Original Research Article

Explaining Human-AI Interaction: An Experimental Study of Transparency and Cognitive Load in AI Decision-Making

Claire Daeun Kim

Bergen County Academies, 200 Hackensack Ave, Hackensack, NJ 07601, United States

ABSTRACT

As artificial intelligence systems increasingly assist or replace human decision-making in high-stakes domains, transparency has become a central design principle intended to support user understanding and trust. However, the effectiveness of transparency may diminish as users may not process the provided information entirely due to cognitive load. Thus, it is unclear whether detailed explanations of AI systems always lead human users to evaluate the system in a positive light. To answer the question, this study conducted an online experiment where 181 participants were given AI-assisted decision-making scenarios with different levels of decision transparency and cognitive load. Contrary to the expectation that transparency may backfire under high cognitive load, the findings show that transparency about the decision process consistently increased perceived trustworthiness and ethicality of AI systems regardless of cognitive load. Moreover, the experiment results find that the level of explanation detail or cognitive load does not alter users' responsibility attribution, pointing to the rigidity of moral responsibility in the AI-assisted decision-making contexts. This study contributes to the literature of AI-human interaction by empirically demonstrating the limited role of cognitive load in shaping human perceptions of AI decision-making. In addition, this study hints at important practical implications for how AI-human interaction should be designed to foster the effectiveness of the interaction.

Keywords: Human–AI Interaction; Cognitive Load; Algorithmic Transparency; Trust in AI; Ethical Perception; Responsibility Attribution; Experimental Study

INTRODUCTION

Decision-making has long been believed to be a specific agentic function of human beings, as humans are the only ones who can take moral responsibility. However, as the advancement of AI systems is increasingly accelerating, it is widely observed that human decision makers receive assistance from AI systems in many areas. Even in high-stakes areas, including healthcare, legal judgment, financial investment, or education, AI systems are widely collaborating with humans or even replacing humans in some cases. For example, the healthcare AI industry is expected to grow by 35% annually by 2030, with the estimated industry size reaching USD 180 billion (1).

As AI-assisted decision-making pervades at an unprecedented speed, the question arises as to how we can foster the trust of human decision makers in AI systems and thereby productive and collaborative AI-human relationships. Although there has been a myriad of design principles suggested that may help increase

Corresponding author: Claire Daeun Kim, E-mail: clairekim.nyop@hotmail.com.

Accepted April 21, 2026
https://doi.org/10.70251/HYJR2348.42465477

the level of trust, theoretical justifications and empirical tests on them are still limited, as the generative AI tools are still nascent. One core argument is that transparency of the AI decision-making process may increase trust in human decision makers. Existing studies suggest that a transparent explanation of the system outcome may increase trust and the sense of fairness and justice (2). As a result, transparency has become a central design principle in many AI guidelines. However, transparency is often treated as if its effectiveness is uniform across contexts and users. In real contexts, users may not always read explanations carefully or engage in the kind of deep reasoning that these approaches often assume.

Transparency of AI systems may not always be perceived in a positive light due to the limitations of human cognition. When humans are mentally busy or the amount of information is too big to process, they often rely on intuitive thinking or heuristics rather than rational analysis (3). Under such conditions of high cognitive load, explanations that require memorization and additional interpretation may be more difficult to process or may influence judgments differently than expected. Despite this potential role of cognitive contexts, limited empirical work has been done on how cognitive load alters users' perception of AI systems' transparency. Existing studies and practice often assume that explanations primarily function as cognitive aids, meaning their benefits should increase when users lack the capacity to evaluate information fully (4). Whether this assumption holds in realistic AI decision contexts remains an open question. In addition, responsibility attribution is a critical yet often underexplored dimension of AI-human interaction. When AI systems are used for decision-making, people may differ in whether they hold themselves or the AI system responsible for outcomes. Prior research shows that individuals sometimes execute what is advised by AI systems and assign the responsibility of consequence to AI systems, although AI systems present only recommendations without making final decisions (5). Importantly, responsibility judgments may rely on different psychological mechanisms than trust evaluations.

Taken together, existing research widely agrees on the role of AI systems' transparency in increasing trust of human users. However, the literature remains divided on whether more detailed explanations are always beneficial to AI-human interaction. On one hand, one line of research demonstrates that increased level of transparency leads to increased trust by helping users better understand how the AI systems work and produce outcomes (2, 6). On the other hand, another line of research warns against too much detailed explanations by showing that overly detailed explanations can produce cognitive overload that makes it difficult for users to process outcomes and thus to lose confidence in the outcomes (4). While the competing roles of transparency are widely discussed in the two lines of literature, they have rarely been brought together in a single study to test how explanation detail interacts with cognitive load. Empirically, some studies examine the relationship between transparency and trust without manipulating cognitive load (7, 8), while others investigate the role of cognitive burden without varying explanation detail (4, 9). In particular, many empirical studies mostly focus on the influence of transparency and cognitive load on functional trust (i.e., how reliable the outcome is), leaving ethical trust (i.e., how ethical the outcome is) and reasonability attribution underexplored.

