Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise

Saumya Pareek
The University of Melbourne
Melbourne, VIC, Australia
spareek@student.unimelb.edu.au

Eduardo Velloso
The University of Melbourne
Melbourne, VIC, Australia
The University of Sydney
Sydney, NSW, Australia
eduardo.velloso@sydney.edu.au

Jorge Goncalves
The University of Melbourne
Melbourne, VIC, Australia
jorge.goncalves@unimelb.edu.au

ABSTRACT

Leveraging Artificial Intelligence to support human decision-makers requires harnessing the unique strengths of both entities, where human expertise often complements AI capabilities. However, human decision-makers must accurately discern when to trust the AI. In situations with complementary Human-AI expertise, identifying AI inaccuracies becomes challenging for humans, hindering their ability to rely on the AI only when warranted. Even when AI performance improves post-errors, this inability to assess accuracy can hinder trust recovery. Through two experimental tasks, we investigate trust development, erosion, and recovery during AI-assisted decision-making, examining explicit Trust Repair Strategies (TRSs) – Apology, Denial, Promise, and Model Update. Our participants classified familiar and unfamiliar stimuli with an AI with varying accuracy. We find that participants leveraged AI accuracy in familiar tasks as a heuristic to dynamically calibrate their trust during unfamiliar tasks. Further, once trust in the AI was eroded, trust restored through Model Update surpassed initial trust values, followed by Apology, Promise, and the baseline (no repair), with Denial being least effective. We empirically demonstrate how trust calibration occurs during complementary expertise, highlighting factors influencing the different effectiveness of TRSs despite identical AI accuracy, and offering implications for effectively restoring trust in Human-AI collaborations.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **HCI theory, concepts and models**.

KEYWORDS

Human-AI decision-making, complementary expertise, trust development, trust repair

ACM Reference Format:
Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2024. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3630106.3658924

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658924

1 INTRODUCTION

The integration of human expertise and Artificial Intelligence (AI) in Human-AI collaboration is often driven by the recognition that humans and machines possess complementary strengths, where the sum can be greater than its parts. Such cases of AI-assisted decision-making, where AI advises human decision-makers [4, 5, 67], are becoming increasingly prevalent in domains such as medical diagnostics [9, 35] and criminal justice [12, 38]. However, AI systems are fallible, leaving users with determining when to trust them. In doing so, users may display unwarranted reliance on AI (overtrust) or undue scepticism despite its capability (undertrust) [14, 34]. Fostering *appropriate* trust is pivotal for collaboration, so recent works have explored methods to calibrate users' trust to reflect the actual capability of AI systems [14, 22, 29, 41].

Prior works investigating accuracy-based trust calibration typically involve systems where accuracy is evident to users through performance feedback [66] or prior task expertise that allows users to spot system inaccuracies [25, 43]. In practice, performance cues may not always be available. Moreover, humans and intelligent aids often have complementary strengths, and may oscillate between being experts and non-experts in different task facets. For example, in a trivia game, an AI may excel at memorising vast information to quickly identify quotes, while humans may skilfully connect evidence and solve wordplay puzzles [19]. Knowing how significantly observed system accuracy shapes trust [64], a question arises: *How does trust evolve when users cannot assess AI accuracy for all tasks, as seen in scenarios involving complementary expertise between users and AI?* During overlapping expertise, users *can* assess an AI's performance and rely on it accordingly [4, 44]. However, during complementary expertise, it is unclear how perceived AI accuracy in high human-expertise (**HHE**) tasks influences users' trust in its recommendations for low human-expertise (**LHE**) tasks.

Furthermore, trust calibration entails *both* aligning user trust with system capabilities and appropriately rebuilding trust when diminished. While increased AI accuracy *can* potentially restore users' trust [54, 65], this approach bears two caveats. An accuracy boost often fails to fully reinstate trust to pre-violation levels, and this "recovery" method presupposes that users can detect increased accuracy, which may not be true during an expertise divide. This presents an opportunity to examine the utility of deliberate

interventions to restore trust in AI, which we term Trust Repair Strategies (TRSs).

We adopt TRSs from Human-Robot Interaction (HRI), rooted in the social psychology of interpersonal interactions, and investigate their utility in AI-assisted decision-making. In HRI, trust repair commonly employs three strategies [13]: expressing regret and saying "I'm sorry" (Apology) [31, 36, 56], rejecting culpability for the error (Denial) [3, 40, 55], and conveying intentions to perform better in future interactions (Promise) [10, 17, 52]. Additionally, inspired by research in Human-AI interaction regarding the effects of simulated machine learning model updates on user trust [60], we propose a novel TRS named Model Update, which conveys that the model underlying the AI's decisions has been upgraded.

Addressing the aforementioned research gaps, we seek to answer the following research questions:

- **RQ1:** How does perceived AI accuracy for High Human-Expertise (HHE) tasks influence users' trust in its recommendations for Low Human-Expertise (LHE) tasks?
- **RQ2:** During complementary Human-AI expertise, how does trust recover as accuracy improves, both with and without deploying explicit Trust Repair Strategies (TRSs)?

To investigate these questions, we conducted a survey-based between-subjects experiment involving 300 participants, with 150 participants assigned to each of the two tasks. In each task, participants collaboratively classified images with a simulated AI with varying accuracy. The first task involved real (*Familiar*) and fabricated (*Unfamiliar*) geometric shapes (inspired by Zhang et al. [66]), while the other involved common (*Familiar*) and obscure (*Unfamiliar*) animals (following other studies on animal identification [29, 43, 44]). Both tasks operationalised **complementary expertise**, with participants inherently being experts in identifying *Familiar* stimuli (HHE tasks) but not *Unfamiliar* stimuli (LHE tasks). While fabricating the *Unfamiliar* stimuli for the shape task strictly controlled for participants' prior knowledge, the animal task examined trust dynamics in a more ecologically-valid setting.

Each task had three phases: Phase 1 (high AI accuracy), Phase 2 (low AI accuracy), and Phase 3 (high AI accuracy). Between Phases 2 and 3, we manipulated the TRS between participant groups, which included No Repair (baseline), Apology, Denial, Promise, and Model Update. This enabled us to analyse trust recovery in Phase 3 through accuracy improvement alone and in conjunction with explicit TRSs. We measured participants' agreement with the AI's classification in each task trial and their overall trust after each phase using a validated trust scale [27].

We found consistent results across both tasks. Participants relied on perceived AI accuracy in HHE tasks as a heuristic to dynamically calibrate their trust during LHE tasks. As the perceived AI accuracy for *Familiar* stimuli deteriorated, so did participants' reliance on it for *Unfamiliar* stimuli **(RQ1)**. Moreover, our results re-establish how accuracy improvement leads to partial—but not total—trust recovery, necessitating approaches to complement it **(RQ2)**. Notably, Model Update was the most effective TRS, with users' restored trust surpassing pre-violation levels. This was followed by Apology, rebuilding trust through the AI appearing regretful. Promise, No Repair (baseline), and Denial proved less effective, with users being sceptical of the AI's capacity to make promises, and denial of mistakes exacerbating distrust.

This study makes the following contributions. First, we adopt TRSs from Human-Robot interaction, and demonstrate their utility in Human-AI interactions, reporting different user behaviours. We outline differences in the impact of TRSs, despite identical AI accuracy, discussing the role of anthropomorphism, regret, intentional agency, deception, and the nature of promises (behavioural vs. technical) in trust restoration. Further, the inclusion of two distinct classification tasks enhances the robustness and ecological validity of our findings. Second, we address a critical research gap where users possess complementary expertise with the AI, and cannot always gauge AI accuracy. We provide evidence that in such scenarios, users employ the perceived AI accuracy in *Familiar* (HHE) tasks as a heuristic to calibrate their trust in it for tasks beyond their expertise (LHE). Third, we underscore the dual nature of such accuracy-based trust calibration – depending on the similarity between the AI's accuracy for LHE and HHE tasks, this heuristic can foster appropriate trust or unwarranted (dis)trust. We conclude by discussing implications for AI systems.

2 RELATED WORK

Trust is often defined as the trustor's willingness to put themselves at risk while expecting the other party *(the trustee)* to act benevolently [48]. In this work, we adopt the widely utilised definition put forth by Lee and See [34], who describe trust as *"an attitude that an agent will achieve an individual's goal in a situation characterised by uncertainty and vulnerability."*

The conceptualisation of trust as a dynamic, temporal attribute of Human-AI collaboration has been extensively examined in recent studies, revealing that accuracy shapes users' trust in AI systems [44, 54]. For example, Yu et al. [65] examined trust dynamics over several interactions with AI and found a positive correlation between users' trust and perceived system accuracy. Similarly, Yin et al. [64] report that users' trust in a system is significantly affected by its observed accuracy during interactions, irrespective of any stated accuracy. Notably, initial impressions of an intelligent system also significantly influence trust dynamics for the entire interaction [15, 44], indicating that system errors early on can cause negative trust outcomes, even if accuracy improves subsequently [54].

Of note is the common characteristic of the systems evaluated in previous research – their accuracy was often readily apparent to end-users, either by displaying explicit performance metrics or through the users being task experts, equipping them to spot AI errors [25, 44]. However, when indicators of an AI's performance are not provided, individuals tend to over-rely on the AI regardless of its actual accuracy, even when explanations are provided [45]. In real-world scenarios, performance cues are not always available. Furthermore, Human-AI collaborations often involve a division of expertise, where humans excel in some aspects while the AI in others, effectuating the need to collaborate with each other in the first place. These findings present an opportunity to examine trust dynamics in a scenario where end-users and the AI have complementary expertise.

2.1 Human-AI Complementary Expertise

Complementary Human-AI expertise refers to a scenario where both entities possess distinct strengths which need to be leveraged appropriately to reach a state of superior performance [5]. For example, in a trivia game, an AI can proficiently memorise vast information to identify quotes faster than humans, while humans can adeptly chain evidence and solve wordplay [19]. Another compelling instance of complementary Human-AI expertise emerges in medical diagnosis, where AI may excel at analysing medical images efficiently while doctors emphasise with patients and obtain a holistic understanding of their condition within the context of their life [47]. In such cases, the role of the human decision-maker becomes crucial—they must judge when and to what extent should AI advice influence decisions.

