Scientific Reports

https://doi.org/10.1038/s41598-026-55728-x

Article in Press

# Reported trust varies with graded value alignment in AI-attributed economic–environmental choices

Received: 5 February 2026

Accepted: 26 May 2026

Published online: 28 May 2026

Cite this article as: Cui L., Sun L. & He G. Reported trust varies with graded value alignment in AI-attributed economic–environmental choices. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-55728-x

Lidan Cui, Lingyun Sun & Guibing He

# Reported trust varies with graded value alignment in AI-attributed economic–environmental choices

**Lidan Cui**[1,2], **Lingyun Sun**[2], **Guibing He**[1,*]

[1]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou 310058, China
[2]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[*]Corresponding author: Guibing He (gbhe@zju.edu.cn)

**Abstract**

As artificial intelligence (AI) systems increasingly support value-laden decisions, users may judge whether an advisor's observable choices align with their own priorities. Prior theories of value similarity predict that value alignment should support trust, but less is known about whether self-reported trust varies across graded, individually calibrated levels of objective value alignment. We tested this question in a controlled economic–environmental trade-off task with 250 participants. Participants first completed matching and titration tasks to estimate their subjective equivalence point between economic gain and additional good-air-quality days. They then observed five pre-programmed choices attributed to a simulated AI advisor. These choices varied by objective value alignment, with four levels defined by distance from each participant's equivalence point, and by decision orientation, economy-leaning versus environment-leaning. Reported trust increased monotonically as AI-attributed choices moved closer to participants' calibrated trade-off points. Perceived value alignment also increased across objective value alignment levels and was closely associated with trust, although this post-task association should not be interpreted as causal mediation. Environment-leaning choices showed a small, less robust trust advantage. These findings characterize objective value alignment as a graded cue for self-reported trust in a controlled AI-attributed decision scenario.

## Introduction

Artificial intelligence (AI) systems are increasingly integrated into human decision-making across domains such as smart city construction [1], scientific research [2], financial risk assessment [3], medical diagnostics [4], and autonomous driving [5]. By leveraging large-scale data and advanced learning algorithms, these systems can enhance decision processes and, in some contexts, improve decision quality and efficiency [6]. However, the real-world benefits of AI depend not only on technical performance but also on whether people are willing to adopt and appropriately rely on AI recommendations [7,8].

Appropriate trust is essential to the acceptance and effective deployment of AI systems [8–10]. Trust shapes whether people follow, ignore, or over-rely on AI recommendations, often beyond what can be explained by objective system performance alone [11]. Miscalibrated trust poses clear risks: insufficient trust can lead to underuse of beneficial AI decision support, whereas excessive trust can produce uncritical acceptance of flawed recommendations, potentially resulting in harmful outcomes [12–15]. As AI systems gain autonomy in high-stakes settings, understanding how trust forms—and how it can be calibrated—remains an important research concern [16].

One difficulty in forming trust is that AI decision-making can be opaque. Machine-learning systems can be difficult to interpret because their internal decision logic may be inaccessible, technically complex, or hard to map onto human-understandable reasons [17,18]. Explainable-AI research has therefore treated interpretability as relevant to whether users can decide when to trust a model's predictions [19]. Trust in automation is also shaped by information about system performance, process, and purpose, as well as by experience with system

behavior over time [12,13]. In value-laden decision contexts, an advisor's observable choices can be especially informative because they reveal how competing priorities are weighted.

This behavioral-cue account connects with established value-based theories of trust. Trust theory links integrity to the perception that a trustee adheres to principles acceptable to the trustor [20], and the Salient Value Similarity (SVS) model links social trust to perceived overlap in salient values [21,22]. Recent human–machine trust theory similarly defines value congruence in terms of whether an advisor weights evidence, constraints, and objectives in a way that coheres with the trustor's value hierarchy [23]. These perspectives provide a theoretical basis for expecting users to evaluate a simulated AI advisor partly through the value priorities implied by its decision pattern.

Concerns about value alignment have also become prominent in AI research and public debate. Documented risks of toxic, biased, and ethically problematic AI outputs have intensified questions about whether AI systems act in ways compatible with human values [24–27]. In this context, value alignment has become a central idea in human-centered AI: AI systems should make decisions and recommendations that are compatible with human values [28–31]. For experimental psychology, this broader agenda raises a more specific question: how do users translate observable value-related behavior into trust judgments in concrete decision contexts?