The present study attempts to fill this gap by examining how explanation transparency and cognitive load jointly shape users' perceptions of AI-assisted decisions in a single empirical study. Specifically, the study evaluates whether explanation detail and cognitive load jointly shape some of the critical factors that influence trust, including perceived ethicality and trustworthiness. In addition, this study also explores the understudied question of whether AI transparency can alter users' responsibility attribution. To do so, this study conducted a between-subjects experiment using a 2 (cognitive load: low vs. high) × 3 (explanation transparency: none vs. simple vs. detailed) factorial design. From those conditions, participants evaluated AI-generated recommendations across three decision scenarios while experiencing a simultaneous cognitive load manipulation.

The paper is structured in the following manner. First, the literature review finds unresolved questions about how transparency and cognitive context can shape users' perception of AI-assisted decision-making and whether they can alter responsibility attribution. The review leads to three hypotheses that the experiment tested. Second, the methodology describes how the experiment was designed and conducted with considerations for methodological rigor. Third, the results of statistical analysis that test the relationship between transparency, cognitive load, trustworthiness, ethicality, and responsibility attribution are presented. Fourth, the implications and limitations of our findings are discussed in detail. Finally, it concludes by suggesting broader implications for human-AI interactions.

LITERATURE REVIEW

Transparency provides insight into the internal workings of AI algorithms and is widely promoted as a means of ensuring the ethical accountability of AI systems (7, 10). It is widely assumed in theories and practice that transparency with detailed explanation about how AI systems reached a conclusion enhances users' trust in the systems (10, 11). When users understand the reasoning behind AI decisions, they can better evaluate whether the system acts on concrete reasoning and moral standards.

For instance, Binns *et al*. (8) find that explanations that clarify the decision-making process allow users to align the use of AI tools with their moral reasoning. This transparency increases users' perceived accountability and justice of AI outputs. Their study made the conclusion by conducting experiments with over 1,200 participants who encountered AI decisions explained in multiple formats: input influence, sensitivity-based, case-based, and demographic-based. When participants could explore multiple explanation styles, which increases the level of transparency, they judged the AI's decisions as fairer and more ethical. But when participants were exposed to only one type of explanation, the perceived ethicality diminished.

However, the cognitive perspective suggests that when users are required to process complex information under time pressure or mental strain, they may rely on guesses rather than analytic reasoning. In other words, the effect of transparency may diminish in a cognitively demanding context. Emerging evidence suggests that the high-transparency principle does not account for the contextual factors that shape human responses to algorithmic outcomes. For example, in Bućinca, Malaya, and Gajos' (4) study of AI-assisted meal planning, participants required to make independent decisions before consulting AI suggestions reported higher confidence in decision-making. Yet, when tasks were cognitively demanding, this decision-making structure offered limited gains in confidence, implying the potential role of cognitive load in shaping human perceptions of AI-assisted decision-making.

AI transparency in the cognitively demanding context may not only be ineffective but also backfire. Kizilcec (7) argues that excessive transparency can cognitively overwhelm users, especially when outcomes are unexpected, potentially undermining users' perception of trustworthiness and ethicality. In the study, 103 MOOC students received peer-assigned grades adjusted by AI systems and were asked to evaluate their confidence in the AI outcomes. When they received a medium level of explanation, which only entailed the algorithm's adjustment for grader bias, the students perceived the grading process as fairer, even if the results violated their expectations. However, when provided with full transparency, including raw peer grades and detailed adjustments, participants often felt overwhelmed and less confident in the system. The results imply that too much detail, in the absence of cognitive capacity, undermined rather than increased ethical perception. Further support comes from Poursabzi-Sangdeh *et al*. (9), who studied nearly 3,800 participants in a real estate valuation task. They found that participants interacting with simple, two-feature models could simulate AI predictions accurately. In contrast, participants assigned to a group with complex, eight-feature models struggled to understand and evaluate the outcomes produced by AI systems. Therefore, the idea that "more transparency leads to higher perceived trustworthiness and ethicality" may not be generalized to different cognitive-load conditions. However, the literature also suggests that the quality and structure of explanation, rather than just its level of transparency, are also crucial. Because users often see the explanation detail in algorithmic outcomes as a signal of system competence (4), a simplified explanation may lead to a lower level of perceived ethicality and trustworthiness. Simple explanations about the decision-making process may provide limited information about system logics and imply systems' incompetence in logical and ethical decision-making. This insight challenges the assumption that there can be a linear and positive relationship between transparency and users' perceived ethicality and trustworthiness.

In conclusion, detailed explanations can increase the perceived ethicality and trustworthiness of AI systems, but only when users have the mental capacity to process them. Under low cognitive load, transparency gives the sense of fair perception and a trustworthy system. Under high cognitive load, even the carefully designed explanations had the possibility to backfire, leading to confusion or superficial judgments. These observations underpin two testable hypotheses that have been inconclusive in the literature: Hypothesis 1. Detailed explanations reduce the perceived ethicality of AI decisions when a user's cognitive load is high, compared to no or simple explanations, and Hypothesis 2. Detailed explanations reduce the perceived trustworthiness of AI decisions when a user's cognitive load is high, compared to no or simple explanations.

While transparency serves as a critical factor that influences human judgment of AI systems and thus fosters effective human-AI interactions, the notion of ethical responsibility proposes an important question: given that transparency can shape a user's perception of ethicality and trustworthiness of algorithmic decisions, is it desirable for the users to take an action in accordance with the decisions without taking ethical responsibility? This is a morally important question because the ethics of technology proposes that humans are morally held responsible for their use of technology. If users' positive perceptions of algorithmic decisions automatically lead to blinded decision-making without ethical responsibility, humans will have the least agency in the era of AI decision-making, undermining the ultimate value of human decision makers.