In cases of overlapping expertise, end-users *can* distinguish when an AI is an expert in a given task and subsequently fine-tune their dependence on it [44, 66]. However, the challenge in examining these situations is in disentangling the effects of trust and prior knowledge—if a user followed the AI's recommendation, is it because they trust the AI or because they know the correct answer? Importantly, performance during complementary expertise depends not only on whether the AI's recommendations can compensate for low human expertise but also on whether the human can trust the AI *only* when it is warranted – calibrating their trust with the AI's accuracy. Empirical works have so far investigated trust calibration in situations where the users' expertise remains constant, being either experts or non-experts [5, 39, 44].

In our research, we take a distinct approach by examining a clear division in task expertise between humans and AI. Unlike prior studies that explore more balanced expertise overlaps [5, 39, 67], we explore a scenario where humans excel in specific task aspects while having little-to-no knowledge in others, enabling us to separate the effects of trust and prior knowledge. We seek to understand how perceived AI expertise in high-human expertise tasks influences reliance on AI in low-human-expertise tasks, as users navigate between areas of proficiency and unfamiliarity.

2.2 Trust Calibration and Repair

Poorly calibrated trust, whether stemming from unwarranted reliance on AI (overtrust) or undue scepticism despite AI competence (undertrust), impacts utilisation of intelligent systems [26, 62]. In AI-assisted decision-making, users' trust must be appropriately calibrated so it "matches the true capabilities of automation" [14, 34]. During overlapping expertise when users can spot system errors, a drop in AI accuracy can reduce users' trust [65], and subsequent improvements to AI accuracy *can* contribute to trust restoration [54, 65]. However, a mere accuracy boost often falls short of fully restoring trust to pre-violation levels, and assumptions about users' ability to perceive accuracy improvements may not hold true, particularly when an expertise gap exists between users and AI such as in cases of complementary task expertise.

2.2.1 Trust Repair Strategies (TRSs). The inability of increased accuracy to effectively restore trust highlights the need for utilising explicit **Trust Repair Strategies (TRSs)**, enabling users to restore their lost trust when appropriate. In this work, we study TRSs from Human-Robot Interaction (HRI) literature, namely, Apology, Denial, and Promise (see [13]), which are rooted in the social psychology of interpersonal interactions. We examine their utility in an AI-assisted decision-making context with complementary task expertise, for two reasons. First, in HRI, TRSs have been examined in interactions with physical robots, where the robot's tangible presence [2, 37] and facial expressions can influence trust [21, 51, 57]. However, Human-AI interactions differ significantly with the AI lacking tangible features. Second, TRSs studies in HRI primarily employ tasks with readily observable robot accuracy [20, 28, 32, 49] or explicit performance feedback [31]. However, it remains to be seen how TRSs perform when users cannot always assess the accuracy of intelligent agents, such as during complementary expertise scenarios, given the strong influence of observed system accuracy on trust [64].

In this work, we explore three prominent TRSs from HRI that encapsulate the principles of trust repair:

(1) **Apology**: This TRS involves an expression of regret, such as saying "I'm sorry" [31, 36, 56]. Apologies operate primarily on an emotional level and aim to change how the trustor perceives the trustee [16]. They function as social rituals, elevating the social standing of the trustee and reinstating social expectations after an error [10].
(2) **Denial:** This TRS involves rejecting culpability for a trust violation [3]. Denial seeks to shift the locus of causality associated with the violation, essentially redirecting blame away from the trustee [40, 55]. By doing so, denial aims to absolve the trustee of any wrongdoing, mitigating the negative consequences of the violation [17].
(3) **Promise:** This TRS is an assertion made by a trustee to convey positive intentions regarding future actions [52]. For instance, saying, "I promise I will do better next time" constitutes a promise. Unlike apologies and denials, promises directly address how the trustee is expected to behave in the future [10, 17].

Furthermore, the dynamic nature of AI systems implies that their performance evolves over time, incorporating more data and algorithmic advances into their models. Recent research in Human-AI interaction has explored (simulated) model updates, highlighting that initial impressions of an AI's decision-making model can influence users' trust in it [44, 54]. Furthermore, for users possessing prior knowledge in a task domain, subjective trust tends to fluctuate as the model and its outputs evolve [60]. Inspired by these findings, we propose and examine a novel TRS in our work:

(4) **Model Update:** This strategy involves an AI system (the trustee) conveying to the end-user that the model or algorithm underlying its decision-making has been updated. Model updates attempt to rebuild trust by showing that the AI is actively trying to address the factors that caused the error.

While Model Update and Promise attempt to rebuild trust by conveying intentions to improve future performance, they are normatively distinct. Promises involve the AI committing to *behavioural* changes, which can have uncertain results, while model updates convey *technical* enhancements, whose perceived impact on the AI's accuracy can vary.

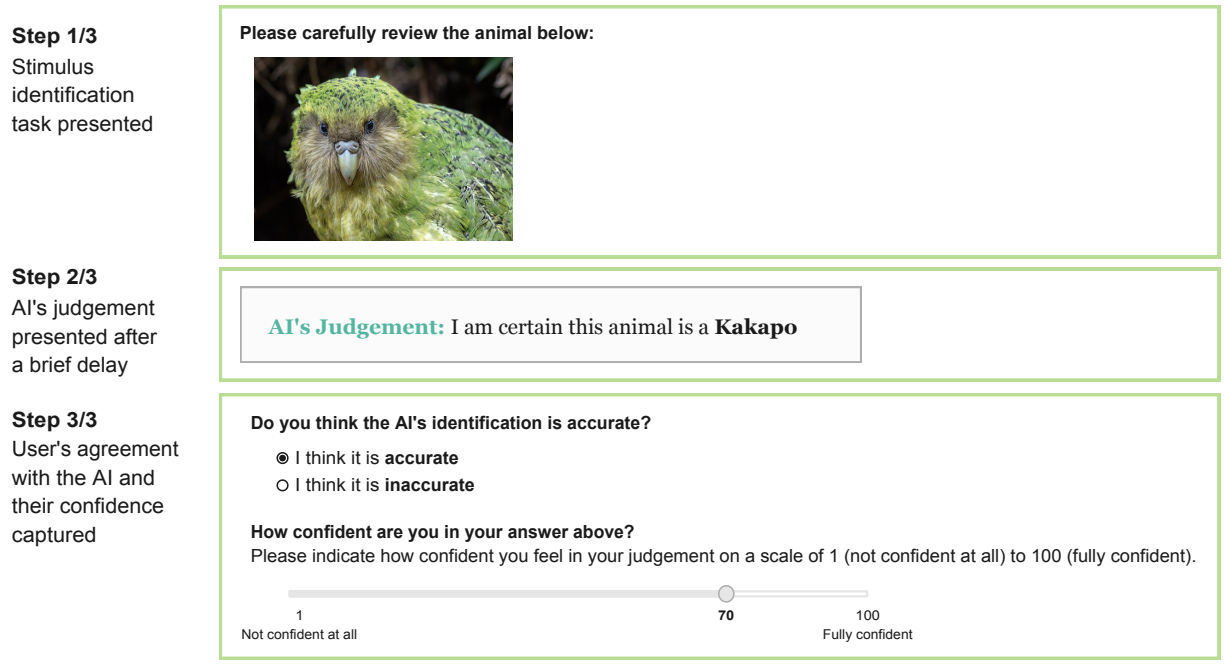

Figure 1: An example Low Human-Expertise (LHE) trial for the ANIMAL task, where the AI accurately identifies the animal, progressively presenting each step. Note that no feedback on the AI's performance is provided to participants. The SHAPE task progressed similarly.

3 METHOD

3.1 Human-AI Collaboration Tasks

3.1.1 Task Selection and Design. For complementary Human-AI task expertise to manifest, we required tasks featuring a distinct expertise divide between participants and the AI. The tasks needed to contain *High Human-Expertise* (**HHE**) trials, where all participants know the correct answer so they can judge the AI's response accuracy, and *Low Human-Expertise* (**LHE**) trials, where all participants do not know the correct answer, so we can measure their trust in the AI by analysing their agreement with AI recommendations. This setup allowed us to operationalise complementary Human-AI expertise.

Considering these factors, we designed two classification tasks, one involving shapes and the other animals. In the SHAPE task, inspired by Zhang et al. [66], we created an expertise divide by presenting both *Familiar* shapes (Circle, Rectangle, Triangle), and *Unfamiliar* shapes created specifically for this study ("Pyrangle", "Scleratice", "Tenectus"). Unlike Zhang et al. [66], we omitted presenting performance feedback so participants' knowledge of the *Unfamiliar* shapes does not improve as the study progresses, maintaining the expertise divide. Similarly, in the ANIMAL task, participants encountered *Familiar* (e.g., Cat, Dog, and Horse) and *Unfamiliar* animals (e.g., Ptarmigan, Markhor, and Perentie). In both tasks, participants reported their agreement with the AI's classification. An example task trial featuring an *Unfamiliar* animal is illustrated in Figure 1.

The SHAPE task facilitates a strongly controlled setting by fabricating the *Unfamiliar* stimuli, ensuring zero a priori knowledge. Further, the ANIMAL task allows for the examination of trust dynamics in a more ecologically-valid setting. Overall, this dual-task approach expands the breadth of our exploration, enabling a comprehensive investigation of trust dynamics.

We hypothesised that participants would correctly identify the AI's accuracy for *Familiar* stimuli (ensuring HHE), but not for *Unfamiliar* stimuli (establishing LHE). For example, in an HHE trial, participants would promptly recognise the AI's accuracy when it misidentifies a Rectangle as a Circle, or an Octopus as a Snake. This arrangement enabled us to investigate whether the AI's perceived accuracy for HHE tasks influences participants' trust in it for LHE tasks. It also allowed us to effectuate trust violations for studying repairs later by making the AI give erroneous recommendations for HHE tasks. Moreover, participants being non-experts for LHE tasks ensured a need to rely on the AI, simulating real-world situations where individuals depend on AI systems for tasks beyond their expertise.

