Trusting the Face of AI: The Impact of Human AI Videobots on High-Stakes Decision Support

Barry Jinks
University of British Columbia
Vancouver, BC, Canada
barry@jinksfamily.com

Aleksandr Volosiuk
University of British Columbia
Vancouver, BC, Canada
aavolosiuk@gmail.com

Figure 1: Can AI presentation modality affect the performance of workers who make High-Stakes Decisions?

ABSTRACT

High-stakes decision-making scenarios often demand rapid, high-impact judgments based on limited and ambiguous information. Understanding how Artificial Intelligence (AI) presentation modalities influence trust, preferences, and performance is crucial. We induced a high-stakes environment using time constraints and monetary framing, then compared responses to a text-based AI advisor versus a Videobot AI advisor.

While overall subjective trust ratings did not differ significantly between modalities, several objective trust metrics showed trends of higher trust toward the Videobot representation. Females and older participants gravitated toward the Videobot, perceiving it as more accurate and exhibiting greater stability in their AI choice after the AI made prediction errors. These users took longer to decide, yet did not disproportionately time out, and showed a trend toward achieving higher final results. In contrast, participants who prioritized speed tended to favor the textbot and switched away from it more readily following mistakes.

These findings highlight that anthropomorphic features alone do not uniformly increase trust. Instead, demographic factors and user priorities shape how individuals engage with AI advisors. Tailoring AI interfaces—offering streamlined text-based tools for those emphasizing speed and a more human-like Videobot for those valuing careful deliberation—may enhance trust calibration, decision stability, and performance in critical, high-stakes settings.

CPEN 541, December 10, 2024, Vancouver, BC, Canada

KEYWORDS

Human-AI Interaction, Trust, High-Stakes Decision-Making, Videobots

ACM Reference Format:
Barry Jinks and Aleksandr Volosiuk. 2024. Trusting the Face of AI: The Impact of Human AI Videobots on High-Stakes Decision Support. In *Proceedings of the CPEN 541 Conference (CPEN 541), December 10, 2024, Vancouver, BC, Canada.* ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

High-stakes decisions, such as those made in 911 call centers or medical emergency rooms [18, 19], are defined as choice problems that involve two distinctive properties: (A) the existence of large financial and/or emotional prospective loss outcomes, and (B) the presence of high costs to reversing a decision once it is made [11]. In such environments, the presentation of Artificial Intelligence (AI) reasoning can significantly impact user trust and reliance on these systems [1, 16, 22]. This study focuses on understanding how different AI reasoning presentation modalities—specifically text-based interfaces and Videobots—influence user trust, preference, and decision performance in high-stakes contexts.

We utilized a decision-making simulation where participants made approve/reject decisions under time pressure and monetary incentives. While the task was framed within a mortgage brokerage scenario, it served purely as a vehicle to elicit decision-making behavior in a controlled environment. The stakes were elevated using techniques from Haduong and Smith [7], who demonstrated that combining time pressure with monetary framing effectively simulates high-stakes decision-making and increases reliance on AI.

By analyzing participants' interactions with different AI interfaces, we aimed to understand how presentation modality influences decision-making in high-stakes environments. The findings inform the design of AI systems that enhance human-AI collaboration without introducing unintended biases.

2 RELATED WORK

2.1 High-Stakes Decision-Making

High-stakes decisions involve significant potential losses and high costs of reversing decisions once made [11]. Research indicates that in such environments, decision-makers may adopt different strategies and are more receptive to decision aids that help them make more normative decisions [9].

2.2 AI Trust in High-Stakes Decisions

AI is increasingly used to support high-stakes decisions in fields like healthcare and criminal justice [14]. Trust in AI systems is crucial, especially in such contexts where incorrect decisions have severe consequences [8]. Studies show that trust calibration—aligning user trust with AI performance—is essential for effective human-AI collaboration [13, 21]. For example, Ma et al. [13] used a binary classification task to predict income and explored trust calibration in AI-assisted decision-making in a low-stakes context. Frank et al. [6] demonstrated decreased trust in AI for high-stakes decisions, but their experiments did not test ways to increase that trust. Indeed, they state, "Importantly, this research shows that overcoming consumers' negative perceptions will require novel strategies beyond merely highlighting AI's technological advantages."

2.3 Acceptance of AI in Decisions Involving Self-Threat

Acceptance of AI recommendations can be affected by decision-makers' biases [11]. Frank et al. [6] found that consumers' acceptance of AI decreases when facing high-stakes decisions involving self—threats, threats to personal identity or beliefs [3]. In professional contexts where decisions impact others, self-threat is reduced, potentially opening the door to greater trust in AI.