Prior empirical work provides important evidence that value similarity is associated with trust in artificial agents and automation. For example, studies have examined human-agent value similarity in scenario-based interaction and moral value similarity in autonomous-car and medical-automation scenarios [32,33]. This

work supports the general expectation that value alignment matters for trust. Building on this work, the present study operationalizes value alignment as an objective, individualized distance from each participant's own behavioral trade-off point and examines whether reported trust varies across ordered degrees of alignment within the same decision domain.

To implement this approach, we operationalized objective value alignment as the experimentally manipulated distance between the advisor's apparent choice threshold and each participant's individualized economic–environmental equivalence point. Participants first completed matching and titration tasks that estimated how they traded off economic benefits against improvements in air quality. The manipulation varied the advisor's apparent choice threshold relative to the participant's calibrated threshold, producing four ordered observable-alignment levels. The highest alignment level represented the closest tested proximity to the participant's equivalence point rather than complete agreement. This design allowed us to test whether reported trust tracks a graded, participant-calibrated observable alignment cue rather than a binary matched-versus-mismatched contrast.

The economic–environmental domain is well suited to this purpose because it is both practically relevant and methodologically tractable. Sustainable-development research treats economic development and environmental protection as central priorities that are often evaluated together [34], and public opinion research has long measured preferences using direct environment-versus-economic-growth trade-off items [35,36]. Similar issues also arise in smart-city and sustainability-related decision support, where AI applications may affect both economic efficiency and environmental outcomes [1]. At the same time, the domain

allows preferences to be expressed through quantifiable outcomes, making it possible to estimate each participant’s trade-off context and construct objective value alignment levels in a controlled way.

In this experiment, participants observed five pre-programmed economic–environmental choices attributed to a simulated AI advisor. The choices were experimenter-controlled stimuli rather than outputs from an interactive or adaptive AI system. Each participant was randomly assigned to one of eight conditions in a 4 × 2 design, crossing objective value alignment level with decision orientation. After observing the AI-attributed choices, participants reported perceived value alignment and trust in the advisor.

Perceived value alignment was included to capture participants' subjective interpretation of the AI-attributed decision pattern. Although objective value alignment was experimentally constructed around each participant’s calibrated threshold, users' trust judgments may depend on whether they register the advisor's choices as consistent with their own priorities. Measuring perceived alignment therefore complemented the objective manipulation by indicating how participants interpreted the advisor's decision pattern.



Taken together, this study contributes to the value-alignment and trust literature by providing a graded and individualized operationalization of objective value alignment in a quantifiable economic–environmental trade-off domain. By calibrating AI-attributed choices to each participant's equivalence point, the design builds on prior work using broad perceived similarity or categorical match-mismatch comparisons and examines whether reported trust varies with incremental proximity between the advisor's decision pattern and the participant's own trade-off point. The focus is therefore on graded, participant-specific

proximity rather than on a broad binary contrast between aligned and misaligned values.

## Results

### Manipulation Check for Perceived Value Alignment

Participants' post-task perceived value alignment increased across the four objective value alignment levels (Figure 1). A 4 × 2 analysis of variance (ANOVA) on perceived value alignment showed a significant main effect of objective value alignment, $F(3, 242) = 77.45$, $p < 0.001$, $\eta^2_p = 0.490$. Bayesian model comparison provided decisive evidence for this effect ($BF_{10} = 1.05 \times 10^{32}$). The main effect of decision orientation was not significant, $F(1, 242) = 0.32$, $p = 0.574$, $\eta^2_p = 0.001$, $BF_{01} = 6.55$, and the objective value alignment × decision orientation interaction was not significant, $F(3, 242) = 1.56$, $p = 0.200$, $\eta^2_p = 0.019$, $BF_{01} = 3.91$. Tukey-adjusted post hoc comparisons indicated that perceived value alignment increased from Level 1 to Level 4, with each successive level differing significantly from the previous one (all adjusted $p < 0.05$; full pairwise statistics are reported in Supplementary Table S1).

These results indicate that participants registered the intended ordering of the AI-attributed decision patterns. Perceived value alignment reflected participants' subjective interpretation of the alignment manipulation, so this measure is treated here as a manipulation check.