Some studies suggest that whether users attribute moral responsibility for decisions to themselves or AI systems largely relies on the context. For example, Dzindolet *et al*. (12) argue that, when human users hold a high level of trust in automation systems, they tend to hold the systems responsible for the outcomes. However, this trust does not emerge in a void. As discussed, with too much contextual information to process, human users tend to lose their trust in AI systems and therefore take responsibility. In contrast, some empirical studies suggest that high cognitive strain hinders human users from evaluating the moral implications of algorithmic decisions carefully; thereby attributing default moral responsibility to the system instead (13). In addition, explanations under cognitive load may unintentionally encourage moral outsourcing, leading users to defer responsibility to AI systems regardless of their understanding (11). These two reasons propose different directions of responsibility attribution under cognitive load in an AI-assisted decision-making context, which requires further investigation.

It is worth noting that other studies fundamentally challenge the idea that responsibility attribution depends on the context and instead suggest it is stable. It is argued that responsibility judgments often rely on stable moral intuitions rather than on situational cues such as explanation detail or cognitive effort. For example, in their Moral Machine Experiment, Awad *et al*. (5) collected large amounts of data about how individuals evaluate multi-dimensional moral dilemmas faced by autonomous vehicles. They found that individual moral judgments are relatively stable across different scenarios. Instead, they found similarities between individuals with similar cultural backgrounds, proposing that individual moral principles are preconditioned regardless of situations. This finding suggests the inconclusive question of whether decision context influences responsibility attribution or not in an AI-assisted decision-making context. This narrative proposes the last testable hypothesis as follows: ***Hypothesis 3****. Under high cognitive load, individuals attribute more ethical responsibility to the AI system than to themselves.*

METHODS AND MATERIALS

The current study aims to test the extent to which transparency in AI explanations and cognitive load of users influence individuals' perceptions of the ethicality, trustworthiness, and responsibility attributions of AI-assisted decision-making systems. As AI tools become increasingly integrated into business, industries, and policy, it is critical to understand how users evaluate AI systems when faced with varying levels of cognitive load and informational clarity.

To address this issue, the study employed an experimental approach using an online experiment platform, Testable, where adult participants, who are familiar with diverse decision-making scenarios in their real lives, went through a series of structured decision scenarios and completed associated evaluations. Before conducting the main experiment, a pilot experiment was built on Testable, and 15 participants were independently recruited from October 29 to October 31, 2025, who completed the pilot experiment, returned their survey answers, and provided feedback. After some refinements reflecting the feedback, the main experiment was conducted from November 13 to December 2, 2025. In total, 184 participants were recruited through the Minds pool on Testable. These individuals volunteered for the study and received a small monetary incentive administered directly through the platform. Only those who were 18 years old or older at the time of participation were allowed to participate because only adults are likely to be exposed to diverse decision scenarios that involve individual responsibility. Because this study aims to investigate general perceptions of AI-decision making, decision criteria other than age were not introduced, as others may restrict the generalizability. However, three participants were excluded from the analysis because they failed to pass the manipulation check. Thus, the final sample consists of 181 participants. The characteristics of the sample, including their age, gender, and regional distribution and familiarity with AI tools, as described in Table 1 in the results section, confirm that the final

sample is well balanced to secure generalizability of the findings. Details of the sample characteristics are discussed in the results section.

The experiment employed a between-subjects 2 × 3 experimental design. This design was used to systemically investigate the two independent variables —AI explanation transparency and cognitive load —on individuals' perceptions of ethicality, trustworthiness, and responsibility attribution in AI-assisted decision-making contexts. For the group assignment, participants were randomly assigned to either a *low-load* (e.g., memorizing a simple 2-digit code while reading an AI-assisted decision-making scenario) or a *high-load* (e.g., memorizing a 7-character alphanumeric string while reading an AI-assisted decision-making scenario) condition. Participants were also assigned to one of three explanation conditions: In the *None* condition, participants were provided with no explanation about how AI systems made a conclusion in the decision scenario. In *the Simple* condition, only a concise rationale without detailed decision criteria was provided. In the *Detailed* condition, the explanation about the system recommendation included specific, technical model logics and fairness considerations.

Once participants joined the experiment and the random group assignment was completed, participants were asked to complete a brief set of questions that may serve as predictors of AI evaluation: numerical comfort, familiarity with AI systems, and need-for-cognition traits. Demographic information, including age, gender, and education level, was also collected to contextualize the sample. These factors were not treated as central variables but served for descriptive statistics of the study population.

Then, participants viewed three AI-assisted decision-making scenarios—loan approval, scholarship recommendation, and content moderation. For example, the loan approval scenario explained a loan applicant's personal profile (e.g., *Jordan, M., 34 years old, works full-time as an office administrator and has been employed in the same company for five years*…) and then presented AI's recommendation on the loan application (in the scenario, it was a rejection). It was followed by a system explanation of how the AI system made the recommendation with varying degrees of detail. For the *None* condition, no explanation was provided. For the *Simple* condition, the explanation provided only a vague decision rule (e.g., *income stability and existing debt level to estimate repayment ability*). For the *Detailed* condition, the explanation provided not only the decision criteria but also the system's numerical evidence in detail (e.g., *the system calculated a predicted default risk of 22*) (for the exact scenario in each condition, please see Appendix 1).