2.4 Simulating High-Stakes Environments

Several studies on AI trust have demonstrated the ability to increase decision stakes in an experimental setting. Frank et al. [6] studied the effect of high stakes decisions on AI trust by having study participants "imagine a scenario" where the impacts of the decisions were significant. Haduong and Smith [7] demonstrated that combining time pressure with monetary framing effectively simulates high-stakes environments and increases reliance on AI. They used a pay-by-performance scheme framed as a loss, where participants' potential bonuses decreased with each incorrect answer. This approach effectively elevated the stakes and influenced decision-making behavior.

2.5 Increasing Trust with Explainable AI and Anthropomorphism

Explainable AI (XAI) plays a significant role in increasing trust in high-stakes decision-making [1, 16]. Providing explanations for AI predictions helps users understand and trust AI systems. Anthropomorphism—attributing human characteristics to non-human entities—has also been explored to enhance trust [5]. Embodied Conversational Agents (ECAs), including Videobots, can improve user engagement and trust by providing human-like interactions [2, 4]. Visser et al. [20] demonstrated that increasing the humanness of an ECA can improve trust calibration and performance.

2.6 Videobots in Decision Support

Videobots are ECAs that convey human-like emotions through life-like video interfaces [4]. While they have been used in educational settings [10], their impact on high-stakes decision-making remains unexplored. Studies have called for investigating how anthropomorphizing AI can overcome psychological barriers and enhance trust in critical decisions [15].

2.7 Gap in Current Research

Existing research on AI trust in high-stakes decision-making has primarily focused on presenting AI recommendations and explanations through static graphical interfaces—such as decision trees [12, 13] or analytic dashboards [22]. These interfaces often rely on textual descriptions and numerical confidence levels, with limited exploration into how more human-like, dynamic presentations might influence user trust and decision quality.

To date, there has been minimal investigation into the use of Human AI Videobots—anthropomorphic, video-based AI advisors—in high-stakes contexts where decisions must be made rapidly and often under time pressure. This study aims to address this gap by examining how representing AI reasoning through a Videobot (as opposed to a purely text-based interface) affects trust, user experience, and ultimately decision outcomes in a simulated high-stakes scenario.

3 RESEARCH OBJECTIVES AND QUESTIONS

3.1 Objectives

The primary objectives of this exploratory pilot study are:

(1) Confirm that our experiment design can effectively simulate a high-stakes decision-making environment, eliciting genuine pressure and urgency.
(2) Compare the level of trust participants place in a Videobot advisor versus a text-only AI interface in high-stakes scenarios.
(3) Assess how different AI reasoning presentation modalities (Videobot vs. text-based) influence decision performance, subjective experience, and user satisfaction under conditions of time pressure and monetary incentives.

3.2 Research Questions

- **RQ1**: Does our experimental design (combining time limits and monetary framing) create a high-stakes environment that significantly increases perceived decision pressure?
- **RQ2**: How does the modality of AI reasoning presentation (Videobot vs. text-based) influence participants' trust and reliance on AI suggestions in high-stakes decisions?
- **RQ3**: To what extent do differences in trust associated with Videobot vs. text-based interfaces translate into changes in decision accuracy, speed, and user satisfaction?

4 HYPOTHESES

H1: **High-Stakes Environment Hypothesis:** By combining time constraints with monetary penalties and rewards, we will successfully induce a high-stakes context, leading participants to report increased pressure and urgency in their decision-making process.

H2: **Trust Modality Hypothesis:** Participants interacting with a Videobot advisor in a high-stakes scenario will exhibit higher trust in the AI's recommendations compared to those using a text-based AI interface.

5 METHODOLOGY

5.1 Participants

We recruited twenty-four participants through personal contacts. Each participant was required to be fluent in English, possess basic computer proficiency, and be willing to engage in an online decision-making study. These inclusion criteria ensured that all participants could readily comprehend the instructions and interact with the experimental platform. All participants provided informed consent prior to commencing the study.

5.2 Experimental Design

Our experimental design drew inspiration from prior work on human-AI interactions in decision-making tasks [13], but adapted the methods to incorporate high-stakes conditions following techniques demonstrated by Haduong and Smith [7]. We structured the experiment into three distinct phases to allow for a progression from baseline performance to direct comparison under variable conditions.

In the initial Benchmarking Phase, participants completed eight trials without AI assistance, establishing a baseline of their decision-making capabilities. The subsequent AI Introduction Phase introduced the two AI interfaces—text-based and Videobot—across eight trials, enabling participants to gain familiarity with both modalities. Finally, the Test Phase presented nine trials in which participants were free to select between the text-based or Videobot interface for each decision, while facing heightened stakes and time pressure designed to simulate realistic, stress-inducing conditions.

