): again, textual explanations were highest, outperforming human feedback ($p = 0.004$) and AI-driven questions ($p = 0.024$), showing that reflective mechanisms, which foster critical thinking, can also highlight uncertainty, thereby reducing comfort. On this measure, performance visualization was more likely to be trusted than were human feedback ($p = 0.036$) or AI-driven questions ($p = 0.025$).

The differences in trust dynamics across conditions were not statistically significant ($p = 0.602$), but the means do tell a story: AI confidence levels showed the largest change ($M = 4.500$, $SD = 1.607$), indicating that users actively adjust their level of trust when they view the AI's level of confidence, whereas human feedback had the smallest change ($M = 3.444$, $SD = 1.674$), suggesting that when users have no visibility of the AI's confidence, they rely more on their own judgment. In between were the questions and performance visualization powered by AI, with scores of $M = 4.056$ and 3.944 respectively, which means there was no significant trust swing either direction.

Knowledge about and accuracy in decision making

for diabetes management

Participants were split into those who had little or no experience with diabetes management (self-rated ≤ 4) and those who had more experience with diabetes management (> 4) and no significant differences in accuracy were identified between these groups by two-sample t-test (all $p > 0.05$). For unfamiliar users ($n = 8$), mean accuracy was 69.4% and for familiar users ($p = 0.398$), it was 72.5%. For AI confidence levels, unfamiliar users ($n = 11$) had a mean of 71.8%, while familiar users had a mean of 77.9% ($p = 0.081$), which was trending toward significance. It is important to note that prior experience in the domain does not systematically shift the accuracy of the decisions made within a condition, either because the domain itself is sufficiently easy to access for the general public or because the support mechanisms are sufficiently strong to raise less-experienced users to the level of more experienced users who may have relied on their background knowledge.

- Evolution of trust patterns under conditions

Self-reported trust scores showed a general pattern, with those who began with higher levels of trust decreasing at the end and those who began with moderate levels increasing at the end in nearly every condition. This was observed in written explanations and even more distinctly in visual explanations, where the level of trust decreased for higher-trust participants following the task. The scores for AI confidence levels were generally stable, with most scores falling near the pre-post diagonal line. The results for human feedback and AI-driven questions were both lower than the baseline, while the results for performance visualization mirrored the overall trend, with initially high-trust participants reporting lower trust after the task.

There was no condition that resulted in a net increase in trust. This lack is worthy of sitting with: it means that even if a mechanism enhanced decision-making, it did not necessarily lead to a greater level of trust in the AI itself. Users might have used these explanations more as learning tools than as instruments for calibrating trust in the system, relying more on them to improve their own mental model and reasoning process. That is, while decision accuracy can increase, trust may not necessarily follow, and people who are initially more trusting might have had higher expectations

that were not satisfied by the system, or by the increased cognitive effort of the specific condition. This decoupling of performance from trust is similar to what has been found in other digital-exchange settings, where trust is developed over time through emotionally charged and repeated interactions, but not through a single, open, and obvious disclosure (Al-Sa'Di et al., 2025b).

Discussion

- The cognitive forcing functions are compared with XAI methods.

We identified six mechanisms and categorized them into two families: XAI methods (textual explanations and visual explanations) and CFFs (AI confidence levels, human feedback, AI questions, and performance visualization).

XAI methods. Detailed explanations using text proved to be a significant factor in boosting both performance and reliability, highlighting the importance of providing users with clear explanations when dealing with AI-generated text. In addition to the above reason, users reported that the system expanded their thinking: one user reported that the system helped him think of ideas and decisions that he wouldn't have had on his own, and another user reported that the system gave him more perspective of the picture. As in previous studies (Bansal et al., 2021), there was a trade-off for that benefit: the increased transparency also increased the cognitive load, and participants struggled to use the system even though they appreciated the transparency. This might be exacerbated by time pressure, as users may not be reading the explanations carefully when they are under time pressure.