### Trust Across Objective Value Alignment Levels and Decision Orientation

Trust in the AI advisor increased monotonically as the AI-attributed choices

moved closer to participants' calibrated economic–environmental equivalence points (Table 1, Figure 2). Perceived familiarity with AI was measured after participants observed the advisor's decisions, so the analysis of covariance (ANCOVA) controlled only for age and gender. The homogeneity-of-regression-slopes check for age was not significant, $F(7, 233) = 1.46$, $p = 0.184$. The ANCOVA showed a significant main effect of objective value alignment on trust, $F(3, 240) = 39.34$, $p < 0.001$, $\eta^2_p = 0.330$. Bayesian model comparison provided decisive evidence for including objective value alignment beyond decision orientation, age, and gender, $BF_{10} = 8.56 \times 10^{17}$. Descriptive statistics for each condition are shown in Table 1.

Tukey-adjusted pairwise comparisons supported the monotonic pattern: all non-adjacent comparisons were significant, and the Level 3-Level 2 and Level 4-Level 3 contrasts were significant. The Level 2-Level 1 contrast was positive but did not reach significance after Tukey adjustment, $p = 0.063$ (Supplementary Table S2 for full post hoc results).

The primary model also showed a small decision-orientation effect, although subsequent sensitivity analyses indicated that this effect was less robust. Across objective value alignment levels, participants reported higher trust in environment-leaning decision patterns than in economy-leaning decision patterns, $F(1, 240) = 5.29$, $p = 0.022$, $\eta^2_p = 0.022$. Bayesian evidence for including decision orientation beyond objective value alignment and covariates was weak, $BF_{10} = 1.47$.

The interaction between objective value alignment level and decision orientation was not significant, $F(3, 240) = 1.40$, $p = 0.242$, $\eta^2_p = 0.017$; Bayesian model comparison favored the additive model over the interaction model, $BF_{01} = 4.72$. Thus, there was no clear evidence that the alignment effect differed by

orientation in this sample.

Age and gender were not significant predictors in the primary ANCOVA (age: $F(1, 240) = 0.30$, $p = 0.582$, $\eta^2_p = 0.001$; gender: $F(1, 240) = 1.82$, $p = 0.178$, $\eta^2_p = 0.008$).

In a full factorial sensitivity model that added post-task perceived familiarity with AI, the objective value alignment effect remained significant (Supplementary Table S3), $F(3, 239) = 40.21$, $p < 0.001$, $\eta^2_p = 0.335$, $BF_{10} = 2.54 \times 10^{18}$. Decision orientation also remained significant, $F(1, 239) = 6.39$, p = 0.012, $\eta^2_p = 0.026$, $BF_{10} = 2.54$. The objective value alignment × decision orientation interaction was not significant, $F(3, 239) = 1.32$, $p = 0.270$, $\eta^2_p = 0.016$, $BF_{01} = 5.17$. Post-task perceived familiarity was positively associated with trust, $F(1, 239) = 8.04$, $p = 0.005$, $\eta^2_p = 0.033$, $BF_{10} = 8.51$. We did not use this model for primary inference because familiarity was measured after the task and its homogeneity-of-regression-slopes check suggested possible nonparallel slopes, $F(7, 232) = 2.64$, $p = 0.012$.

A second full factorial sensitivity model excluded the reverse-coded trust item (Supplementary Table S3). The objective value alignment effect remained, $F(3, 240) = 43.94$, $p < 0.001$, $\eta^2_p = 0.355$, $BF_{10} = 5.79 \times 10^{19}$. The small orientation effect was attenuated, $F(1, 240) = 3.21$, $p = 0.075$, $\eta^2_p = 0.013$, $BF_{01} = 1.80$. The interaction remained unsupported, $F(3, 240) = 1.72$, $p = 0.163$, $\eta^2_p = 0.021$, $BF_{01} = 3.18$. Balance, exclusion, and model-assumption checks are reported in Supplementary Table S4. Together, the sensitivity analyses indicated that the graded alignment effect was robust to including post-task familiarity and to excluding the reverse-coded trust item, whereas the smaller decision-orientation effect was less stable.

## Monotonic Trend in Reported Trust Across Objective Value Alignment Levels

To examine whether the monotonic alignment-trust pattern was consistent with linearity, we fitted sequential polynomial regression models. In the linear model, objective value alignment was positively associated with trust after controlling for decision orientation, age, and gender, $B = 0.49$, $SE = 0.04$, $t(245) = 10.86$, $p < 0.001$. This model explained 34.3% of the variance in trust. Adding quadratic and cubic terms did not meaningfully improve fit (quadratic: $\Delta R^2 = 0.002$, $p = 0.412$; cubic: $\Delta R^2 < 0.001$, $p = 0.717$). Bayesian Information Criterion (BIC)-approximated Bayes factors favored the simpler trend models (Supplementary Table S5). Given the four ordered alignment levels and bounded Likert outcome, these analyses indicate that the data were consistent with a linear pattern within the tested range, but they do not establish the functional form definitively.