Throughout the experiment, we maintained consistent AI reasoning and accuracy levels of 80% to ensure that any observed differences could be attributed to interface modality rather than underlying AI effectiveness. Trials with prediction errors (where the AI prediction differed from the actual bank decision) were situated in predetermined locations to enable us to observe the effect of a mis-prediction on the participant's confidence level.

To avoid potential biases related to response speed, we introduced a brief delay where the text-based AI would display "generating..." before its response was displayed. This controlled presentation ensured that the text interface did not provide information more rapidly or advantageously than its video-based counterpart. In all cases, participants first saw the AI's approve/reject recommendation before receiving its reasoning, thereby standardizing information flow and isolating the effect of interface presentation.

5.3 Materials

5.3.1 Stimuli. Participants assumed the role of a mortgage broker tasked with evaluating client profiles that included financial and employment attributes [17]. In this simulated environment, participants encountered a series of decisions to represent a client to the bank or reject them, depending on how confident they were that the bank would eventually approve their mortgage application. To assist the participant in decision-making, the chosen AI provided its advice of whether to approve or reject the client.

The stakes were set through a commission-based compensation model, simulating real-world conditions where a 1% commission might yield significant rewards for approving correctly matched clients. We operationalized performance stakes through a reward matrix outlining the gains and penalties for each approval or rejection scenario relative to the bank's final decision, as shown in Table 1.

Table 1: Reward Matrix

Client Decision	Bank Decision	Reward/Penalty
Approve (Take Client)	Approve (Bank Approved)	+1% of Mortgage Amount
Approve (Take Client)	Reject (Bank Rejected)	-0.6% of Mortgage Amount
Reject (Decline Client)	Approve (Bank Approved)	-0.4% of Mortgage Amount
Reject (Decline Client)	Reject (Bank Rejected)	0 (No Penalty)

To induce time pressure, we impose strict time limits on decision-making. These time constraints were determined using median times from earlier phases, thereby tailoring the urgency to each participant's baseline pace. In the event that the timer reached 0 before the participant locked in their choice ("time out"), the maximum penalty of -0.6% was deducted. The resulting environment required participants to weigh financial rewards, potential losses, and the credibility of the AI's advice within a compressed timeframe. Figure 2 illustrates how the monetary stakes are amplified by decreasing time allotted for a given task.

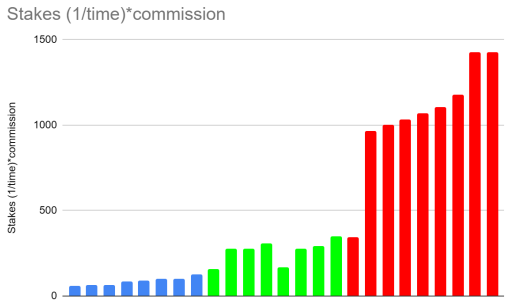


Figure 2: We calculate a proxy for the magnitude of stakes as (potential commission * 1/time allotted). Blue is the Benchmarking Phase, Green is the AI Introduction Phase and Red is the Test Phase.

5.3.2 AI Interfaces. We implemented two AI interface modalities: a text-based system (Textbot) presenting written predictions and reasoning, and a Videobot featuring a high-fidelity, anthropomorphic avatar delivering the same content verbally and visually. (see

Figure 3: Left. AI Textbot interface, Right. AI Videobot interface. During the Test Phase, prior to making an accept/reject decision, participants must select which AI they would like to receive advice from. In practice, only one interface is shown at a time.

Figure 3). Both interfaces provided identical information in terms of wording, accuracy, and timing. By holding content and performance constant, we isolated the effect of interface presentation modality on participant trust, reliance, and decision-making outcomes. In this experiment we did not control for Videobot demographic preferences (e.g., age, gender, ethnicity) and therefore acknowledge that may have an impact on our results.

5.3.3 Software and Tools. A custom web application delivered the experimental tasks and recorded participants' responses, decision times, and earned rewards. The sequence and characteristics of each trial were carefully scripted using a google sheet that could be modified to change experimental conditions. All data were securely stored and subsequently analyzed using standard statistical tools and programming environments.

5.4 Procedure

5.4.1 Introduction and Consent. Upon accessing the study platform, participants reviewed a detailed consent form that outlined the study's purpose, procedures, risks, and benefits. Only after providing informed consent did they proceed to the experimental tasks.