However, there was no significant improvement in performance when the visual explanations were added. There is mixed literature here: some studies show that if it is well designed, visual explanations increase satisfaction and trust (Yang et al., 2020), while others indicate that visual explanations are subject to misinterpretation, particularly by non-expert users, and that they are not as informative as detailed text. Our findings suggest the latter: satisfaction and complexity remained unchanged, and trust did not shift, although the underlying task was a visual task. One of the participants explicitly described the resulting hesitation: "Sometimes I would be torn whether I should change when I opted to go in a different direction than AI. It appears that visual cues are not sufficient to

reliably calibrate trust.

The XAI conditions together represent a minimalist example of how transparency and cognitive load are related, and how these mechanisms can actually aid, provided that they are optimally designed to not overload the user. The reflective, user-centred approach to designing interfaces that practitioners use to solve other interface problems can be useful in solving that trade-off (Al-Sa'di & Alsamarraie, 2023).

- Cognitive forcing functions

Previous studies have suggested that AI confidence should be made explicit, as it aids in assessing when and how to trust the system (Ma et al., 2023; Bućinca et al., 2021), and our findings support this: AI confidence levels significantly affected performance, suggesting that explicit uncertainty provides users with something they can critically engage with. This increase in perceived accuracy did not significantly increase trust, but the increase in perceived accuracy indicates that trust levels are a useful cognitive scaffold that can be used to make more informed judgments without overloading the system. Nevertheless, as there was some lingering doubt about the trustworthiness of the confidence level in the signal, participants expressed some residual uncertainty about how to use the signal — "I wasn't sure if I should trust the confidence level," said one, and another added, "...but most of the time I found it hard to believe the AI suggestions" (Zhang et al., 2020), reflecting a broader concern in the literature that confidence scores alone may not be sufficient for deep trust calibration, and that richer explanation may be required for truly complex decisions.

Human feedback, especially the self-reporting variety, produced mixed results for trust calibration, in line with prior findings that self-report mechanisms can raise cognitive load and, when users are asked to reflect on their own confidence, introduce skepticism. Our data fit this pattern: no significant gains in performance, satisfaction, or trust, and participant comments pointed to two distinct failure modes, disengagement ("I had already made my mind up by the third phase") and eroded confidence in the system ("I felt like the AI was unreliable"), a reaction consistent with broader findings that poorly calibrated feedback loops are a common source of user frustration in interactive systems (Al-Sa'di & Litayem, 2023). Structured self-

assessment can work well when it is built around a validated instrument (Al-Sa'di et al., 2023a), but the ad hoc confidence input asked of participants here offered no such scaffolding. Feedback mechanisms, in short, can promote reflection, but only if designed with real care about how much complexity and skepticism they introduce along the way; giving managers and system designers concrete strategies for building that care into a mechanism's rollout is itself an active design challenge (Al-Sa'di et al., 2023b).

Reflective questions have been studied as a way to pull users from fast, heuristic System 1 judgments toward slower, analytical System 2 processing (Panjabi et al., 2024; Kahneman, 2011), and CFF research more broadly shows that forcing users to slow down can reduce overreliance (Bućinca et al., 2021). Our results are consistent with the engagement half of that story but not the trust half: AI-driven questions raised engagement while measurably lowering trust and reliability, likely because reflection made the AI's mistakes more visible, as the pizza-related quotes above illustrate directly. System complexity dropped significantly at the same time, suggesting the questions helped focus attention even as they cost the system some credibility. Designing reflective prompts that build critical thinking without undermining trust remains an open challenge.

Performance visualization has been studied as a way to help users build an accurate mental model of AI behavior and thereby justify their own decisions more confidently. In our data it showed borderline significance for performance, suggesting it nudges users toward more deliberate engagement without fully getting there; participant comments ("I used it to reflect," "the bar chart helped give me a visual 'faith' in the AI") support this reading. The mechanism functions as a cognitive aid, but the modest effect size suggests it would benefit from more actionable insight layered on top, consistent with prior work emphasizing that visual aids need careful design to avoid either overwhelming users or falling short of real impact (Al-Sa'di & Al-Samarraie, 2022).