For the experimental conditions to remain the same across the three scenarios, the same level of explanations (*none, simple*, or *detailed*) and cognitive load (*low* or *high cognitive load*) were kept the same across the three scenarios for a specific participant. This design was intentional because allowing participants to move between conditions would introduce demand characteristics, in which participants guess the research purpose, which in turn can alter their responses. Keeping both manipulations constant prevents participants from noticing the experimental condition and helps ensure the responses reflect genuine reactions of the participants to the stimuli.

After reading a decision scenario, participants were asked to evaluate the ethicality and trustworthiness of the AI system that made the decision along the 11 measures revised from Choung *et al*. (14). Their study developed 11 items of trust in smart technologies on a 7-point Likert scale: six items pertain to human-like trust, and five items pertain to functionality trust in smart technologies. As these two domains well represent the theoretical concepts of perceived ethicality and trustworthiness, respectively, a revised set of these 11 items was used for participants to evaluate the ethicality and trustworthiness of the AI system. In particular, six items were used to measure perceived ethicality and five for perceived trustworthiness. For example, the first perceived ethicality item states "*the AI system cares about the well-being of the people affected by its recommendations*" and asks the participants to indicate the degree of agreement from 1 (=strongly disagree) to 7 (=strongly agree). In the meantime, the first perceived trustworthiness item states "*the AI system works well when making decisions like this one,*" and then also asks the same evaluation. By averaging the scores across the items, it was able to measure a specific participant's perceived ethicality and trustworthiness of AI-assisted decision-making. As the measures were adapted to the study's context accordingly, it was necessary to check the reliability of the measures by seeing whether these items exhibit internal consistency. The adapted measures showed high level of internal consistency in the current sample. Cronbach's alpha was 0.97 for perceived ethicality and 0.94 for perceived trustworthiness, confirming that the adapted measures represent their intended theoretical constructs consistently.

Responsibility attribution was measured on a 100-point scale, where *0* means full human responsibility, and *100* represents full AI system responsibility for the final decision made with the assistance of the AI recommendation. The survey always followed each scenario, so it could appropriately measure participants' perceptions of each scenario and responsibility attribution separately. Appendix II presents all the items used to measure perceived ethicality and trustworthiness.

To make sure the group manipulations, the experimental design used two strategies together. While the level of explanation can be directly manipulated in the scenario, the cognitive load is a subjective experience that cannot easily be observed. Thus, the experiment set the time limit for the reading of the scenario by considering the time required to memorize a 2-digit (e.g., 82) code for the low cognitive load condition or a 7-character alphanumeric string (e.g., AF345Z1) and reading the scenario. It was short enough time to put cognitive load in conducting memorizing and reading tasks at the same time, and long enough to complete the reading. For example, for group 1 (simple explanation and low cognitive load), 25 seconds were given to memorize the code and read the scenario. For group 6 (detailed explanation and high cognitive load), whose tasks were the most demanding, 75 seconds were given. Then, a code recall task followed the scenario with the code, whose result verifies whether the participants experienced cognitive load from the code memorization task or simply skipped the reading and task. When the answer in the recall task is significantly different from the original code, the participant was classified as failing the manipulation and thus excluded from the sample.

To check whether the manipulation of cognitive load worked or not, at the end of the experiment, the following question was asked in a 7-point Likert scale: "How mentally demanding did you find the overall task?" An independent-samples t-test showed that participants in the high cognitive load condition (group 1,2 and 3) perceived the task significantly more demanding (M = 4.77, S.D. = 1.78) than those in the low cognitive load condition (group 4,5, and 6) (M = 4.02, S.D. = 1.76) ($t(179) = -2.83$, $p = .0051$, $d = 0.42$). This result indicates successful manipulation of cognitive load.

At the end of the study, participants were presented with a full debriefing explanation about the actual aims of the research. Because the study employed mild masking of the true research purpose to reduce demand characteristics, the debrief was essential in restoring full transparency and ensuring ethical completeness. The debrief also clarified how the collected data would be used in academic analysis and reporting.

As this study involved an experiment with human participants, certain ethical issues had to be addressed. Although the content of the study had a very low risk of trauma trigger or lasting physical or mental impacts, anonymity was a central issue. Thus, participants' names or personal information were not disclosed in the analysis process, and only relevant pieces of background information that helped answer the research question were included. Furthermore, informed consent was collected from all the participants at the beginning of the study. At this stage, all important details of the study were clearly explained. The respondents were also informed that they could withdraw from the study during the process.

The collected data were analyzed by using two-way ANOVA to test whether group conditions (*none* vs. *simple* vs. *detailed* and *low* vs. *high cognitive load*) influence users' perceived ethicality and trustworthiness of AI systems and responsibility attribution. All the analysis was conducted in the open-source statistical software, R.

RESULTS

To test how the level of explanation and cognitive load influence perceived ethicality, trustworthiness, and responsibility attribution, the current study presents two analyses. First, descriptive statistics describe demographic characteristics of the sample and the general tendency of participants' perception of the scenarios. Second, inferential statistics with the ANOVA approach report statistical significance found in any difference across the group conditions to confirm the suggested hypotheses.