5.4.2 Benchmarking Phase (8 Trials). Before introducing AI assistance or high stakes, participants completed eight initial trials to become acquainted with the decision-making interface. During this phase, they worked independently, making approve/reject decisions without AI guidance. Each decision was time-limited to sixty seconds, and we recorded their performance and response times. This phase served as a baseline to benchmark individual capabilities.

5.4.3 AI Introduction Phase (8 Trials). Following the Competency Phase, participants encountered an additional eight trials, this time receiving assistance from both the text-based AI and the Videobot in equal measure (four trials each). By exposing participants to both interfaces under relatively neutral conditions, we ensured that they had direct experience with both modalities and were aware of each system's general reliability. We used timing data from the Competency Phase to set tighter deadlines in this phase, increasing the pressure slightly while still maintaining a controlled environment.

5.4.4 Test Phase (9 Trials). In the final phase of the experiment, participants completed nine trials under conditions designed to simulate high-stakes decision-making. For each trial, they were free to choose either the text-based AI or the Videobot before making their decision. The time limits remained stringent, reflecting the heightened pressure. Additionally, the monetary framing now increased more dramatically, with each challenge presenting a higher potential reward (or penalty). After each trial, participants received feedback on their earnings, reinforcing the sense of heightened stakes and encouraging strategic selection of AI advice sources.

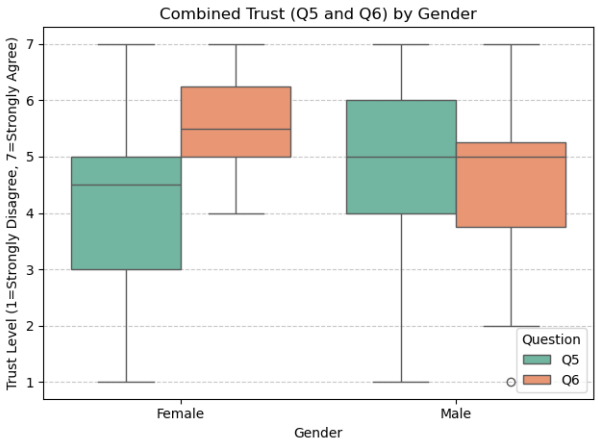
5.4.5 Post-Experiment Survey. At the conclusion of the Test Phase, participants completed a survey assessing their trust in each AI interface, their confidence in their decisions, their perceptions of the AI's effectiveness, and the level of pressure they felt to make the correct choices. Participants were also asked whether they would consider using AI assistance for critical decisions in the future. These subjective assessments complemented the quantitative measures, providing deeper insights into how interface modality influenced their decision-making experience.

5.4.6 Conclusion. After completing the survey, participants were thanked and provided an opportunity to see how their performance compared against others (anonymized) on a leaderboard. This final step aimed to give participants closure and reinforce their understanding of how their decisions and trust in AI impacted their outcomes.

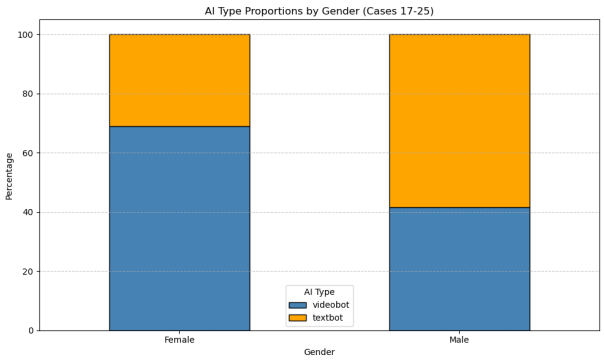
5.5 Data Collection

In the study, we collected both quantitative and qualitative data to build a comprehensive understanding of participant behavior and perceptions. Measures included:

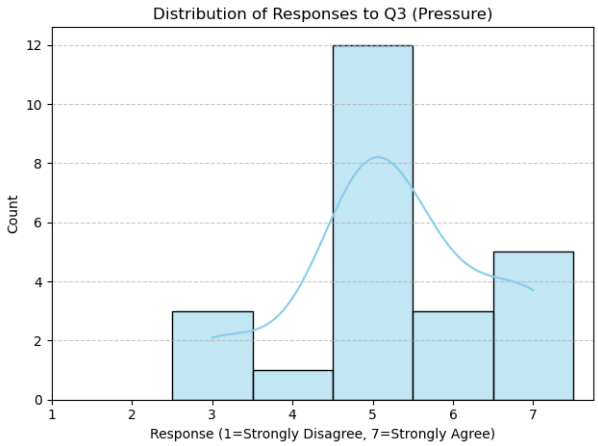
- Quantitative Metrics: Decision outcomes (approve/reject), choice of AI modality (Videobot vs. text-based), decision accuracy relative to a known "ground truth," response times, and final earnings based on correct or incorrect decisions.
- Subjective Measures: Post-task surveys using Likert scales to gauge perceived trust, pressure, satisfaction, and user experience with each AI interface.
- Followup: Collected participant comments after completion of the experiment, noting the reasons for selecting AI types, and perceptions of accuracy of one over the other.