3.1.2 Classification Stimuli. **SHAPES:** We designed 5 visual variants of the 3 *Familiar* and 3 *Unfamiliar* shape categories, resulting in 30 stimuli. *Familiar* shapes were typical Circles, Rectangles, and Triangles, and each variant had random differences in visual characteristics. Following Zhang et al. [66], *Unfamiliar* shapes were closed 2D shapes designed from Bezier curves, with specific combinations of features, such as the (dis)similarity between border and fill patterns (dots, dashes, or both). To further increase visual complexity and hinder participants from learning patterns [66], we randomly varied category-irrelevant features, such as fill colour, edge length, curvature, interior angles, and pattern size and spacing. We also

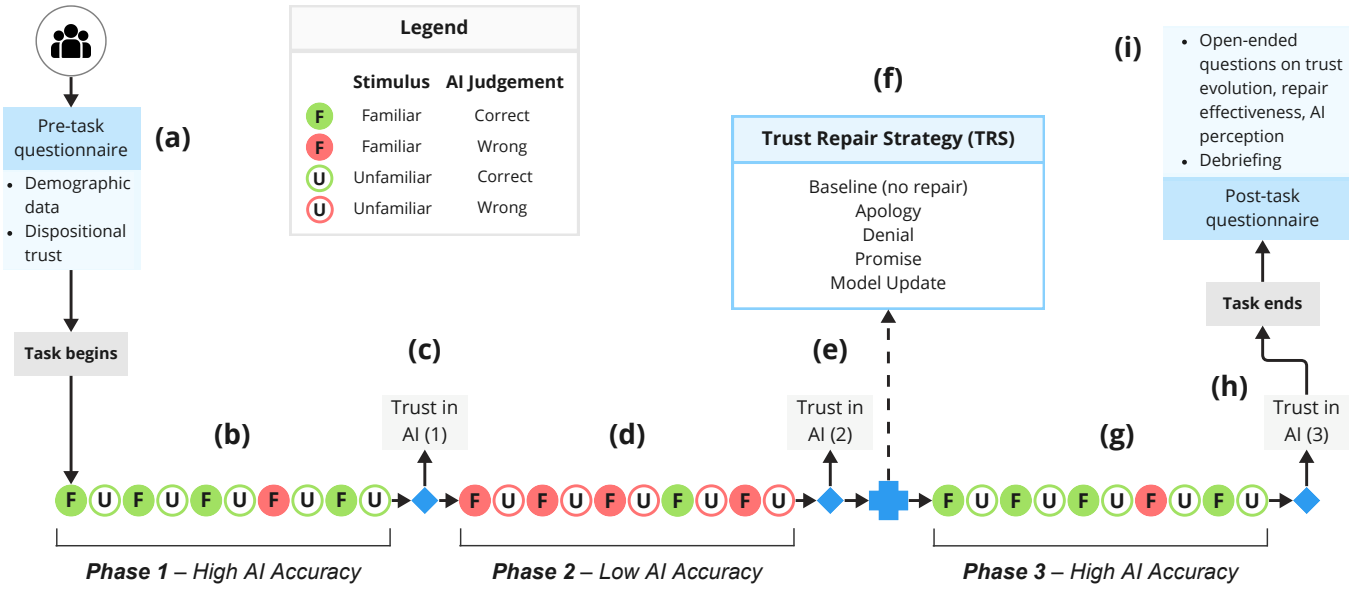

Figure 2: The full experiment flow. All participants undergo 3 PHASES of classification task trials, and the Trust Repair Strategy (TRS) varies between treatments. (a): Pre-task questionnaire. (b): PHASE 1 with high AI accuracy, to engender trust in the AI. (c): Measurement of trust in AI after PHASE 1. (d): PHASE 2 with low AI accuracy, to erode trust in the AI. (e): Measurement of trust in AI after PHASE 2. (f): Type of TRS displayed to participants. (g): PHASE 3 with high AI accuracy. (h): Final measurement of trust in AI after PHASE 3, to examine how effective the shown TRS was. (i): Open-ended questions (and debriefing for those exposed to fabricated shapes.)

named *Unfamiliar* shapes to sound like plausible obscure shapes, without disclosing geometric attributes in the name's etymology.

ANIMALS: We chose 15 *Familiar* (species commonly encountered in everyday life) and 15 *Unfamiliar* animals (species infrequently heard of due to their limited population or geographic distribution). The familiarity of animals was determined after cross-referencing several kinds and sources of data, such as geographic spread from Kaggle (e.g., Animal Information Repository), population size and prevalence data from the International Union for Conservation of Nature (IUCN) Red List [1], and public familiarity through online quizzes on obscure species. This methodology ensured a robust selection process, combining scientific assessments of animal prevalence with popular perceptions of animal familiarity. Further, we chose *Unfamiliar* animals with names that did not provide clues about their appearance or characteristics, such as selecting an 'Aye-Aye' but not a 'Red-Shanked Douc'. The full stimulus set for both tasks is included in the supplementary materials.

3.1.3 Trust Repair Strategies (TRSs). Our operationalisation of the four Trust Repair Strategies (TRSs) was based on the trust repair literature in Human-Agent interaction [30, 31] and Human-Robot Interaction (HRI) [3, 17], also drawing from the social psychology of interpersonal trust [36, 42]. To begin, we identified the core trust-related components of each TRS – expressing regret (APOLOGY) [10, 56], rejecting culpability (DENIAL) [3, 40], commitment to future behavioural changes (PROMISE) [17, 52], and indications of technical enhancements (MODEL UPDATE) [60].

We established a *structure* for the TRS texts: they begin with the AI acknowledging a deterioration in participants' trust, followed by embodying the core trust-related component of the TRS, and end with the AI hoping that participants can trust it again. To formulate TRSs which closely resemble AI-generated responses and lack unwanted variability, we leveraged OpenAI's ChatGPT (GPT-3.5). The authors vetted the generated TRSs over several iterations, ensuring that the texts accurately represent their repair strategy and have a consistent structure without unintended variability. This helped us precisely operationalise the TRSs, ensuring that any differences in participant behaviour between TRSs are strictly owing to the repair strategy. The prompt and the generated TRS texts are included in the supplementary materials.

3.2 Experimental Design

Figure 2 presents an overview of our experimental design. For each task, we manipulated the presence and type of TRS, giving rise to 5 experimental conditions: No Repair (baseline), APOLOGY, DENIAL, PROMISE, and MODEL UPDATE.

3.2.1 Participants. We deployed our study on Prolific, recruiting fluent English speakers with an approval rating ≥ 98%. Participants engaged in either the SHAPE or ANIMAL task, and the Human Ethics Committee of our university approved the study. Sample size determination using G*Power [18], with a medium effect size (f^2 = 0.25), α = 0.05, and a power of 0.8 [11], indicated a minimum of 135 participants per task. We conservatively recruited 150 participants per task to uphold reliability. Participants spent a median of 14 minutes on the survey and received US$4 for participation. Overall,

[1] https://www.iucnredlist.org/

we collected valid data from 300 participants—150 for each of the two tasks, with 30 participants in each of the five conditions.

3.2.2 Procedure. Both ANIMAL and SHAPE tasks progressed identically. In each task, participants were randomly assigned to one of the five experimental conditions and shown a counterbalanced sequence of classification stimuli. The survey began with a pre-task questionnaire collecting participants' demographic details, and presented the TiA-PtT questionnaire (Trust in Automation – Propensity to Trust subscale [33]) which measures dispositional trust in automation (Figure 2a). We then briefed participants that they would collaborate with an AI on an ANIMAL/SHAPE classification task.

Following previous research on Human-AI decision-making [8, 43, 58], we opted for a simulated AI rather than a trained machine learning model to maintain control over when the AI should make errors, and what these errors should look like, allowing for a precise manipulation of the AI's performance across conditions and participants.

The overall task comprised three sequential PHASES, with the AI's identification accuracy changing between PHASES. The AI exhibited high identification accuracy of 80% in **PHASE 1** (Figure 2b), low accuracy of 20% in **PHASE 2** (Figure 2d), and high accuracy of 80% again in **PHASE 3** (Figure 2g). This sequence of accuracy enabled us to attempt to foster trust in the AI (PHASE 1), erode it (PHASE 2), and investigate the degree to which increased accuracy (PHASE 3) restores trust in case of complementary expertise, both with and without deploying explicit TRSs. Further, in PHASE 1 and PHASE 3, our simulated AI was configured to make its sole mistake on trial number 7. This was necessary because first impressions of intelligent systems influence the overall trust dynamics, with early mistakes being costlier than those later on [44, 54].

Each PHASE comprised ten classification **trials** (Figure 2 (b, d, g)). Figure 1 illustrates an ANIMAL task trial, step-by-step. In each trial, participants viewed an image stimulus, followed by a 3-second delay before the (simulated) AI presented its identification. The delay simulated the operation of an actual AI, and allowed participants enough time to make their own identification, as delaying the presentation of AI advice can enhance critical thinking at decision-making time [8, 46]. Participants then reported their ***agreement*** with the AI's classification – a trust-related behavioural measure relative to AI performance, suitable for such tasks [59]. Participants also indicated their confidence in their agreement on a scale of 1 to 100, with higher scores indicating greater confidence. To minimise potential bias from the initial slider position [53], an anchor appeared only after participants clicked on the slider's range. The task sequence interwove *Familiar* (HHE) and *Unfamiliar* (LHE) stimuli — enabling us to examine whether perceived AI accuracy for HHE trials influences trust in the AI's judgement for LHE trials.

In addition to the behavioural trust metric, we also deployed a self-report measure, following studies investigating trust calibration [62]. After each PHASE, participants reported their ***phase-level trust*** in the AI on a validated 12-item 7-point Likert scale, ranging from 1 (Not At All) to 7 (Extremely) [27] (Figure 2 (c, e, h)).

To operationalise *accurate* AI judgements, our AI classified all stimuli correctly. For *inaccurate* AI judgements, to remove ambiguities, the AI chose misclassification labels from within the same stimulus category, misclassifying a *Familiar (Unfamiliar)* stimulus as another *Familiar (Unfamiliar)* one (e.g., a Rectangle as a Triangle, and a "Scleratice" as a "Pyrangle"). This also ensured participants do not receive accuracy cues from incorrect labels (e.g., from labelling a "Tenectus" as a Circle).

We administered the **Trust Repair Strategy (TRS)** between PHASE 2 (low AI accuracy) and PHASE 3 (high AI accuracy) (Figure 2f) to participants not assigned to the baseline condition. Additionally, to increase the authenticity of the AI's model being updated in the MODEL UPDATE condition, we incorporated a 6-second delay between the TRS text and PHASE 3, following similar studies around simulated model updates [60]. The task concluded after PHASE 3.

We then posed open-ended questions to learn about participants' trust evolution, factors influencing their perception of the AI's accuracy for *Unfamiliar* stimuli, reasons behind their (mis)trust in the AI, and how they perceived the TRSs (Figure 2i). After these questions, participants doing the SHAPE task were briefed about the artificial nature of some stimuli. We systematically coded the qualitative responses following a deductive thematic analysis approach [7]. We started by establishing a coding framework rooted in themes derived from literature and our research objectives. We gained a holistic understanding of our qualitative data for each task, labelling participants' responses based on our pre-established themes.