Table 1 reports the background information of 181 experiment participants. The sample is gender-balanced with an average *age* of 35. The distribution of participants' region of residence is also well balanced, with the largest portion of the sample from Europe (n=66, 36.36%), ensuring the generalizability of the result. More than 80% of the participants hold *bachelor's degrees or higher*, indicating that they are likely to be exposed to diverse AI tools and decision-making scenarios in real life. Confirming this assumption, the average self-reported scores on *familiarity with AI tools* and *comfort with numbers* exceed 5 out of 7. The participants also reported a high preference for *complex, time-consuming,* and *difficult tasks,* with the average scores of 4.97, 5.17,

and 5.31, respectively. Therefore, the sample was suitable for the experiment, engaging various decision-making scenarios with moral and technical aspects.

The descriptive statistics of the main dependent variables show how transparency of AI systems and cognitive load of the decision contexts influence participants' perceived trustworthiness and ethicality, and responsibility attribution, which is summarized in Table 2. Contrary to the expectations formulated in the first and second hypotheses, detailed explanation always led to higher levels of perceived trustworthiness (M= 5.27, S.D.= 0.98 in low cognitive load and M= 5.47, S.D.= 1.05 in high cognitive load) and ethicality (M= 5.03, S.D.= 1.16 in low cognitive load and M= 5.19, S.D.= 1.30 in high cognitive load) compared to no and simple explanation conditions, regardless of cognitive load. Interestingly, simple explanation consistently yielded lower levels of perceived trustworthiness (M= 4.58, S.D.= 1.19 in low cognitive load and M= 4.68, S.D.= 1.28 in high cognitive load) and ethicality (M= 4.32, S.D.= 1.23 in low cognitive load and M= 4.25, S.D.= 1.63 in high cognitive load) compared to no explanation. Although the statistical significance of the differences should be confirmed by using inferential statistics, the

Table 1. *Descriptive Statistics of Background Information.*

	Mean	Median	S.D.	min	max	Total
Age	36.85	35	11.00	18	73	181
Familiarity with AI tools	5.47	6	1.29	1	7	181
Comfort with Number	5.80	6	1.23	2	7	181
Thinking Preference: Complex	4.97	5	1.45	1	7	181
Thinking Preference: Time-consuming	5.18	5	1.38	1	7	181
Thinking Preference: Difficult	5.31	6	1.38	1	7	181
Gender	**Man**		**Woman**		**Other**	181
	92 (53.80%)		83 (45.11%)		2 (1.09%)	
Education	**Less than High School**	**High School or Equivalent**	**College or Technical School**	**Bachelor's Degree**	**Master's or Higher**	181
	1 (0.54%)	21 (11.41%)	12 (6.52%)	104 (56.52%)	46 (25.00%)	
Region	Africa	Asia	Europe	North America	South America	181
	45 (24.86%)	36 (19.89%)	66 (36.46%)	30 (16.57)	4 (2.21%)	

Table 2. *Descriptive Statistics of Perceived Trustworthiness, Ethicality, and Responsibility Attribution across Six Groups.*

Load	Transparency	n	Mean (Trustworthiness)	S.D.	Mean (Ethicality)	S.D.	Mean (Responsibility)	S.D.
Low	**None**	32	4.78	1.52	4.44	1.54	73.21	25.99
Low	**Simple**	31	4.58	1.19	4.32	1.23	65.87	22.48
Low	**Detailed**	28	5.27	0.98	5.03	1.16	72.64	23.10
High	**None**	30	4.74	1.10	4.67	1.36	67.14	29.99
High	**Simple**	30	4.68	1.28	4.25	1.63	53.89	32.91
High	**Detailed**	30	5.47	1.05	5.19	1.30	76.86	24.18

results of the descriptive statistics reveal that, regardless of cognitive load, detailed explanation enhances users' perceived trustworthiness and ethicality, while simple explanation often produces negative effects on the human perceptions of AI systems.

A two-way ANOVA analysis reveals the statistical effect of transparency, cognitive load, and their interactions on human perception. Significant main effects of transparency on trustworthiness ($F(2, 175) = 6.22$, $p = .002$, $\eta^2_p = .066$) and on ethicality ($F(2, 175) = 5.49$, $p = .005$, $\eta^2_p = .059$) were found, confirming the general patterns that emerged in descriptive statistics. However, the main effects of cognitive load on perceived trustworthiness ($F(1, 175) = 0.24$, $p = .623$, $\eta^2_p = .001$) and ethicality ($F(1, 175) = 0.28$, $p = .598$, $\eta^2_p = .002$) were found insignificant, implying cognitive load alone does not significantly influence the human perceptions of AI systems. Finally, to see whether explanation detail backfires under high cognitive load as hypothesized, the statistical significance of the interaction between transparency and cognitive load should be checked in the two-way ANOVA analysis. Contrary to expectations, no significant interaction effects were found both for trustworthiness ($F(2, 175) = 0.14$, $p = .866$, $\eta^2_p = .002$) and ethicality ($F(2, 175) = 0.20$, $p = .823$, $\eta^2_p = .002$).