- Under-reliance and over-reliance on AI

We also examined how reliance shifted between Phase 2 (AI support without a designed mechanism) and Phase 3 (AI support with one), measuring the degree to which users followed correct versus incorrect AI suggestions in each

phase.

Textual explanations produced only a minor, non-significant shift in reliance calibration, suggesting detailed reasoning helps without dramatically altering how much users lean on the AI (Yang et al., 2020). Visual explanations showed a small, non-significant increase, hinting that users may lean on the visual aid without necessarily scrutinizing whether the AI's underlying suggestion is correct. AI confidence levels, by contrast, produced a significant improvement in reliance calibration: transparent confidence appears to help users align more closely with the AI when it is right and pull back when it is wrong, underscoring the value of quantifiable certainty for minimizing both under- and over-reliance (Cao et al., 2024; Ma et al., 2023). Human feedback left reliance calibration largely unchanged, consistent with users struggling to map their own confidence onto the AI's accuracy in any consistent way. AI-driven questions produced a significant improvement in reliance calibration, with users more likely to follow correct suggestions and reject incorrect ones after reflecting, though this benefit needs to be weighed against the added cognitive load and risk of overthinking that comes with it. Performance visualization showed a small, non-significant decrease in reliance calibration, suggesting the comparison data alone, without more direct interpretive support, was not enough to stop users from following the AI even when it erred.

Across these patterns, mechanisms built around transparency and interpretability (textual explanations, AI confidence levels) and around reflective engagement (AI-driven questions) appear most effective at curbing overreliance on incorrect AI output. Visual explanations and performance visualization, by comparison, seem to offer too little interpretive guidance on their own, leaving users more exposed to the AI's errors. The upshot is that reliance calibration benefits from mechanisms that go beyond surface-level transparency to offer either clear, actionable reasoning or a genuine opportunity for reflection. Balancing cognitive load and reliance calibration Our reliance-dynamics analysis confirms the EIL-based account from the Results section, but from a different perspective. Again, AI confidence levels are useful, allowing users to tune their reliance by adjusting their reliance to be closer to the correct AI output, while being less dependent on the

incorrect suggestions, consistent with previous research on the importance of quantifiable uncertainty (Logg et al., 2019). Visual explanations and performance visualization, on the other hand, appear to provide insufficient support to allow users to exit from incorrect AI suggestions, a general trend of overreliance on automation where transparency is low or cognitive load is high (Goddard et al., 2012; Vasconcelos et al., 2023; Logg et al., 2019). With the intention of stimulating reflection and active engagement, human feedback does not necessarily lead to better calibration or accuracy, in line with previous studies showing that reflective mechanisms can underdeliver, especially for comparative-reasoning tasks of high cognitive demand (Bućinca et al., 2021). The most interesting paradox is that of reduced reliance calibration along with increased accuracy, which indicates a cognitive load effect on reflection that forces users to be more critical during their experience, while simultaneously introducing some doubt about the AI's reasoning (Bućinca et al., 2021). The heart of our findings is this central design puzzle: reflective engagement is a facilitator for deeper thinking, but only when the engagement is calibrated carefully enough to not tip into over-scrutiny (Al-Sa'di & Miller, 2026).

Limitations and Future Work

These results were limited in their generalizability in several ways. First, the study focuses on a specific and high-stakes domain (diabetes meal-planning), and other domains (finance, law, education) have different cognitive demands and stakes, so future work should test the mechanisms in other domains to see how well the patterns transfer. Second, we measured reliance calibration in just one, relatively brief, controlled session; though this is a typical strategy in the field (Bućinca et al., 2021; Ma et al., 2023), it does not reflect the way reliance changes over longer, more realistic sessions, as those that occur in adaptive online-course platforms are designed for sustained user engagement (Al-Sa'di & Khrypko, 2024). Real-world use likely involves users shifting between over- and under-reliance as they build familiarity with a system over time, and as the system's demonstrated value to them changes; the trust-building process observed in repeated peer-to-peer digital exchanges offers one relevant analogue (Al-Sa'di et al., 2025a). Third, our sample skewed toward current AI-familiar adult users, and how