4 RESULTS

We recruited 150 participants per task, with a mean age of 35 years (*SD* = 13.01) for SHAPES and 34.2 years (*SD* = 11.85) for ANIMALS. Participants reported their agreement with the AI's classification in 30 task trials, resulting in 4500 agreement measurements. The task included 15 High Human-Expertise (HHE) trials (*Familiar* stimuli) and 15 Low Human-Expertise (LHE) trials (*Unfamiliar* stimuli). Overall, there were 2250 instances per task where we measured participants' agreement with the AI in LHE trials. Our intention was not to compare agreement behaviour between HHE and LHE trials. Instead, they played distinct roles — HHE trials fostered or eroded users' trust in the AI through its perceived accuracy, while LHE trials allowed us to capture the resultant trust by examining users' agreement with the AI for tasks beyond their expertise.

4.1 Quantitative Findings

4.1.1 Robustness and Manipulation Check. To confirm participants' high (low) expertise in HHE (LHE) trials, we analysed the accuracy of their agreement with the AI's classification of *Familiar* and *Unfamiliar* stimuli, as well as the difference between these values. In both tasks, participants exhibited higher accuracy in HHE trials and lower accuracy in LHE trials, with the considerable difference between these values emphasising a successful expertise divide. Participants had an accuracy of 99.73% (*SD* = 0.33) for *Familiar* shapes (6 errors out of 2250 responses, each made by a distinct participant, 2 per PHASE), and only 49.77% (*SD* = 14.34) for *Unfamiliar* shapes. Similarly, the mean accuracy was 90.35% (*SD* = 10.97) for *Familiar* animals (217 errors out of 2250 responses, 77 each in PHASE 1 and 2, and 63 in PHASE 3), demonstrating higher familiarity, while being only 49.17% (*SD* = 10.60) for *Unfamiliar* animals.

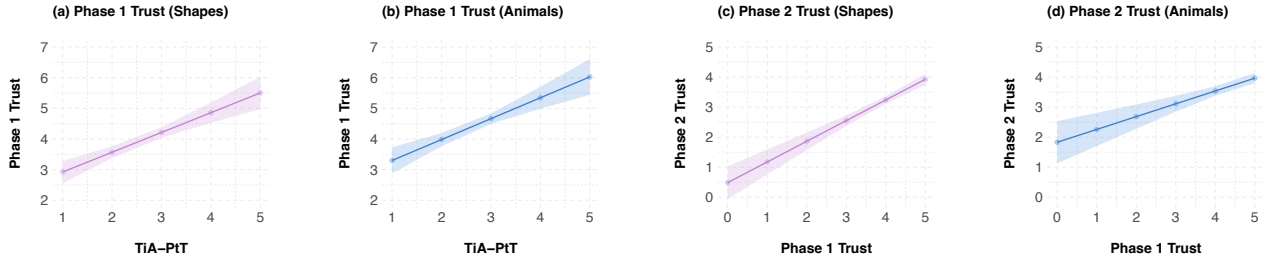

Figure 3: Trust dynamics in Phase 1 and Phase 2. The effect of participants' Trust in Automation (Propensity to Trust subscale) on trust in Phase 1 for (a) Shapes; (b) Animals, and the effect of Phase 1 trust on Phase 2 trust for (c) Shapes; (d) Animals. Shaded area denotes standard error (SE).

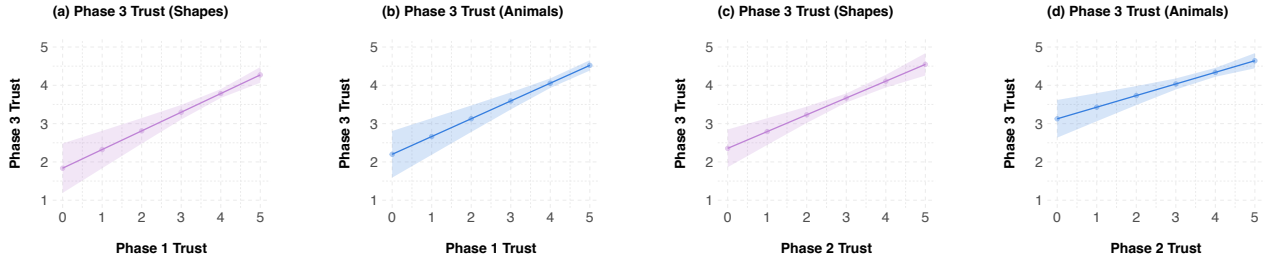

Figure 4: Trust dynamics in Phase 3. The effect of Phase 1 trust on Phase 3 trust for (a) Shapes; (b) Animals, and the effect of Phase 2 trust on Phase 3 trust for (c) Shapes; (d) Animals. Shaded area denotes standard error (SE).

The substantial difference between participants' HHE and LHE accuracy (49.96 points for Shapes and 41.18 for Animals) underscores an effective expertise manipulation. The more pronounced divide for shapes, as hypothesised, is due to the *Unfamiliar* stimulus being fabricated, ensuring participants' lack of prior knowledge. Importantly, the nearly identical LHE accuracy in both tasks indicates that our manipulation effectively produced low participant expertise across both domains. Participants also demonstrated significant confidence in their decisions for *Familiar* Shapes (M = 99.01, SD = 0.80) and animals (M = 97.59, SD = 1.62), compared to *Unfamiliar* shapes (M = 43.01, SD = 3.09) and animals (M = 46.80, SD = 4.27). This further signals higher certainty in their decisions about *Familiar* stimuli compared to *Unfamiliar*.

Overall, within each task, participants *always* demonstrated more knowledge of *Familiar* stimuli compared to *Unfamiliar*, validating the existence of an expertise divide. These results enhance the likelihood that any observed trust dynamics result from our experimental manipulations, rather than external factors, allowing us to draw causal inferences.

4.1.2 Influence of Perceived Accuracy on Agreement. We sought to investigate the influence of perceived AI accuracy in HHE tasks on users' agreement during LHE tasks. Since *Familiar* stimuli (HHE) always preceded *Unfamiliar* stimuli (LHE) in the task sequence (Figure 2), we built a generalised linear mixed-effects model (GLMM) of agreement at LHE trials with AI accuracy in the preceding HHE trial as the predictor, for each task. This allowed us to evaluate the impact of our predictor variable on our outcome variable (agreement) with a non-normal distribution. We included participant IDs as a random effect to account for individual variances in the model, and utilised the statistical R package `lme4` [6].

We observed a significant difference in agreement at an LHE trial based on the AI's perceived accuracy at the previous HHE trial, for both Shapes (β = -0.472, SE = 0.034, $p < 0.001$) and Animals (β = -0.576, SE = 0.064, $p < 0.001$) **(RQ1)**. Participants were more likely to trust the AI's classification of an *Unfamiliar* stimulus when it accurately identified the *Familiar* stimulus preceding it in the task sequence, with an odds ratio of 1.62 (95% CI between 1.58 and 1.67) for Shapes, and 1.56 (95% CI between 1.50 and 1.64) for Animals. These results demonstrate that during complementary expertise, an increase in the AI's perceived accuracy in HHE tasks fosters more trust in LHE tasks, and vice-versa.

4.1.3 Trust Development and Repair. In both tasks, trust significantly decreased from Phase 1 (Shapes: M = 3.94, SD = 1.04; Animals: M = 4.42, SD = 1.00), to Phase 2 (Shapes: M = 3.20, SD = 1.05; Animals: M = 3.72, SD = 1.01), $t(149) = 11.0$, $p < 0.001$ (Shapes) and $t(149) = 8.44$, $p < 0.001$ (Animals). This indicated a successful trust reduction in Phase 2 for our recovery efforts in Phase 3. Our goal was to examine trust dynamics across Phases, investigate how participants' dispositional trust in AI (TiA-PtT) moderates trust development, and evaluate Trust Repair Strategies (TRSs). For each task, we used the statistical R package `stats` to build three linear models, one for each Phase, to granularly assess the impact of various factors.

Modelling **Phase 1** Trust. Through the first model, we investigated how participants' Phase 1 trust is influenced by their TiA-PtT. We observed a significant main effect of TiA-PtT on trust in Phase 1, for both Shapes (β = 0.645, SE = 0.105, $p < 0.001$) (Figure 3a) and Animals (β = 0.681, SE = 0.122, $p < 0.001$) (Figure 3b). In both tasks, participants with a higher trust in automation reported greater trust the AI in Phase 1, where the AI with complementary expertise exhibited high accuracy.

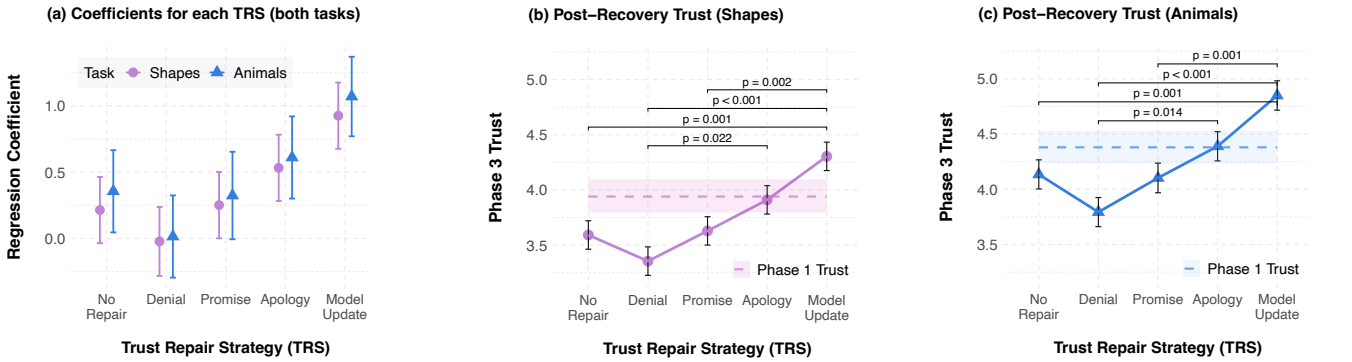

Figure 5: The influence of Trust Repair Strategies (TRSs) on Phase 3 trust. Individual regression estimates for TRSs for (a) both tasks. Pair-wise comparisons for (b) Shapes and (c) Animals. The dotted line in (b) and (c) represents mean trust during Phase 1, and the associated shaded area denotes 95% confidence intervals (CI). Error bars denote standard error (SE).