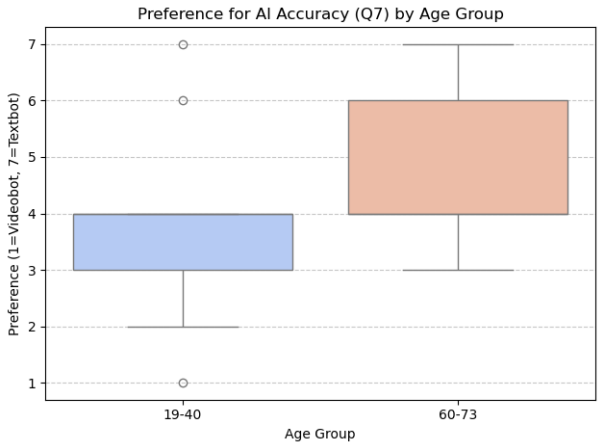
(a) Combined Trust (Q5 and Q6) by Gender



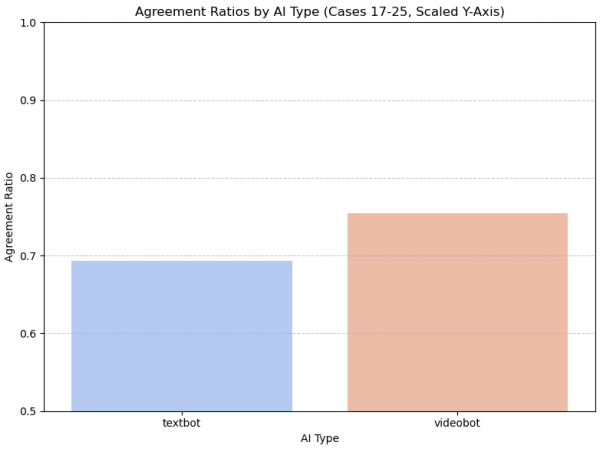
(b) AI Type Proportions by Gender (Cases 17-25)



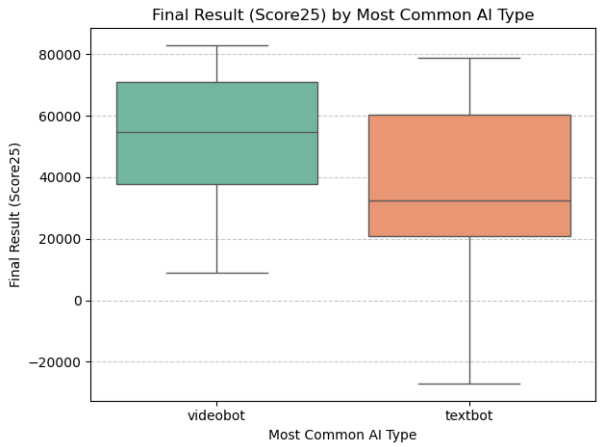
(c) Distribution of Responses to Q3 (Pressure)



(d) Preference for AI Accuracy (Q7) by Age Group



(e) Agreement Ratios by AI Type (Cases 17-25, Scaled Y-Axis)



(f) Final Score by Most Common AI Type

Figure 4: Key measures of trust, pressure, preferences, and outcomes under high-stakes conditions. (a) shows trust ratings (Q5, Q6) by gender, (b) illustrates differences in AI type usage by gender, (c) confirms that participants felt increased pressure (Q3), (d) highlights nuanced age-related differences in accuracy preference (Q7), (e) shows agreement ratios by AI type, and (f) compares final score by the most commonly used AI type.

6 RESULTS

Figure 4 presents six key visualizations from our study: (a) Combined Trust (Q5 and Q6) by Gender, (b) AI Type Proportions by Gender (Cases 17–25), (c) Distribution of Responses to Q3 (Pressure), (d) Preference for AI Accuracy (Q7) by Age Group (after outlier removal), (e) Agreement Ratios by AI Representation Type, and (f) Final Score by Most Common AI Representation Type. Appendix A details the survey that was administered at the end of the Test Phase.