The descriptive statistics of responsibility score (1= fully human, 100= fully AI) exhibit a lower mean score of high cognitive load groups ($M = 65.96$, S.D. =30.43) than that of low cognitive load groups ($M = 70.53$, S.D. =23.93), as shown in Table 3. This result contrasts with the expectation that users under high cognitive load may attribute more decision responsibility to AI systems in the decision-making context. However, an ANOVA analysis revealed no significant main effect of cognitive load on responsibility attribution ($F(1, 175) = 1.35$, $p = .248$, $\eta^2_p = .008$), indicating no significant statistical difference between the two conditions. In contrast, there was a significant main effect of transparency ($F(2, 175) = 4.87$, $p = .009$, $\eta^2_p = .053$), implying users provided with a detailed explanation may attribute more responsibility to AI systems.

Table 3. *Descriptive Statistics of Responsibility Score in High vs. Low Cognitive Load Groups.*

Load	n	Mean (Responsibility)	S.D. (Responsibility)
High	90	65.96	30.43
Low	91	70.53	23.93

DISCUSSION

This study investigated how explanation transparency and cognitive load both influence human perceptions of an AI system's decision-making. Participants evaluated decisions made under varying levels of explanation detail (none, simple, detailed) and cognitive load (low and high). The central finding showed that explanation transparency has a positive effect on perceived trustworthiness and ethicality, regardless of the level of cognitive load. Responsibility attribution, in contrast, remained relatively stable across all cognitive conditions. The results suggest that the level of transparency is a central factor in how trustworthy AI systems are perceived. This finding is contrary to some theoretical predictions that cognitive load interferes with the process of judgments (15). Instead, it is consistent with the previous findings that detailed explanations can act as a signal of the competence of an AI system in addition to an information-processing aid (2, 8).

The persistence of transparency under all cognitive conditions further suggests that explanations may operate through surface-level judgment cues rather than through deep cognitive processing. Users may not fully analyze the content of detailed explanations and instead infer that a system that provides extensive justification is more trustworthy and ethical. However, prior theories often assume that explanations require effortful cognitive processing and that high cognitive load should therefore diminish their effectiveness or even produce a backfire (4). Given the present findings, there is a deviation suggesting that users may rely on heuristic cues rather than analytic reasoning when evaluating AI decision-making. This pattern is consistent with dual-process theories of decision making, which explain that individuals rely more heavily on intuitive cues when cognitive resources are constrained (3). Particularly, the manipulation of cognitive load was successful in the experiment, as demonstrated in the methodology. Thus, the insignificant main effects of cognitive load on perceived trustworthiness and ethicality should not be interpreted as a failure in experiment manipulation. Rather, it should be interpreted as the dominating role of transparency as heuristic cues rather than as information aid.

Surprisingly, simple explanations frequently produced the lowest trustworthiness and ethicality ratings. It is often thought that minimal explanations are a compromise between transparency and cognitive efficiency. But the present results suggest that simple

explanations may inadvertently undermine human confidence in AI decisions. One possible interpretation is an expectation-violation effect. Simple explanations may raise users' expectations for transparency. But when the AI system seems unable to provide sufficient substance, users have skepticism (8). Similarly, prior research has shown that partial transparency can reduce trust when users perceive explanations as incomplete or misleading (6, 16). Rather, omitting an explanation ensures that unmet expectations are not set.

It is worth noting that, although the main effects of transparency were statistically significant on perceived trustworthiness and ethicality, the effect sizes were relatively small (η^2_p = .066 and .059, respectively) in a conventional sense. However, they fall within the modest effect size range in behavioral research. For example, Welkowiz *et al*. (17) provide standard benchmarks for partial eta squared (η^2_p) with 0.06 being considered a medium effect size, meaning the effect explains 6% of variance, in behavioral research. Thus, the effects represent meaningful differences in human perception across different levels of transparency. Furthermore, in the context of human-AI interaction, even a small shift in perceived trustworthiness and ethicality is meaningful because this change drives individual tendency to rely on AI systems in decision-making, and the collection of such individual behavioral shifts can yield large overall change at the organizational and societal level.

Responsibility attribution did not vary significantly across cognitive load conditions. This could signal that responsibility judgments are relatively stable and less susceptible to situational differences. Participants appeared to rely on their pre-existing moral intuitions about accountability rather than adjusting to situational factors. This finding aligns with previous work, affirming that responsibility judgments are more deeply rooted in individual moral reasoning and social norms than judgments of fairness, which are more context-sensitive (5, 18). The present findings suggest that transparency functions not only as an informational aid for judgment but also as a competence cue that shapes users' perceptions. Transparency may therefore remain significant in the design of AI systems even when users face cognitively demanding tasks. This result also points to the need for conceptual differentiation. While the informational aid emphasizes comprehension and information processing, the legitimacy cue highlights the social and moral values conveyed by transparency (2, 15).

From a theoretical point of view, future research should explicitly distinguish AI explanation as an informational aid and as a competence cue. By distinguishing the two, AI research can better understand how users interpret and respond to outputs that AI systems generate. The result also proposes straightforward practical implications. Well-crafted, detailed explanations can actually enhance trust even in cognitively demanding contexts. Thus, AI design, especially for systems that interact with human judgment, should be equipped with significant transparency rather than defaulting to minimal explanations. The results caution against oversimplified explanations, which may be worse than no explanation at all, particularly in high-stakes scenarios.