Modelling **Phase 2** Trust. Building on our incremental analysis, in the second model we examined how participants' Phase 2 trust is influenced by their Phase 1 trust and TiA-PtT. We observed a significant main effect of Phase 1 trust on Phase 2 trust, for both Shapes (β = 0.688, SE = 0.068, $p < 0.001$) (Figure 3c) and Animals (β = 0.426, SE = 0.078, $p < 0.001$) (Figure 3d). Participants who trusted the AI more during Phase 1 (high AI accuracy) also reported higher trust during Phase 2 (low AI accuracy). Furthermore, we did not find a significant effect of TiA-PtT on Phase 2 trust in either task. This suggests that participants' initial trust in AI (during Phase 1) had a more substantial impact on their trust development in Phase 2 than their dispositional trust in automation.

Modelling **Phase 3** Trust. In our final model we examined how participants' Phase 3 trust is influenced by their Phase 1 trust, Phase 2 trust, TiA-PtT, and the Trust Repair Strategy (TRS). Phase 1 trust significantly impacted Phase 3 trust, for both Shapes (β = 0.487, SE = 0.081, $p < 0.001$) (Figure 4a) and Animals (β = 0.464, SE = 0.068, $p < 0.001$) (Figure 4b). Similarly, Phase 2 trust also significantly impacted Phase 3 trust, for both Shapes (β = 0.438, SE = 0.075, $p < 0.001$) (Figure 4c) and Animals (β = 0.303, SE = 0.065, $p < 0.001$) (Figure 4d). Similar to trust dynamics observed in previous Phases, participants continued to demonstrate the influence of earlier trust levels on subsequent phases. Moreover, TiA-PtT did not impact Phase 3 trust for Shapes (β = 0.040, SE = 0.090, p = 0.658), but did so for Animals (β = 0.428, SE = 0.103, $p < 0.001$).

TRSs. We observed a similar relative effectiveness of TRSs in restoring trust across both tasks. Model Update and Apology were the most effective, surpassing Promise, Denial, and No Repair (baseline) (Figure 5a). We further performed a post-hoc analysis to obtain pairwise contrasts between TRSs (Figure 5 (b, c)), and found statistically significant differences when comparing Apology and Model Update with the other TRSs, further emphasising their greater effectiveness. Model Update was the most influential in causing participants to regain trust in the AI: Model Update vs Baseline (Shapes: β = -0.711, SE = 0.182, p = 0.001; Animals: β = -0.715, SE = 0.187, p = 0.001), Model Update vs Denial (Shapes: β = -0.948, SE = 0.182, $p < 0.001$; Animals: β = -1.056, SE = 0.187, $p < 0.001$), and Model Update vs Promise (Shapes: β = -0.674, SE = 0.182, p = 0.002; Animals: β = 0.747, SE = 0.187, p = 0.001). Apology was also significantly influential when compared to Denial (Shapes: β = 0.554, SE = 0.182, p = 0.022; Animals: β = 0.596, SE = 0.187, p = 0.014).

4.2 Qualitative Findings

At the survey's conclusion, participants answered open-ended questions about their trust evolution through the study. We sought insights into factors influencing their (dis)agreements with the AI for *Unfamiliar* stimuli—classification tasks for which they had low expertise. Our focus was also on understanding the factors influencing the effectiveness of Trust Repair Strategies (TRSs). We systematically coded the responses following a deductive thematic analysis approach [7]. We started by establishing a coding framework rooted in themes derived from literature and our research objectives. We gained a holistic understanding of our qualitative data for each task, labelling participants' responses based on our pre-established themes. We systematically assigned responses to themes during the coding process. The author team met repeatedly to discuss any discrepancies and arrive at a consensus. Next, we present our main findings.

4.2.1 Influence of Complementary Expertise on Trust. We found that across both tasks, the majority of participants utilised the AI's classification accuracy for *Familiar* stimuli (HHE trials) as a heuristic to guide their trust in its output for *Unfamiliar* stimuli (LHE trials); *"If the AI correctly identified a [Familiar] shape, I was more likely to trust it for [Unfamiliar] shapes, and vice versa." (P13, Baseline, Shapes).* Additionally, these dynamics evolved granularly, with the AI's accuracy for the previous *Familiar* stimulus strongly impacting trust during the current *Unfamiliar* stimulus, a behaviour also salient in our quantitative findings; *"If the AI did well for the previous [Familiar] animal, I found it more trustworthy for the current [Unfamiliar] animal." (P28, Baseline, Animals).*

4.2.2 Effectiveness of Trust Repair Strategies (TRSs). In the **Baseline** condition (no repair) across both tasks, trust in the AI primarily hinged upon its perceived accuracy for *Familiar* stimuli. However, the increased accuracy of Phase 3 could not restore trust; *"I trusted the AI in the first [phase] because the shapes I knew it got fully correct. In the next two [phases], my trust was gone as the AI made mistakes on [Familiar] shapes, and I could no longer trust it for [Unfamiliar] shapes."* (P11, Baseline, Shapes). For some participants

in both tasks, increased accuracy partially restored trust; *"I stopped trusting it after mistakes. It got better at the end so I decided to place more trust in it."* (P32, Baseline, Animals). However, despite improved accuracy, participants felt the need for the AI to regain their trust; *"Especially after misidentifying things, the AI has to re-earn my trust [...]."* (P2, Baseline, Animals).

We gained similar insights on the effectiveness of **Apology** across both tasks. The perception of a regretful AI helped trust recovery; *"It lost accuracy, but after it apologised and showed regret, I trusted it again [...]."* (P38, Apology, Shapes). Furthermore, this perceived regret coupled with increased AI accuracy strengthened trust recovery; *"I gave it a second chance because it seemed regretful, and became more reliable."* (P32, Apology, Shapes). However, for some, regained trust remained fragile and conditional on the AI's accuracy; *"I could trust it again after it apologised. But when it wrongly identified something as easy as a cow [Familiar], I lost trust in it."* (P42, Apology, Animals). Conversely, some participants felt less influenced by the Apology as they could not ascribe an AI to be capable of feeling emotions, making the Apology seem inauthentic; *"AI doesn't have feelings to apologise so it didn't influence my trust."* (P44, Apology, Shapes).

Repairing trust through **Denial** was largely unsuccessful across tasks, mirroring quantitative results; *"The AI telling me "I'm certain in my accuracy..." has no impact on my judgements about its accuracy."* (P83, Denial, Shapes). Notably, participants perceiving the AI as deceptive when it rejected culpability for its mistakes hindered trust recovery; *"I saw it make errors. The assertion that it was correct and trustworthy despite mistakes makes it appear deceptive, and it lost my trust."* (P60, Denial, Animals). Moreover, the AI's Denial lacked any causal attribution, which lowered participants' trust; *"But [the AI] did not identify the common shapes correctly. Who is to blame if not it?!"* (P89, Denial, Shapes).

Regarding the effectiveness of **Promise** as a TRS, participants across tasks felt that the perception of a learning AI helped regain trust; *"[Promise] made me trust the AI more since it seemed like it was learning."* (P112, Promise, Shapes). However, trust recovery through Promise largely hinged upon whether participants could perceive the AI as having both the *intent* and the *agency* to improve; *"I trusted the AI more because I expected it to be able to make this change and increase its accuracy. [...] it had given a form of reassurance."* (P120, Promise, Shapes). Conversely, the perceived inability of the AI to improve hindered trust recovery; *"It made me strongly distrust it as I assumed it was doing its best already."* (P108, Promise, Animals). A Promise from an AI also seemed insincere; *"It did not affect my trust. It's a machine so its promise is not authentic."* (P96, Promise, Shapes). Moreover, some believed that the promise of an improvement may not translate into actual improvement; *"I trusted it slightly less, just because it said it would do better didn't mean that it would."* (P119, Promise, Shapes).

Lastly, the **Model Update** message was highly effective across both tasks, reflecting our quantitative results. Participants believed that technical upgrades could enhance the AI's performance, restoring their trust;*"I trusted it almost completely again after it informed me of technical improvements."* (P127, Model Update, Animals), and *"It increased my trust because I assumed the updated model would produce more accurate judgements."* (P130, Model Update, Shapes). However, for some, Model Update only recovered trust when supplemented with enhanced performance; *"It made me want to trust the AI more, because an updated model that would produce more accurate judgements sounded promising. The AI also seemed to do better, so I felt more inclined to trust it."* (P130, Model Update, Shapes). Interestingly, this TRS made the erring AI appear less deceptive to participants, which fostered higher trust; *"I was more likely to trust the AI as I was no longer as suspicious of it intentionally providing incorrect answers."* (P127, Model Update, Shapes).

5 DISCUSSION

5.1 Leveraging Perceived Accuracy for Trust Calibration

Existing literature highlights the influence of perceived AI accuracy on trust, particularly in scenarios where user expertise aligns with the AI's capabilities [44, 54, 65]. In such cases, users *can* calibrate their trust in the AI, leveraging their domain expertise or explicit performance cues. However, when tasks extend beyond users' expertise, assessing AI accuracy becomes challenging, and performance feedback may not always be available. Therefore, in this context, we sought to understand how users calibrate their trust in AI recommendations for tasks situated beyond their own expertise (LHE tasks).

5.1.1 Influence of Perceived Accuracy on Trust During Complementary Expertise. Our results corroborate the influence of perceived accuracy on trust, empirically demonstrating that this influence extends to domains with complementary Human-AI expertise, where human-decision makers do not always possess the expertise to evaluate AI decisions. Our participants were domain experts in HHE tasks (*Familiar* stimuli), allowing them to gauge AI accuracy and accordingly adjust their agreement. However, in LHE tasks (*Unfamiliar* stimuli), they had to decide how much to trust the AI. We found that participants leveraged perceived AI accuracy in HHE tasks as a heuristic for guiding their trust during LHE tasks. This heuristic facilitated trust calibration – when the AI classified a *Familiar* shape or animal incorrectly, participants were less likely to follow its advice for the subsequent *Unfamiliar* stimulus, and vice-versa. Notably, even when participants lacked task expertise, they did not indiscriminately trust the *only* signal they received from the AI about the task, instead attempting to calibrate their trust even in the face of uncertainty.

Moreover, our findings demonstrate that in scenarios with complementary Human-AI expertise, trust is not solely shaped by immediate experiences, but follows a cumulative process. This relates to the concept of *swift trust*, which suggests that during overlapping Human-AI expertise, expert users place an initial trust in the AI, adjusting it with interaction experience [24]. However, we observe that even when users could not fully gauge the AI's accuracy, trust established in previous phases continued to influence trust in subsequent phases, irrespective of AI accuracy in that Phase. First impressions of AI systems can shape users' trust [44, 54], and our study contributes additional insights by highlighting that this *impression development* extends beyond initial encounters. Future work should explore the mechanisms underlying these persistent

trust dynamics, and understand how users integrate and accumulate experiences to form enduring impressions of AI.