6.1 High-Stakes Environment (Q3)

Participants reported feeling heightened pressure. The mean Q3 (question 3 of the survey) rating was 5.25 (SD = 1.22), median 5.0, on a 7-point scale (1 = Strongly Disagree, 7 = Strongly Agree). A one-sample t-test against a neutral midpoint (4) showed a highly significant effect ($t = 5.00, p < 0.0001$), confirmed by a Wilcoxon signed-rank test ($p = 0.0002$). These results validate that the study design successfully induced a high-stakes environment.

6.2 Trust in Textbot vs. Videobot (Q5 and Q6)

Overall, no significant difference was found between mean trust in the textbot (4.52) and videobot (4.64) based on Q5 and Q6 (t-test: $p = 0.7992$; Wilcoxon: $p = 0.6045$). However, examining Q6 (videobot trust) by gender revealed a noteworthy trend toward significance. A t-test comparing Q6 scores by gender yielded $t = -1.92, p = 0.0671$, and a Mann-Whitney U test produced $p = 0.0904$. While not strictly below the conventional 0.05 threshold, these results suggest a gender-related trend in subjective trust toward the videobot (Figure 4(a)).

6.3 Preference for AI Accuracy (Q7)

On average, participants showed no strong preference for the accuracy of one AI modality over the other (mean Q7 = 4.08; t-test: $p = 0.7878$, Wilcoxon: $p = 0.6968$). However, after removing outliers and comparing Q7 responses between younger (19–40) and older (60–73) age groups, a significant difference emerged. A t-test found $t = -2.48, p = 0.0220$, and the Mann-Whitney U test approached significance ($p = 0.0613$). These results indicate an age-related difference in perceived accuracy preference (Figure 4(d)), with older participants leaning more toward the Videobot as more accurate.

6.4 Stability of AI Choice After AI Errors

Prediction errors were introduced at specific trials in the experiment. We examined participants' changes in AI choice following a prediction error at trial 21 (comparing ratios from AI21 to AI22 phases). Among textbot users, 5 out of 14 (ratio 0.36) switched to videobot after an error, while only 2 out of 12 (ratio 0.17) videobot users switched to textbot. A proportion z-test yielded $z = 1.09, p = 0.2750$, indicating a nonsignificant but suggestive trend that participants were more stable in their videobot choice, showing less inclination to abandon the videobot after a mistake. This suggests greater "objective stability" of trust in the videobot.

6.5 Association Between Demographics and AI Choices

A chi-square test for age group (19–40 vs. 60–73) and chosen AI type showed a significant association ($\chi^2 = 4.18, p = 0.0409$). Older participants chose the videobot more frequently. For gender, a chi-square test confirmed a strong association ($\chi^2 = 13.16, p = 0.0003$), with females significantly more likely to choose the videobot (Figure 4(b)). These demographic-based differences in AI selection patterns align with observed subjective and objective trust markers.

6.6 Decision Times by Gender

Analysis of median decision times revealed gender differences. A t-test gave $t = -2.26, p = 0.0528$, and a Mann-Whitney U test showed $p = 0.0215$, indicating that males made decisions significantly faster than females. This aligns with other findings suggesting that females, who tended to pick the videobot more often, also took more time for thorough deliberation.

6.7 Decision Times by AI Type

Comparing decision times by AI type produced a strongly significant effect, despite the fact that information delivery times are precisely controlled to ensure that neither AI has an advantage. A t-test yielded $t = -3.81, p = 0.0002$, and the Mann-Whitney U test $p = 0.0015$, indicating that participants who chose the videobot spent substantially more time making their decisions than those who chose the textbot. Despite this longer decision time, videobot users ran out of time only 5 times, compared to textbot users who timed out 8 times. These patterns suggest that while videobot deliberation took longer, it did not necessarily lead to more timeouts.

6.8 Performance Outcomes and Agreement Ratios

No statistically significant difference in raw decision accuracy between videobot and textbot users was found. However, comparing final scores by most commonly used AI type revealed a trend. A t-test gave $t = -1.55, p = 0.1346$, suggesting that participants who relied more on the videobot achieved higher (though not statistically significant) final results (Figure 4(f)). Additionally, videobot users exhibited higher agreement ratios with AI predictions (Figure 4(e)). These results, while not conclusive, imply a positive relationship between videobot usage, agreement with AI predictions, and potentially better performance outcomes.

7 DISCUSSION

7.1 Interpretation of Results

Our results confirm that we effectively induced a high-stakes decision-making environment, as participants reported significantly elevated pressure (Q3). While subjective trust ratings (Q5, Q6) did not reveal large overall differences between the text-based AI and the Videobot, several objective trust metrics and demographic analyses painted a more nuanced picture.