## Exploratory Analysis of Perceived Value Alignment and Trust

As an exploratory descriptive check, perceived value alignment covaried with both the objective value alignment manipulation and reported trust. Because perceived alignment and trust were measured in the same post-task questionnaire, and because perceived alignment also served partly as a manipulation-check measure, this analysis was not interpreted as temporal or causal mediation. The indirect association was significant in a bootstrap analysis with 5,000 resamples, $ab = 0.390$, 95% CI [0.296, 0.489]. The full descriptive model is reported in Supplementary Table S6.

## Discussion

Using an economic–environmental decision task, we found three main patterns. First, reported trust increased as pre-programmed choices attributed to a simulated AI advisor moved closer to participants' calibrated trade-off points. Second, perceived value alignment tracked the intended ordering of the manipulation and was closely associated with reported trust, although this exploratory association should not be interpreted as causal mediation. Third, environment-leaning decision patterns received slightly higher trust than economy-leaning patterns, but this effect was small and less robust across sensitivity analyses. Overall, the findings characterize objective value alignment as a graded cue for self-reported trust within a controlled AI-attributed decision scenario.

The increase in reported trust across higher levels of objective value alignment is consistent with established work linking shared values, integrity, trustworthiness, and human–machine value alignment to trust [20,22,23,32,33,37]. The present study adds to this literature by operationalizing value alignment as a graded, participant-specific distance from each individual's economic–environmental trade-off point. This allowed reported trust to be examined across ordered alignment cues rather than only through broad perceived similarity or a binary match–mismatch comparison.

This pattern suggests that, in this controlled task, reported trust was sensitive to ordered differences in objective value alignment. Economic–environmental trade-offs are a useful setting because they connect environmental protection with economic development, two value domains central to sustainability research [34]. In this experiment, the trade-off could also be represented numerically, allowing choices to be calibrated around each participant's own

equivalence point. In practical decision-support settings, alignment may often be a matter of degree rather than a perfect match. The present results suggest that users can be sensitive to these incremental differences when the relevant trade-off is concrete, repeated, and easy to compare with their own preferences.

Perceived value alignment provides information about how participants understood the choice sequence. It tracked the objective value alignment manipulation and was closely associated with trust, which is consistent with participants noticing the value-relevant pattern in the AI-attributed choices. At the same time, perceived alignment and trust were measured in the same post-task questionnaire and are conceptually close. The indirect-association analysis is therefore best read as descriptive evidence that perceived alignment covaried with both the manipulation and trust, not as a causal mediation test.

The association between perceived alignment and trust is consistent with the possibility that value legibility matters: users may need enough behavioral or explanatory cues to form a coherent impression of what a system appears to prioritize. Work on explanation and interpretability shows that clearer decision logic can support users' understanding of AI systems and collaboration with them [38–40]. In the present study, this point remains a hypothesis rather than a directly tested mechanism, because explanation, transparency, and value-legibility cues were not manipulated. Future work should directly test whether making value-relevant trade-offs more transparent improves users' ability to evaluate system alignment and calibrate reliance.

Several features of the data and design constrain interpretation. First, the monotonic alignment–trust pattern should not be read as a precise functional form. The four alignment levels and bounded Likert trust measure limit the ability to

distinguish a linear pattern from threshold, saturation, or other monotonic patterns. The data are therefore most appropriately interpreted as showing a graded increase within the tested range. Studies with more alignment levels, repeated observations, and behavioral outcomes would be needed to characterize the shape of this relationship more precisely.

Second, the decision-orientation effect was smaller and more context-dependent than the alignment effect. Participants reported slightly higher trust in environment-leaning than economy-leaning AI-attributed choices, but this effect was attenuated when the reverse-coded trust item was excluded. Environmental protection may have carried stronger moral salience in this task, consistent with work on protected or sacred values [41–43], and the asymmetric air-quality framing may have made environmental benefits more vivid than economic benefits. Broader public concern about climate and environmental protection provides a relevant contextual backdrop for this possibility [35,36,44]. For these reasons, the orientation result is better viewed as a context-bound baseline difference rather than a general preference for environment-leaning AI-attributed choices.