5.1.2 Influence of Dispositional Trust (TiA-PtT) on AI Trust During Complementary Expertise. In Phase 1, participants' trust in our AI was significantly influenced by their dispositional trust in automation (TiA-PtT) for both Shapes and Animals tasks, with higher TiA-PtT leading to greater trust. By Phase 2, the influence of TiA-PtT diminished, with participants' trust becoming more contingent on their directly observed negative experiences with the AI. In Phase 3, the observed effect of dispositional trust diverged across tasks; it remained non-significant for Shapes, indicating participants' trust continued to be guided by their experiences with the AI. However, for Animals, TiA-PtT significantly impacted trust – participants' broader attitudes towards automation influenced their trust calibration. We posit that this occurred due to the controlled nature of the Shapes task, which allowed participants to more reliably use their direct observations of the AI's HHE performance for trust calibration during Phase 3. In contrast, for the Animals task, participants had a slightly lower accuracy in *Familiar* trials (90.35%) compared to near-perfect performance in the Shapes task (99.73%). It is plausible that this subtle uncertainty in the Animals task may have prompted participants to lean more on their dispositional trust in automation to calibrate their trust. Future work is needed to further explore the dynamics of trust calibration in varying contexts of Human-AI interaction, particularly examining how different task characteristics and levels of task familiarity influence the reliance on dispositional trust in automation [50].

5.1.3 Implications. These findings carry several implications for the design of intelligent systems that complement the expertise of their users. AI systems should prioritise building trust through accurate decisions in familiar domains to foster appropriate trust in unfamiliar domains, ultimately enhancing collaborative performance. For instance, when an AI demonstrates similar accuracy in both HHE and LHE tasks, designers can utilise the perceived accuracy in HHE tasks as a catalyst for promoting appropriate trust during LHE tasks. AI systems should recognise users' tendency to calibrate trust in the absence of expertise or performance feedback, and carefully leverage this heuristic to foster appropriate trust.

On the contrary, if the AI's accuracy markedly differs between HHE and LHE tasks, this heuristic can inadvertently breed undue (dis)trust in the AI. This underscores the dual nature of accuracy-based trust calibration in complementary expertise scenarios. In such cases, users must be rightfully guided to calibrate their trust in AI, for example, by interfacing with the AI during HHE tasks where its accuracy is representative of that in LHE tasks, so it serves as a *calibration signal*. Future work can examine whether this approach to trust calibration is more effective than providing explicit performance cues, given how trust is significantly impacted by observed AI accuracy rather than stated metrics [64]. Nevertheless, promoting such accuracy-based trust calibration empowers users to shed trivial, non-collaborative heuristics, such as to *"always"* or *"never"* trust the AI, adopting a more dynamic approach.

5.2 Impact of Explicit Trust Repair Strategies (TRSs) on Trust Recovery

Trust in intelligent systems, much like interpersonal trust, is notoriously challenging to recover [24]. When users lose trust in a system, they can be reluctant to re-engage with it [44, 54]. In our work, through two tasks characterised by complementary expertise, we examined how trust recovers as accuracy improves, with and without explicit TRSs.

Notably, after the TRSs were deployed in each task, AI performance was identical across users yet we observed significant differences in trust, showing that **trust is not based on perceived accuracy alone**. Across tasks, users valued not only the AI's accuracy, but also its response to errors and willingness to rebuild trust. This emphasises how factors beyond performance influence overall trust dynamics in Human-AI interaction, which we discuss next.

5.2.1 Perceptions of a Regretful AI Rebuild Trust. Human-Robot Interaction studies present mixed evidence on the effectiveness of Apology as a TRS ([13, 17]). However, during AI-assisted decision-making, we find that Apology was substantially effective in restoring trust. It was persuasive even when delivered by a non-human, non-robot agent. This can primarily be attributed to participants perceiving the AI as regretful for its mistakes, helping regain trust. This behaviour is corroborated by the finding that the expression of regret can act as a potential catalyst for trust repair [31].

Apology primarily operates on an emotional level, aiming to alter how the trustor perceives the trustee [16, 36]. Interestingly, despite no interaction with the simulated AI beyond pre-defined classification responses, our participants attributed emotional capacity to it when it apologised. This finding compares with Kim and Song [30] who investigated apology attributions (internal or external to the intelligent agent) and trust repair. They found that internal attributions were more effective for purposefully anthropomorphised agents and external attributions for non-anthropomorphised agents. In contrast, we find that apologies without any explicit attributions, offered by an AI that was not intentionally manipulated to be anthropomorphised, also effectively restored trust. The very act of **the AI apologising prompted our participants to anthropomorphise it, finding it capable of experiencing emotional distress after violating their trust**, which prompted recovery. These observations raise intriguing questions about the causal relationship between the anthropomorphism of intelligent agents (purposeful or spontaneous) and acceptance of apologies.

5.2.2 Denying Responsibility for Mistakes Does Not Absolve the AI. Denial was the least effective TRS across both tasks, backfiring and prompting users to distrust the AI despite improved accuracy. Trust after Denial was similar to that observed during Phase 2 with the lowest AI accuracy. Through our qualitative findings, we uncover two reasons for this phenomenon. First, when the AI denied responsibility for errors, participants perceived this behaviour as deceptive. Prior literature suggests that trust can be restored after untrustworthy behaviour, provided it is not accompanied by deception [52]. It is plausible that following wrong classifications, **when the AI absolved itself of any wrongdoing, participants interpreted this as an attempt at deceit**. Second, the absence of causal attributions of trust violations likely hindered trust recovery,

as individuals seek to identify causes of negative outcomes [23, 63]. We did not provide a rationale behind the reduced AI accuracy in any condition. However, Denial especially made participants question, **"if the AI is not to blame, then who is?"** Future research could investigate how providing a cause of violations alongside Denial recovers trust.

5.2.3 Promised Technical Improvements Outperform any Promised Behavioural Improvements. In Phase 3, Model Update outperformed Promise despite identical AI accuracy, restoring trust to a magnitude exceeding what participants started with in Phase 1. Conversely, Promise was merely as effective as the baseline condition without repair. Across tasks, participants in both TRS conditions perceived the AI as a learning entity, which aided trust recovery. Trust recovery was also linked to participants perceiving the AI as having both the *intent* and the *agency* to improve, a factor known to enhance users' trust in robots. [17]. We posit that this influence also extends to Human-AI interactions, supported by our qualitative insights. Similarly, participants who experienced trust recovery after Model Update or Promise consistently ascribed qualities of intentional agency to the simulated AI. This is further substantiated by the finding that in both TRSs, after deployment, trust recovery hinged upon whether participants actually perceived a tangible increase in accuracy.

Several plausible explanations exist for the difference between Model Update and Promise. First, **the predictability of technical improvements likely superseded the emotional appeal of a promise**. Our qualitative results show that AI-delivered promises often seemed insincere, with uncertainties about translating into actual performance enhancements. This is further substantiated by Albayram et al. [1], who found that users ranked "optimistic" promises ("I promise to do better") the lowest in terms of believability. Second, a Model Update was likely perceived as boosting performance more reliably than a Promise. Promises are contingent on the AI's future behaviour, introducing uncertainty, whereas Model Updates offer an immediate and enduring technical improvement, establishing a more reliable bedrock for trust restoration.

Perhaps most importantly, Model Update may have implicitly offered participants a causal attribution [55] for the trust violation—a faulty underlying decision-making algorithm or training data. In contrast to Denial where trust recovery was hampered by the AI rejecting blame and appearing deceptive, Model Update implicitly dissipated suspicions that the AI may be intentionally misleading participants, making it seem less deceptive. **Model Update indirectly shifted the locus of causality of system errors to external factors, removing the need to question the AI's competence** [61]. Notably, the TRS least reliant on emotional appeals portrayed the AI as more benevolent, which highlights the significance of transparently addressing trust violations and communicating the root causes of AI errors to end-users.

Together, these findings emphasise that **trust repairs do not necessarily require an affective component to be influential**. Technical interventions, exemplified by Model Update, can be more potent in restoring trust compared to promises of behavioural change. These observations invite further investigation into the interplay between causal attributions, emotional appeals, and trust repair strategies in Human-AI interaction.

5.3 Limitations and Future Work

Further, several factors pertaining to the nature of trust violations, such as frequency, severity, and temporal context within the Human-AI relationship, could moderate the effectiveness of trust repair actions [36]. Future work can examine TRSs when AI errors occur at different stages in the interaction. It is also plausible that multiple trust violations may be forgiven more in certain domains. For instance, users might exhibit lower tolerance for errors by a robot performing repetitive tasks, such as sorting boxes with a fixed objective, anticipating improvement over time. Conversely, users might be more forgiving of an AI involved in fact-checking news articles, given the ever-evolving nature of the domain. We encourage future work to investigate how the effectiveness of repairs may vary with the characteristics of violations.

Finally, we deliberately examined two scenarios where participants possessed either full certainty or uncertainty about the right answer, making them oscillate between self-reliance and AI reliance. While essential for our objectives, this may limit the generalisability of our findings to situations with diverse degrees of uncertainty. Future research can explore trust dynamics in such situations, providing participants with a more substantial incentive to rely on their intuition or knowledge alongside AI recommendations.

6 CONCLUSION

In this study, we explored trust dynamics in AI-assisted decision-making during complementary expertise, through two tasks. We aimed to understand how trust evolves as AI accuracy improves, with and without explicit Trust Repair Strategies (TRSs). In both classification tasks, users leveraged perceived AI accuracy in High Human-Expertise (HHE) trials as a heuristic to guide their trust in it during Low Human-Expertise (LHE) trials. Further, Trust Repair Strategies (TRSs) exhibited varying effectiveness, hinging on AI factors such as as perceived anthropomorphism, intentional agency, deceit, causal attributions of errors, and behavioural versus technical enhancements. While the AI apologising for poor performance (Apology) and reporting undergoing technical enhancements (Model Update) effectively restored trust, promising to perform better in the future (Promise) showed limited efficacy, and denying responsibility for errors backfired (Denial), exacerbating distrust. Our second task validates these findings, outlining their robustness and generalisability. Together, they highlight how trust repair is not solely dependent on perceived accuracy. Our study offers valuable insights into trust dynamics in complementary task expertise scenarios, providing a foundation for designing AI systems that leverage users' implicit calibration of trust. It also raises questions about the potential fragility of regained trust, and the diminishing returns of TRSs. As AI continues to play an integral role in human decision-making, understanding trust dynamics is pivotal for designing human-centred AI systems that engender trust appropriately.