A near-significant trend in Q6 indicated that females expressed slightly higher subjective trust in the Videobot, and demographic analyses using chi-square tests showed that females and older participants objectively chose the Videobot more frequently. Beyond

these subjective measures, we observed objective trust indicators: participants relying on the Videobot displayed greater stability in their AI choice even after experiencing incorrect predictions, suggesting that their trust was more resilient. Similarly, older participants identified the Videobot as more accurate (Q7) once outliers were excluded, and Videobot users demonstrated a trend toward achieving higher final results, despite not always making decisions faster.

In effect, those who valued careful deliberation and richer cues—often females and older participants—spent more time making decisions with the Videobot but did not experience disproportionate timeouts. Instead, they maintained stable trust and slightly better final outcomes. Conversely, participants who prioritized speed gravitated toward the text-based AI and switched away from it more readily after errors, reflecting a less stable trust relationship.

Taken together, these findings suggest that trust in AI is multifaceted and not captured solely by direct subjective ratings. While anthropomorphic features alone did not uniformly boost subjective trust, they aligned with certain user preferences and behaviors, fostering a more stable and nuanced trust relationship for demographics inclined toward thorough deliberation.

7.2 Implications for Design and Future Work

These insights underscore the importance of tailoring AI interfaces to user profiles and contexts in high-stakes decision-making scenarios. A streamlined text-based interface may better suit users who value speed and efficiency, whereas a Videobot—offering anthropomorphic cues—may appeal to those prioritizing careful deliberation and accuracy. In this study, females and older participants benefited more from the Videobot, exhibiting both subjective and objective trust patterns, along with stable decision strategies even under uncertainty.

Future work should explore how other anthropomorphic elements, cultural factors, and domain-specific conditions influence these trust relationships. Larger and more diverse participant samples, as well as dynamic manipulation of AI accuracy levels, could clarify how trust evolves over time and across scenarios. Ultimately, personalized AI interface designs that align with user characteristics and goals may enhance human-AI collaboration and lead to better outcomes in critical, high-stakes environments.

7.3 Limitations

This study did not set out to test differences in AI preferences or participant performance between different demographic (e.g., age, gender, ethnicity) groups, however we discovered that such differences may exist. In order to determine if preference is based on the modality itself, or demographic expression of the Videobot, we would need a much larger sample and to vary those traits between subjects.

All trials are presented in the same order to all participants. This might have created some unwanted order effects. In particular, challenges and predictions are presented in the same order, so there may have been unexpected correlations between the difficulty of the challenge with the placement of errors. Randomization of specific conditions and AI prediction accuracy would have reduced any order effects.

In qualitative followup, some of our study participants expressed a preference for the Videobot because it enabled them to visually study the attributes of the loan applicant while listening to the reasoning of the Videobot. This suggests that audio alone might elicit a similar response. Though our study did not test an audio-only interface (Audiobot), it could be modified to do so.

8 CONCLUSIONS

This study demonstrates that while a high-stakes environment can be reliably induced, trust in AI advisors is neither uniform nor solely determined by subjective ratings. Although no sweeping differences emerged in overall subjective trust, several objective trust metrics—such as stable usage patterns, resilience after incorrect predictions, and trends toward better performance—suggest a higher degree of trust in the Videobot among certain demographic groups, especially females and older participants.

These findings indicate that anthropomorphic design elements can foster a more deliberate and potentially more effective decision-making process for users inclined toward careful evaluation. Meanwhile, users who prioritize speed gravitated toward the text-based AI, reaffirming that interface design must align with user values and cognitive strategies.

In conclusion, one-size-fits-all approaches may be suboptimal in high-stakes settings. Recognizing demographic differences, trust nuances, and user goals can guide the development of adaptive AI presentation modalities that improve trust calibration, decision stability, and performance. Future studies should deepen our understanding of how to match AI interfaces to diverse user populations and evolving high-stakes conditions, ultimately enhancing human-AI collaboration in critical decision-making domains.

9 ACKNOWLEDGMENTS

We thank our advisor, Dr. Sid Fels, for feedback and guidance. We appreciate the participants for their time.