the coming generation of digitally native users engages with these same mechanisms remains an open question worth tracking (Ahmed & Ahmad, 2023). Longitudinal studies would help clarify how reliance actually evolves, and whether support mechanisms need to adapt as users become more experienced. Finally, our interface was implemented as a controlled web prototype; deploying comparable mechanisms on resource-constrained or embedded platforms raises its own robustness and security considerations that a production system would need to address (Litayem & Al-Sa'di, 2023).

6. CONCLUSION

In this study, six decision-support mechanisms were studied for their influence on trust, engagement, and decision-making in a high-stakes context, diabetes meal planning, each of which aimed to influence either fast, intuitive (System 1) or slower, deliberative (System 2) processing. Mechanisms based on transparent and interpretable reasoning, such as textual explanations and AI confidence levels, enhanced decision accuracy by enabling users to better match their own decisions with the correct AI suggestions. Human feedback and AI-generated questions were among the CFF mechanisms that fostered deeper engagement, but did not come without a price: additional cognitive load that measurably impaired performance. No single mechanism was able to enable both System 1 and System 2 thinking simultaneously, suggesting that there is a true trade-off in interface design between intuitiveness and depth of thought. Visual explanations and visualization of performance were low effort supports but did little to enhance critical evaluation, and did less to improve trust and reliance; mechanisms based on reflection were more deeply analyzed but had the potential to overload users as the cognitive demand increased. The results suggest adaptation in decision support systems, taking into consideration task complexity, user expertise and frequency of decision making. In our data, mechanisms that had manageable EIL and had an acceptable level of interpretability of their output seemed to strike the best balance between engagement and performance when collaborating with humans. Future research should focus on dynamically adapting interfaces to the complexity of explanations based on user behavior and task

difficulty, as well as longitudinal research to investigate how users' dependency on AI evolves over time and across settings. Further research in this area could pave the way for AI systems to go beyond mere accuracy in decision-making to truly enable safe and sustainable human-AI collaboration in critical environments.

REFERENCES

- Ahmed, A., & Ahmad, E. (2023). The future of New Zealand tertiary education: Generation alpha. *New Zealand Quality. Research and Innovation*.
- Al-Sa'di, A. (2022). User centred design methods in animal centred design: A systematic review. In 2022 International Conference on Theoretical and Applied Computer Science and Engineering (ICTASCE).
- Al-Sa'di, A., Ahmad, E., Yamjal, P., & McPhee, C. C. A. (2023b). Strategies to enhance educators' soft skills: The role of managers. In 2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) (pp. 1-6). IEEE.
- Al-Sa'di, A., & Al-Samarraie, H. (2022). A Delphi evaluation of user interface design guidelines: The case of Arabic. *Advances in Human-Computer Interaction, 2022*, Article 5492230.
- Al-Sa'di, A., & Alsamarraie, H. (2023). Design thinking mindset and self-awareness of user experience practitioners: The case of New Zealand. SSRN. <https://doi.org/10.2139/ssrn.4361224>
- Al-Sa'di, A., & Khrypko, Y. (2024). User experience for online course strategies: EduFlex framework. In 2024 9th International Conference on Business and Industrial Research (ICBIR) (pp. 1587-1592). IEEE.
- Al-Sa'di, A., & Litayem, N. (2023). User frustration: Shaping the experience of automotive and computing. In 2023 IEEE 3rd International Conference on Computer Systems (ICCS) (pp. 56-61). IEEE.
- Al-Sa'di, A., & Miller, D. (2023). Exploring the impact of artificial intelligence language model ChatGPT on the user experience. *International Journal on Technology, Innovation, and Management (IJTIM)*, 3(1), 1-8.
- Al-Sa'di, A., & Miller, D. (2026). Supporting novice creativity in design education through human-centred explainable AI. *Theoretical and Applied Ergonomics*, 2(2), 4.
- Al-Sa'di, A., Yamjal, P., Ahmad, E., Panjabi, R., Allott McPhee, C., & Guler, O. (2023a). Assessing educators' soft skills: Developing a self-assessment instrument. *Administrative Sciences*, 13(9), 208.
- Al-Sadi, A., Moses, T., & Altawabeyeh, S. (2022). iAWHINA: Towards designing an offline disaster mobile application. In *Proceedings of the ISCRAM Asia Pacific Conference 2022*.
- Al-Sa'Di, A., Puri, S. K., & Panjabi, R. (2025a). Trust, negotiation, and skill acquisition in social commerce: Evidence from Facebook Marketplace. In 2025 6th International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 1-5). IEEE.
- Al-Sa'Di, A., Puri, S. K., & Panjabi, R. (2025b). Tracing the