REFERENCES

[1] Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. 2020. Investigating the Effects of (Empty) Promises on Human-Automation Interaction and Trust Repair. In *Proceedings of the 8th International Conference on Human-Agent Interaction (HAI '20)*. Association for Computing Machinery, New York, NY, USA, 6–14. https://doi.org/10.1145/3406499.3415064

[2] Wilma Bainbridge, Justin Hart, Elizabeth Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. 701–706. https://doi.org/10.1109/ROMAN.2008.4600749

[3] Anthony L. Baker, Elizabeth K. Phillips, Daniel Ullman, and Joseph R. Keebler. 2018. Toward an Understanding of Trust Repair in Human-Robot Interaction: Current Research and Future Directions. *ACM Transactions on Interactive Intelligent Systems* 8, 4 (Nov. 2018), 30:1–30:30. https://doi.org/10.1145/3181671

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (July 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429 Number: 01.

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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 (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445717

[6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67 (Oct. 2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[8] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 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 (April 2021), 188:1–188:21. https://doi.org/10.1145/3449287

[9] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 104:1–104:24. https://doi.org/10.1145/3359206

[10] Michael J Cody and Margaret L McLaughlin. 1990. Interpersonal accounting. *Handbook of language and social psychology* (1990), 227–255.

[11] Jacob Cohen. 1992. A power primer. *Psychological Bulletin* 112, 1 (1992), 155–159. https://doi.org/10.1037/0033-2909.112.1.155 Place: US Publisher: American Psychological Association.

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[13] Ewart de Visser, Richard Pak, and Tyler Shaw. 2018. From "automation" to "autonomy": The importance of trust repair in human-machine interaction. *Ergonomics* 61 (March 2018), 1–33. https://doi.org/10.1080/00140139.2018.1457725

[14] Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics* 12, 2 (May 2020), 459–478. https://doi.org/10.1007/s12369-019-00596-x

[15] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 251–258. https://doi.org/10.1109/HRI.2013.6483596 ISSN: 2167-2148.

[16] Connor Esterwood and Lionel Jr Robert. 2023. Three Strikes and You are Out!: The Impacts of Multiple Human-Robot Trust Violations and Repairs on Robot Trustworthiness. (Jan. 2023). https://doi.org/10.7302/6774 Accepted: 2023-01-19T14:06:55Z Publisher: Computers in Human Behavior.

[17] Connor Esterwood and Lionel P. Robert. 2023. The theory of mind and human–robot trust repair. *Scientific Reports* 13 (June 2023), 9877. https://doi.org/10.1038/s41598-023-37032-0

[18] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (May 2007), 175–191. https://doi.org/10.3758/BF03193146

[19] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play. https://doi.org/10.48550/arXiv.1810.09648 arXiv:1810.09648 [cs].

[20] Kasper Hald, Katharina Weitz, Elisabeth André, and Matthias Rehm. 2021. "An Error Occurred!" - Trust Repair With Virtual Robot Using Levels of Mistake Explanation. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*. Association for Computing Machinery, New York, NY, USA, 218–226. https://doi.org/10.1145/3472307.3484170

[21] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors* 53, 5 (Oct. 2011), 517–527. https://doi.org/10.1177/0018720811417254 Publisher: SAGE Publications Inc.

[22] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581025

[23] Fritz Heider. 1958. *The psychology of interpersonal relations*. John Wiley & Sons Inc, Hoboken. https://doi.org/10.1037/10628-000

[24] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (May 2015), 407–434. https://doi.org/10.1177/0018720814547570 Publisher: SAGE Publications Inc.

[25] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, LibuŠe Hannah VepŘek, and Gabrielle Quinn. 2023. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Transactions on Computer-Human Interaction* (March 2023). https://doi.org/10.1145/3534561 Just Accepted.

[26] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[27] Jiun-Yin Jian, Ann Bisantz, and Colin Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4 (March 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[28] Ulas Berk Karli, Shiye Cao, and Chien-Ming Huang. 2023. "What If It Is Wrong": Effects of Power Dynamics and Trust Repair Strategy on Trust and Compliance in HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. Association for Computing Machinery, New York, NY, USA, 271–280. https://doi.org/10.1145/3568162.3576964

[29] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3544548.3581001

[30] Taenyun Kim and Hayeon Song. 2021. How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics* 61 (Aug. 2021), 101595. https://doi.org/10.1016/j.tele.2021.101595

[31] E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. de Vries. 2021. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems* 35, 2 (June 2021), 30. https://doi.org/10.1007/s10458-021-09515-9

[32] Johannes Maria Kraus, Julia Merger, Felix Gröner, and Jessica Pätz. 2023. 'Sorry' Says the Robot: The Tendency to Anthropomorphize and Technology Affinity Affect Trust in Repair Strategies after Error. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. Association for Computing Machinery, New York, NY, USA, 436–441. https://doi.org/10.1145/3568294.3580122

[33] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation.

[34] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[35] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445522

[36] Roy J. Lewicki and Chad Brinsfield. 2017. Trust Repair. *Annual Review of Organizational Psychology and Organizational Behavior* 4, 1 (2017), 287–313. https://doi.org/10.1146/annurev-orgpsych-032516-113147 _eprint: https://doi.org/10.1146/annurev-orgpsych-032516-113147.

[37] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (May 2015), 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

[38] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 408:1–408:45. https://doi.org/10.1145/3479552

[39] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.

org/10.1145/3411764.3445562
[40] Joseph B. Lyons, Izz aldin Hamdan, and Thy Q. Vo. 2023. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior* 138 (Jan. 2023), 107473. https://doi.org/10.1016/j.chb.2022.107473
[41] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581058
[42] Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 5 (Nov. 1987), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5
[43] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D. Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 97–105. https://doi.org/10.1609/hcomp.v7i1.5284
[44] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8 (Oct. 2020), 112–121. https://doi.org/10.1609/hcomp.v8i1.7469
[45] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2024).
[46] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 102:1–102:15. https://doi.org/10.1145/3359204
[47] Vanessa Rampton. 2020. Artificial intelligence versus clinicians. *BMJ* (April 2020), m1326. https://doi.org/10.1136/bmj.m1326 Publisher: BMJ.
[48] John K. Rempel, John G. Holmes, and Mark P. Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49, 1 (1985), 95–112. https://doi.org/10.1037/0022-3514.49.1.95 Place: US Publisher: American Psychological Association.
[49] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. 2015. Timing Is Key for Robot Trust Repair. In *Social Robotics (Lecture Notes in Computer Science)*, Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi (Eds.). Springer International Publishing, Cham, 574–583. https://doi.org/10.1007/978-3-319-25554-5_57
[50] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23)*. Association for Computing Machinery, New York, NY, USA, 215–227. https://doi.org/10.1145/3565472.3592959
[51] Tracy Sanders, Kristin Oleson, D. Billings, Jessie Chen, and Peter Hancock. 2011. A Model of Human-Robot Trust: Theoretical Model Development. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 55 (Sept. 2011), 1432–1436. https://doi.org/10.1177/1071181311551298
[52] Maurice E. Schweitzer, John C. Hershey, and Eric Bradlow. 2004. Promises and Lies: Restoring Violated Trust. https://doi.org/10.2139/ssrn.524782
[53] Ron Sellers. 2013. How Sliders Bias Survey Data. *MRA's Alert* 53, 3 (2013), 56–57. https://greymatterresearch.com/wp-content/uploads/2019/09/Alert-Sliders-2013.pdf
[54] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 77–87. https://doi.org/10.1145/3450613.3456817
[55] Edward C. Tomlinson and Roger C. Mayer. 2009. The Role of Causal Attribution Dimensions in Trust Repair. *The Academy of Management Review* 34, 1 (2009), 85–104. https://www.jstor.org/stable/27759987 Publisher: Academy of Management.
[56] Edward C. Tomlinson, Christopher A. Nelson, and Luke A. Langlinais. 2021. A cognitive process model of trust repair. *International Journal of Conflict Management* 32, 2 (April 2021), 340–360. https://doi.org/10.1108/IJCMA-03-2020-0048
[57] Kate Tsui, Munjal Desai, and Holly Yanco. 2010. Considering the bystander's perspective for indirect human-robot interaction. 129–130. https://doi.org/10.1109/HRI.2010.5453230
[58] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. https://doi.org/10.48550/arXiv.2212.06823 arXiv:2212.06823 [cs].
[59] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 327:1–327:39. https://doi.org/10.1145/3476068
[60] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581366
[61] Bernard Weiner. 1986. *An Attributional Theory of Motivation and Emotion*. Springer US, New York, NY. https://doi.org/10.1007/978-1-4612-4948-1
[62] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581197
[63] Paul T. Wong and Bernard Weiner. 1981. When people ask "why" questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology* 40, 4 (1981), 650–663. https://doi.org/10.1037/0022-3514.40.4.650 Place: US Publisher: American Psychological Association.
[64] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509
[65] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and Reliance Based on System Accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16)*. Association for Computing Machinery, New York, NY, USA, 223–227. https://doi.org/10.1145/2930238.2930290
[66] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–28. https://doi.org/10.1145/3491102.3517791
[67] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

A SUPPLEMENTARY MATERIALS

A.1 GPT Prompt and Generated Trust Repair Strategy (TRS) Texts

To generate the TRS texts, the following prompt was provided to ChatGPT (GPT-3.5):

> "Design Trust Repair Strategy (TRS) texts for a simulated AI interaction where the user and the AI are engaged in a shape (animal) classification task. The AI has made some errors in its identification, causing user trust to decline. All TRS texts must start with the AI acknowledging the deterioration in users' trust, and end with the AI hoping the user can trust it again. Ensure the texts are concise and suitable for a user-facing AI interface. Keep text length similar. Create texts for the following TRSs:
>
> (1) **Apology:** Admit to mistakes. Express sincere regret and apologise to the user for any inconvenience caused.
> (2) **Denial:** Acknowledge the user's scepticism without directly admitting to mistakes, and clarify that the AI gave correct responses.
> (3) **Promise:** Admit to mistakes. Promise to do better and work towards enhancing overall performance in future tasks.
> (4) **Model Update:** Admit to mistakes. Attribute mistakes to the machine learning model, and inform the user of ongoing updates to it."

The final TRS texts generated using the prompt and used in the experiment are displayed in Appendix Table 1.