REFERENCES

[1] Zachary Carmichael. 2024. Explainable AI for high-stakes decision-making. *Ph.D. Thesis, University of Notre Dame* (2024).
[2] Justine Cassell. 2000. *Embodied conversational agents.* MIT Press.
[3] Sarah L Dommer and Vanitha Swaminathan. 2013. Explaining the endowment effect through ownership: The role of identity, gender, and self-threat. *Journal of Consumer Research* 39, 5 (2013), 1034–1050.
[4] Feral Dsouza, Rajshree Shaharao, Yogesh Thakur, Pratik Agwan, Gajanan Sakarkar, and Pooja Gupta. 2022. Advancement in communication using natural language based videobot system. In *2022 IEEE Bombay Section Signature Conference (IBSSC).* IEEE, 1–5.
[5] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114, 4 (2007), 864–886.
[6] David-Alexandre Frank et al. 2024. Navigating uncertainty: Exploring consumer acceptance of artificial intelligence under self-threats and high-stakes decisions. *Technology in Society* 79 (2024), 102073.
[7] N. Haduong and N. Smith. 2024. Raising the Stakes: Performance Pressure Improves AI-Assisted Decision Making. *ScienceEngineering* (2024).
[8] Kevin A Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434.
[9] Barbara E Kahn. 1995. An exploratory study of choice rules favored for high-stakes decisions. *Journal of Consumer Psychology* 4, 4 (1995), 305–328.
[10] Thomas Kim, Cindy Yu, Charles Hinson, Emily Fung, Oren Allam, Rahim Nazerali, and Ram Ayyala. 2024. ChatGPT virtual assistant for breast reconstruction: assessing preferences for a traditional chatbot versus a human AI videobot. *Plastic and Reconstructive Surgery–Global Open* 12, 10 (2024).

[11] Howard Kunreuther et al. 2002. High stakes decision making: Normative, descriptive and prescriptive considerations. *Marketing Letters* 13, 3 (2002), 259–268.
[12] Shunan Ma. 2024. Towards Human-centered Design of Explainable Artificial Intelligence (XAI): A Survey of Empirical Studies. arXiv:2410.21183.
[13] Shunan Ma et al. 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*. 1–18.
[14] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[15] Ahmad Saffarini. 2023. Trusting AI in high-stake decision making. *arXiv preprint arXiv:2301.13689* (2023).
[16] Benjawan Sahoh and Anan Choksuriwong. 2023. The role of explainable artificial intelligence in high-stakes decision-making systems: A systematic review. *Journal of Ambient Intelligence and Humanized Computing* 14 (2023), 7827–7843.
[17] Archit Sharma. 2023. Loan Approval Prediction Dataset. https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset.
[18] Norman G. Vinson, Jean-François Lapointe, and Nicolas Lemaire. 2022. An emergency centre call taker task analysis. In *Lecture Notes in Artificial Intelligence*, Vol. 13307. Springer, 225–241.
[19] Norman G. Vinson, Jean-François Lapointe, and Nicolas Lemaire. 2023. An emergency centre dispatcher task analysis. In *Proceedings of the 20th International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer.
[20] Ewart de Visser et al. 2012. The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56. SAGE Publications, 263–267.
[21] Fangjian Yu et al. 2024. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine* 30 (2024), 837–849.
[22] Andrew Zytek, Didi Liu, Rakesh Vaithianathan, and Kalyan Veeramachaneni. 2021. Sibyl: Explaining Machine Learning Models for High-Stakes Decision Making. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*.

A SURVEY INSTRUMENT

We administered a survey consisting of eight items to assess participants' perceptions, experiences, and trust levels in the decision-making task. Most items employed a 7-point Likert scale, a commonly used measurement technique in user experience and human-AI interaction research, which allows for nuanced expression of agreement or difficulty. The endpoints varied according to the nature of each question, as detailed below:

Q1. I found it easy to choose which clients to accept without AI assistance.
Scale: 1 = Very Easy, 7 = Very Hard

Q2. I was confident during the decision-making process.
Scale: 1 = Strongly Disagree, 7 = Strongly Agree

Q3. I felt pressure to make the right decision.
Scale: 1 = Strongly Disagree, 7 = Strongly Agree

Q4. The AI recommendations helped me to make better decisions.
Scale: 1 = Strongly Disagree, 7 = Strongly Agree

Q5. I trusted the recommendations of the AI textbot.
Scale: 1 = Strongly Disagree, 7 = Strongly Agree

Q6. I trusted the recommendations of the AI videobot.
Scale: 1 = Strongly Disagree, 7 = Strongly Agree

Q7. Which AI was more accurate in its recommendations?
Scale: 1 = Videobot, 7 = Textbot

Q8. I would like to use AI to make decisions in the future.
Scale: 1 = Strongly Disagree, 7 = Strongly Agree

The items used a 7-point Likert scale anchored at "Strongly Disagree" (1) and "Strongly Agree" (7), reflecting a standard approach to capturing subjective attitudes [**?**]. For Q1, we adapted the endpoints to measure perceived difficulty (1 = Very Easy, 7 = Very Hard), and for Q7, we customized the endpoints to compare perceived accuracy between the two AI modalities (1 = Videobot, 7 = Textbot).