- emotional pathways of Facebook Marketplace. In 2025 6th International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 1–4). IEEE.
- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99, Article 101805.
- Arnold, M., Goldschmitt, M., & Rigotti, T. (2023). Dealing with information overload: A comprehensive review. *Frontiers in Psychology*, 14, Article 1122200.
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM.
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21.
- Cao, S., Liu, A., & Huang, C.-M. (2024). Designing for appropriate reliance: The roles of AI uncertainty presentation, initial user decision, and user demographics in AI-assisted decision-making (arXiv:2401.05612). *arXiv*. <https://doi.org/10.48550/arXiv.2401.05612>
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.
- Colberg, S. R., Sigal, R. J., Yardley, J. E., Riddell, M. C., Dunstan, D. W., Dempsey, P. C., Horton, E. S., Castorino, K., & Tate, D. F. (2016). Physical activity/exercise and diabetes: A position statement of the American Diabetes Association. *Diabetes Care*, 39(11), 2065–2079.
- Ellahham, S. (2020). Artificial intelligence: The future for diabetes care. *The American Journal of Medicine*, 133(8), 895–900.
- Evert, A. B., Dennison, M., Gardner, C. D., Garvey, W. T., Lau, K. H. K., MacLeod, J., Mitri, J., Pereira, R. F., Rawlings, K., & Robinson, S. (2019). Nutrition therapy for adults with diabetes or prediabetes: A consensus report. *Diabetes Care*, 42(5), 731–754.
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127.
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621.
- Lambe, K. A., O'Reilly, G., Kelly, B. D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review. *BMJ Quality & Safety*, 25(10), 808–820.
- Litayem, N., & Al-Sa'di, A. (2023). Exploring the programming model, security vulnerabilities, and usability of ESP8266 and ESP32 platforms for IoT development. In *2023 IEEE 3rd International Conference on Computer Systems (ICCS)* (pp. 150–157). IEEE.
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103.
- Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). ACM.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
- Panjabi, R., Ahmed, A., & Ahmad, E. (2024). Advantage of using AI in education to enhance learning: Review and analysis. In *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)* (pp. 1–6). IEEE.
- Sarsam, S. M., Al-Samarraie, H., Bahar, N., Shibghatullah, A. S., Eldenfria, A., & Al-Sa'Di, A. (2021). Detecting real-time correlated simultaneous events in microblogs: The case of men's Olympic football. In *International Conference on Human-Computer Interaction* (pp. 368–377). Springer International Publishing.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4), 591–611.
- Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1–38.
- Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution* (pp. 196–202). Springer.
- Yamjal, P., & Ahmed, A. (2022). Strategies for retention and completion in vocational education: Faculty perspectives. *Journal of Management and Business Education*, 5(4), 247–265.
- Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 189–201). ACM.
- Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (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* (pp. 295–305). ACM.
- Zuchowski, L. C., Zuchowski, M. L., & Nagel, E. (2024). A trust based framework for the envelopment of medical AI. *npj Digital Medicine*, 7(1), 230.