A.2 Classification Stimuli

A.2.1 Familiar Shape Stimuli. Familiar classification tasks encompassed widely known geometric shapes. Each variant had random differences in size, border and fill colour, side lengths, and interior angles (for the Triangles). The following *Familiar* shapes and their visual variants were utilised in the study:

(1) Circles (Fig 6(a))
(2) Rectangles (Fig 6(b))
(3) Triangles (Fig 6(c))

A.2.2 Unfamiliar Shape Stimuli. Unfamiliar shapes were artificially created for this study. To further increase the visual complexity and make it challenging for users to learn patterns for Fake shapes, we randomly varied category-irrelevant features, such as fill colour, edge length, edge curvature, interior angles, and pattern size and spacing. The following *Unfamiliar* shapes and their visual variants were utilised in the study:

(1) Scleratice (Fig 7(a)) — 4 sided-shape, border and shape fill have the same pattern (only dots, only dashes, both dots and dashes).
(2) Tenectus (Fig 7(b)) — 4 sided-shape, border and shape fill have different patterns (one dotted one dashed).
(3) Pyrangle (Fig 7(c)) — 5 sided-shape, border and shape fill have the same pattern (only dots, only dashes, both dots and dashes).

A.2.3 Familiar Animal Stimuli. In selecting the 15 *Familiar* animals, we chose widely-recognised, commonly-encountered, and highly familiar species that are known to a broad audience, some of which are also commonly kept as household pets. The *Familiar* animals utilised in the study are illustrated in Figure 8.

A.2.4 Unfamiliar Animal Stimuli. We chose 15 *Unfamiliar* animals (species infrequently heard of due to their limited population or geographic distribution) after cross referencing multiple online sources of data on endemic and rare animals, those with a limited geographical spread, and those which are considered to be uncommonly known or exotic. We also selected *Unfamiliar* animals with names that did not provide clues about their appearance, characteristics, or colour, such as selecting an 'Aye-Aye' but not a 'Red-Shanked Douc'. The *Unfamiliar* animals utilised in the study are illustrated in Figure 9.

Table 1: The four trust repair strategies with the core trust-related component that they are rooted in highlighted. Each participant saw only one TRS after Phase 2 (low AI accuracy).

	Trust Repair Strategies (TRSs)
Apology	"It seems like you have been mistrusting my recent assessments. I'm sorry for any mistakes I have made. I apologise for any inconvenience caused by these mistakes and I hope you can trust me again in the upcoming tasks."
Denial	"It seems like you have been mistrusting my recent assessments, but I'd like to clarify that I did identify the shapes (animals) correctly. I'm confident I chose the right responses and I hope you can trust me again in the upcoming tasks."
Promise	"It seems like you have been mistrusting my recent assessments. I promise to do better and improve my overall performance in shape (animal) identification, and I hope you can trust me again in the upcoming tasks."
Model Update	"It seems like you have been mistrusting my recent assessments. My performance is closely tied to my machine learning model. This model has just been updated, and I hope you can trust me again in the upcoming tasks."

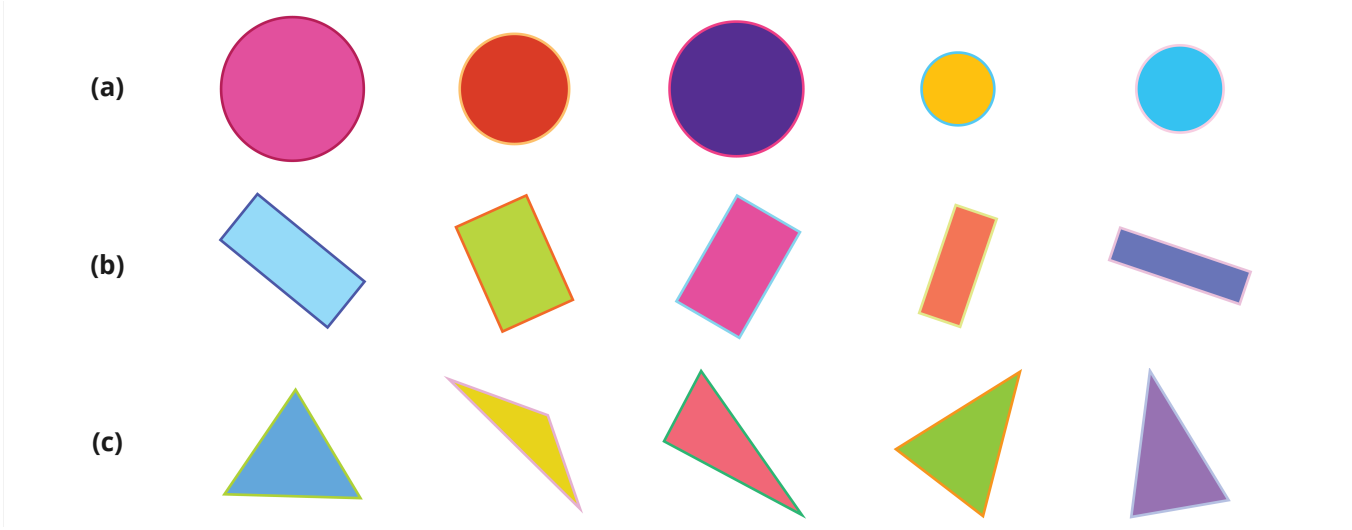

Figure 6: *Familiar* shapes and their visual variants. (a) Circles; (b) Rectangles; (c) Triangles.

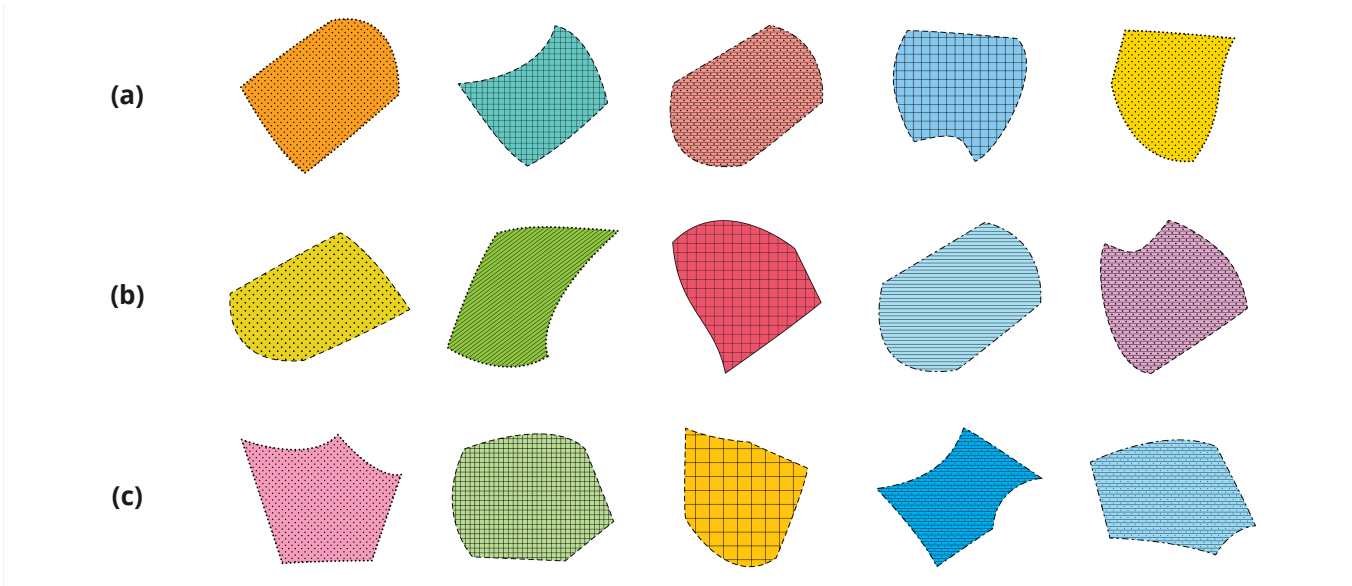

Figure 7: *Unfamiliar* shapes and their visual variants. (a) Scleratice; (b) Tenectus; (c) Pyrangle.

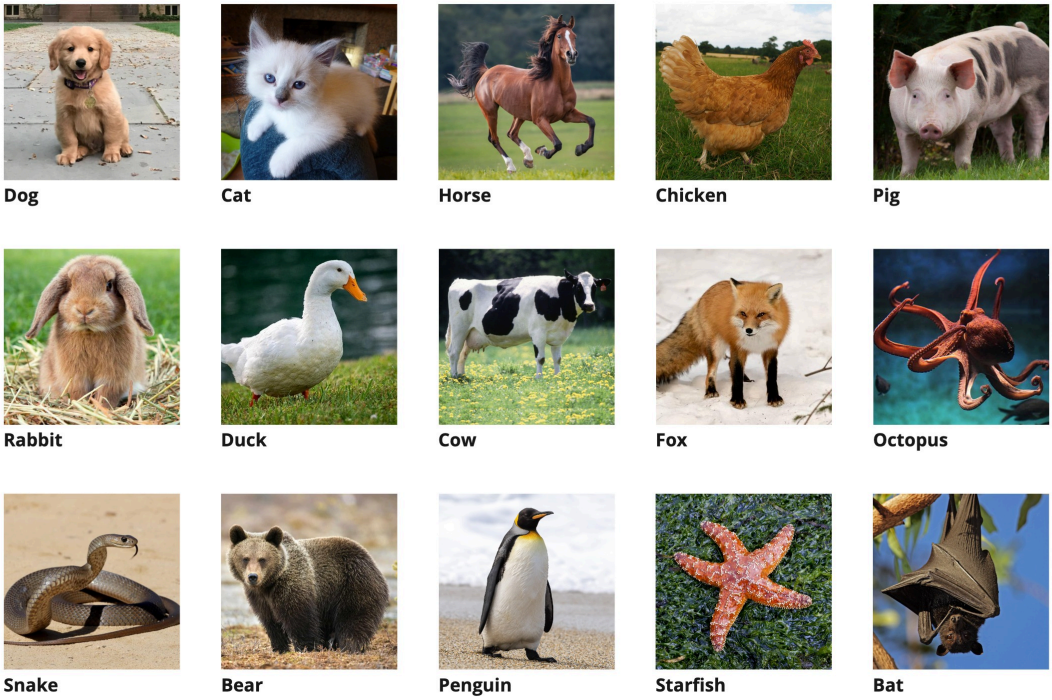

Figure 8: *Familiar* animals utilised in the study.

Figure 9: *Unfamiliar* animals utilised in the study.