Third, the calibration results also qualify the interpretation of individualization. The matching value $Y$ varied widely across participants, whereas the final day-based equivalence point $Z$ was concentrated near the center of the titration range (Supplementary Table S8 and Supplementary Figure S1). Thus, the procedure individualized the monetary scale more strongly than it individualized the final $Z$ values. The alignment manipulation was still calibrated around each participant's trade-off context, but the final equivalence points were less dispersed than the initial monetary valuations.

These findings offer a focused theoretical and applied contribution. Theoretically, they extend established value-congruence accounts of trust to a controlled AI-attributed decision context by operationalizing objective value alignment as graded, participant-calibrated proximity to an individual trade-off point. The association between perceived alignment and trust further suggests that value legibility may matter: objective value alignment may inform trust most clearly when users can infer what a system appears to prioritize. Applied to AI decision support, the findings suggest that alignment should not be treated only as a binary property of matched versus mismatched values. Designers may instead benefit from mechanisms that reduce value distance incrementally, communicate the trade-offs behind recommendations, and allow users to inspect or adjust value-relevant parameters. Such design features may help users evaluate the system's value stance more clearly, although future work is needed to test whether they improve calibrated reliance in interactive AI systems.

Several limitations should be noted. First, the task limits ecological validity. Participants evaluated a short sequence of pre-programmed choices attributed to a simulated AI advisor; they did not interact with an actual AI model, face uncertainty, observe adaptive behavior, or receive outcome feedback. Trust was measured as a post-task self-report rather than as behavioral reliance during ongoing human–AI collaboration, a distinction emphasized in automation research [12–15]. Although AI-attributed or vignette-style paradigms can be useful for isolating responses to controlled decision cues [45], they do not reproduce the dynamics of real AI use, where performance feedback, errors, opacity, and repeated interaction may shape trust differently.

Second, the alignment manipulation bundled several sequence-related cues. Because choices appeared in a fixed order, value distance, agreement frequency, and timing of disagreement were not fully separable. The results should therefore be interpreted as reflecting responses to graded observable-alignment sequences, rather than as isolating the effect of value distance alone. Although the sequence of agreement and disagreement was mirrored across economy-leaning and environment-leaning conditions, future studies should randomize decision order or independently manipulate agreement frequency and agreement position.

Third, generalizability is limited by the sample and decision domain. Participants were recruited from a Chinese online platform, and perceived AI alignment may vary across cultural contexts and may not always track objective value alignment in the same way [46]. The observed association between objective and perceived alignment should therefore be tested in more culturally diverse samples. The study also focused on a single economic–environmental value domain. Although this domain is useful because trade-offs can be quantified and individualized, human values are multidimensional and may conflict in less easily quantifiable ways.



Finally, since the study was not preregistered, the indirect-association, polynomial, and sensitivity analyses should be treated as exploratory. Future preregistered studies should test whether the present graded alignment–trust pattern generalizes to interactive AI systems, behavioral reliance outcomes, and other value domains.

Overall, the findings support a view of value alignment as an observable, graded cue for self-reported trust in an AI-attributed decision task. Reported trust increased across higher objective-alignment levels in simulated advisor choice

sequences, perceived value alignment tracked this manipulation, and decision orientation showed a small, context-bound association with trust. Future preregistered studies using interactive AI systems, behavioral reliance measures, outcome feedback, and culturally diverse samples are needed to test how far this pattern generalizes.

# Methods

## Experimental Design and Participants

We used a 4 × 2 between-subjects factorial design to examine how objective value alignment relates to reported trust in an AI-attributed decision task. The dependent measure was trust in the AI advisor. The independent variables were objective value alignment (four levels, from Level 1 to Level 4) and AI decision orientation (economy-leaning versus environment-leaning). In this study, objective value alignment refers to the experimentally imposed proximity between the advisor's apparent choice threshold and the participant's calibrated economic–environmental trade-off threshold. Participants' perceived value alignment with the advisor was assessed after the task and served as the manipulation-check measure and as an exploratory association variable; analyses involving this measure were treated as descriptive and non-causal.

Sample size was determined a priori using G*Power 3.1 [47]. For our two-factor analysis of variance design with a significance level of $\alpha = 0.05$, statistical power of $1\text{-}\beta = 0.90$, and a medium effect size of $f = 0.25$ [48], the minimum required sample size was 231 participants to ensure detection of both main effects and interaction effects.