Future research should test the strength of these findings across more diverse populations, particularly users with lower familiarity or confidence in AI systems. Studies involving real-time decision-making contexts would be useful in clarifying whether transparency effects continue to persist. Additionally, future work should also explore alternative explanation formats, such as visual or interactive explanations, to determine whether different forms reinforce or weaken the normative signaling function of transparency. Finally, a more in-depth investigation is needed to understand why responsibility judgments resist situational manipulation.

CONCLUSION

The history of human–machine interaction has often been characterized by technical progress, including the accuracy and efficiency of technologies. However, the emergence of modern AI systems marks a shift toward a new question: the central question is not whether the machine can perform well enough, but whether human beings can trust what it produces.

This is where transparency is an essential trait. Users do not simply want accurate outcomes. They want to know how those outcomes came to be. Specifically, they want explanations that feel grounded and aligned with human reasoning. This is why AI systems should invite users into the decision-making logic rather than leaving them outside. And when systems clarify both how they arrived at a decision and why that process is justified, trust becomes something earned rather than assumed.

Ultimately, transparency is not just a feature. It is the foundation for trust, ethical judgment, and responsible deployment. As the boundaries between human and machine decision-making continue to blur, transparent AI systems offer a path toward collaboration rather than

conflict. They enable users to scrutinize, challenge, and understand the systems that affect their lives. And in doing so, they help ensure that as AI advances, it does so in a way that strengthens, rather than erodes, the relationship between humans and the technologies they create.

This study began by asking the question of how we can improve the trustworthiness and perceived ethicality of AI systems, because they are critical to human-AI interactions. While this study particularly focused on the role of transparency and cognitive load in an AI-assisted decision-making environment, there exist multiple areas to explore what factors may shape human-AI interaction spaces. As AI systems experience technological advancement at an unprecedented pace, future economic and societal prosperity heavily relies on how humans effectively collaborate with AI systems. Without fully understanding what factors influence human interactions with AI systems, the human quest of AI-based economy and society is left incomplete. Therefore, scholars and practitioners alike should work together to further explore human-AI interaction spaces.

CONFLICT OF INTEREST

The author declares no conflicts of interest related to this work.

REFERENCES

1. Kumar U. Why explainable AI is becoming essential to clinical decision-making. ET Edge Insights. 2025;
2. Miller BT. But Why ? Understanding Explainable Artificial Intelligence. *XRDS Crossroads, ACM Mag Students*. 2019; 25 (3): 20-5. https://doi.org/10.1145/3313107
3. Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux; 2011.
4. Buçinca Z, Malaya MB, Gajos KZ. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. In: ACM on Human-computer Interaction. 2021; p. Article 188. https://doi.org/10.1145/3449287
5. Awad E, Dsouza S, Kim R, Schulz J, *et al*. The Moral Machine Experiment : 40 Million Decisions and the Path to Universal Machine Ethics. *Nature*. 2018; 7729 (October): 59-64. https://doi.org/10.1038/s41586-018-0637-6
6. Langer M, Oster D, Speith T, Hermanns H, *et al*. What do we want from Explainable Artificial Intelligence (XAI)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif Intell [Internet]*. 2021; 296 (2021): 103473. Available from: https://doi.org/10.1016/j.artint.2021.103473
7. Kizilcec RF. How Much Information ? Effects of Transparency on Trust in an Algorithmic Interface. In: CHI Conference on Human Factors in Computing Systems. 2016; p.1-6. https://doi.org/10.1145/2858036.2858402
8. Binns R, Kleek M Van, Veale M, Lyngs U, *et al*. It's Reducing a Human Being to a Percentage; Perceptions of Justice in Algorithmic Decisions. 2018; https://doi.org/10.31235/osf.io/9wqxr
9. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H. Manipulating and Measuring Model Interpretability. In: CHI Conference on Human Factors in Computing Systems. 2021; p.1-67. https://doi.org/10.1145/3411764.3445315
10. Ananny M, Crawford K. Seeing without knowing : Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc*. 2018; 20 (3): 973-989. https://doi.org/10.1177/1461444816676645
11. Floridi L, Cowls J. A Unified Framework of Five Principles for AI in Society. *Harvard Data Sci Rev*. 2019; 1 (1): 1-14. https://doi.org/10.2139/ssrn.3831321
12. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. *Internaltional J Human-Computer Stud*. 2003; 58: 697-718. https://doi.org/10.1016/S1071-5819(03)00038-7
13. Risko EF, Gilbert SJ. Cognitive Offloading. *Trends Cogn Sci*. 2016; 20 (9): 676-88. https://doi.org/10.1016/j.tics.2016.07.002
14. Choung H, David P, Ross A. Trust and ethics in AI. *AI Soc [Internet]*. 2022; 38 (2): 733-745. Available from: https://doi.org/10.1007/s00146-022-01473-4
15. Shin D. The effects of explainability and causability on perception , trust , and acceptance : Implications for explainable AI. *Int J Hum - Comput Stud [Internet]*. 2021; 146 (April 2020): 1-10. Available from: https://doi.org/10.1016/j.ijhcs.2020.102551
16. Eslami M, Kumaran SRK, Sandvig C, Karahalios K. Communicating Algorithmic Process in Online Behavioral Advertising. In: CHI Conference on Human Factors in Computing Systems. 2018; p.1-13. https://doi.org/10.1145/3173574.3174006
17. Welkowitz J, Cohen BH, Ewen RB. Introductory Statistics for the Behavioral Sciences. 6th ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2025.
18. Malle BF. Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics Inf Technol*. 2016; 18 (4): 243-56. https://doi.org/10.1007/s10676-015-9367-8

APPENDIX I. AI-ASSISTED DECISION-MAKING SCENARIOS

Here in the appendix, the examples of one decision-making scenario, load application, are presented. To generate the experiment vignettes, the author prompted a generative AI tool, ChatGPT, to generate hypothetical decision-making scenarios in a loan application with varying levels of explanations. The AI tool was used to ensure that the explanation styles of AI-assisted decision scenarios look as close as possible to real AI prompts, and thus increase the realism of vignettes. To ensure that any piece of personal profile used in the scenarios does not distort participants' perception of ethicality and trustworthiness, the author asked the tool to create a neutral profile without, for example, any racial or educational information susceptible to stereotype.

The following are the three (None, Simple, Detailed) scenarios presented to each group. Cognitive load was manipulated by the code presented at the top of the scenario page, which participants were asked to memorize while reading the scenario.

Example 1. Scenario for No Explanation Groups

Code: 36 (for low cognitive load group) / Code: A7K2941 (for high cognitive load group)

Please remember the code you see above while reading the passage below. You'll answer 1–2 questions about the code and this passage.

Scenario: An AI system was used by a financial institution to review loan applications and make lending recommendations for applicants.

Applicant: Jordan M., 34 years old, works full-time as an office administrator and has been employed in the same company for five years. Jordan recently applied for a $15,000 personal loan to consolidate credit-card debt.

System Recommendation: The AI system recommended rejecting the current applicant's loan request.

System Explanation: No additional information about how the recommendation was made was provided.

This screen will advance automatically after a short time. But you can also proceed voluntarily by clicking NEXT.

Example 2. Scenario for Simple Explanation Groups

Code: 36 (for low cognitive load group) / Code: A7K2941 (for high cognitive load group)

Please remember the code you see above while reading the passage below. You'll answer 1–2 questions about the code and this passage.

Scenario: An AI system was used by a financial institution to review loan applications and make lending recommendations for applicants.

Applicant: Jordan M., 34 years old, works full-time as an office administrator and has been employed in the same company for five years. Jordan recently applied for a $15 000 personal loan to consolidate credit-card debt.

System Recommendation: The AI system recommended rejecting the current applicant's loan request.

System Explanation: The system analyzed income stability and existing debt level to estimate repayment ability. Its recommendation was based on data patterns learned from many past loan outcomes. The model's accuracy was periodically checked by bank analysts to ensure reliable performance.

This screen will advance automatically after a short time. But you can also proceed voluntarily by clicking NEXT.

Example 3. Scenario for Detailed Explanation Groups

Code: 36 (for low cognitive load group) / Code: A7K2941 (for high cognitive load group)

Please remember the code you see above while reading the passage below. You'll answer 1–2 questions about the code and this passage.

Scenario: An AI system was used by a financial institution to review loan applications and make lending recommendations for applicants.

Applicant: Jordan M., 34 years old, works full-time as an office administrator and has been employed in the same company for five years. Jordan recently applied for a $15,000 personal loan to consolidate credit-card debt.

System Recommendation: The AI system recommended rejecting the current applicant's loan request.

System Explanation: The AI reviewed the applicant's income stability, debt-to-income ratio, and credit-repayment history. It used a model trained on about 50,000 previous loan decisions and is regularly audited for fairness and accuracy. The system calculated a predicted default risk of 22%, which exceeded the bank's approval threshold. To promote fairness, the model adjusted decision cut-offs to balance false approvals and false rejections across age and gender groups.

This screen will advance automatically after a short time. But you can also proceed voluntarily by clicking NEXT.

APPENDIX II. MEASURES OF PERCEIVED ETHICALITY AND TRUSTWORTHINESS OF AI SYSTEMS

Perceived Ethicality revised from Human-Like Trust in AI in Choung *et al.* (2022) (6 items; α = .92 in original study)

	Revised Item	Original Item
1	The AI system cares about the well-being of the people affected by its recommendations.	"Smart technologies care about our well-being."
2	The AI system seems sincerely concerned about providing fair and appropriate outcomes.	"Smart technologies are sincerely concerned about addressing the problems of human users."
3	The AI system tries to be helpful and does not operate out of self-interest.	"Smart technologies try to be helpful and do not operate out of selfish interest."
4	The AI system is truthful in the information it provides or uses.	"Smart technologies are truthful in their dealings."
5	The AI system keeps its commitments and delivers consistent recommendations.	"Smart technologies keep their commitments and deliver on their promises."
6	The AI system behaves honestly and would not misuse information about users.	"Smart technologies are honest and do not abuse the information and advantage they have over users."

Perceived Trustworthiness revised from Functionality Trust in AI in Choung *et al.* (2022) (5 items; α = .91 in original study)

	Revised Item	Original Item
1	The AI system works well when making decisions like this one.	"Smart technologies work well."
2	The AI system has the capabilities needed to make accurate recommendations.	"Smart technologies have the features necessary to complete key tasks."
3	The AI system is competent in evaluating this type of decision.	"Smart technologies are competent in their area of expertise."
4	The AI system is reliable in its decision-making process.	"Smart technologies are reliable."
5	The AI system is dependable when providing recommendations.	"Smart technologies are dependable."