We recruited 280 participants through Credamo (www.credamo.com), a data collection platform functionally similar to Amazon Mechanical Turk. Credamo integrates with WeChat and requires account authentication, which minimizes the risk of duplicate or automated (bot) responses. Participants were randomly assigned to one of eight experimental conditions (35 per condition). After excluding participants who failed attention checks (involving recall of AI decision patterns), the final analysis included 250 participants (88 males; $M_{age}$ = 28.66, $SD$ = 5.82, range: 19-49 years), representing an 89.29% valid response rate. Exclusions by condition were as follows: economy-leaning Level 1 = 4, Level 2 = 5, Level 3 = 3, and Level 4 = 2; environment-leaning Level 1 = 7, Level 2 = 4, Level 3 = 2, and Level 4 = 3. Exclusions did not show clear imbalance across the eight cells, $\chi^2(7) = 5.20$, $p = 0.636$. A Monte Carlo Fisher exact test gave the same conclusion, $p = 0.639$ (Supplementary Table S7).



## Ethics approval and consent to participate

This work was approved by the Ethics Committee of the Department of Psychology and Behavioral Sciences at Zhejiang University (Approval Number: 027, Date: February 29, 2024) and conducted in accordance with the principles of the Declaration of Helsinki. Informed consent was obtained from all participants prior to participation.

## Experimental Procedure

The experiment proceeded in three sequential phases (Figure 3).

*Phase 1: Determining Participants' Subjective Equivalence Point for Economic–Environmental Values*

To precisely quantify participants' relative weighting of economic versus environmental values, we employed matching and titration procedures commonly used to estimate subjective equivalence points in judgment and decision-making research. The matching task elicited each participant's monetary value Y for 35 good air quality days, and the titration task used this individualized Y value to estimate a day-based reversal point Z. This Z value served as the participant's subjective equivalence point, indicating the threshold at which economic and environmental considerations were perceived as equally valuable.

(1) Matching Task

First, participants provided a monetary matching value Y, defined as the economic benefit they judged subjectively equivalent to a 35-day increase in good air quality. This step individualized the monetary scale used in the subsequent titration task.



The task was described as follows (see Supplementary Note: Economic–Environmental Matching Scale for detail):

*Imagine you are a decision-maker for a city who needs to evaluate trade-offs between economic development and environmental protection. Please judge how much economic growth in Plan B would make it subjectively equivalent to Plan A. Fill in the appropriate number on the blank line.*

***Plan A:*** *Increase the city's annual number of good air quality days by 35 days compared to before implementation, but with no change in annual economic benefits.*

***Plan B:*** *Increase the city's annual economic benefits by ____ 10-million-yuan units compared to before implementation, but with no change in annual good air quality days.*

(2) Titration Task

Building on the matching task, we used a titration task to more precisely measure each participant's subjective equivalence point. The titration task consisted of 15 decision pairs, requiring participants to choose one plan from each pair. The questionnaire structure was as follows (see Supplementary Note: Economic–Environmental Titration Scale for detail):

***Plan A:*** *Increase the city's annual number of good air quality days by X days compared to before implementation (X representing 15 values from 5 to 75, increasing in increments of 5), but with no change in annual economic benefits.*

***Plan B:*** *Increase the city's annual economic benefits by Y 10-million-yuan units compared to before implementation, but with no change in annual good air quality days (Y fixed as the value provided by the participant in the matching task).*



We quantified each participant's subjective equivalence point $Z$ by identifying the preference reversal point between Plan B and Plan A. As $X$ increased, we calculated $Z$ as the midpoint between the highest $X$ value at which the participant preferred Plan B and the lowest X value at which the participant preferred Plan A. This equivalence point represents the number of good air quality days perceived as equivalent to the specified economic benefit $Y$. For example, if a participant selected Plan B when offered $X \leq 40$ days but switched to Plan A when $X \geq 45$ days, we computed the subjective equivalence point as $Z = (40 + 45)/2 = 42.5$ days.

In the final sample, the matching value $Y$ was strongly right-skewed ($M$ = 550.81, $SD$ = 1933.60, median = 15.00, range = 1.30–15000.00, expressed in units of 10-million-yuan). This skewness was expected given that $Y$ was elicited through an open-ended matching task and reflected participants' subjective monetary valuation of 35 additional good air quality days. Some participants assigned very high monetary values to city-wide air-quality improvement; because these responses were task-consistent and the participants passed the attention checks, they were retained rather than treated as invalid outliers. The participant-level subjective equivalence point $Z$ was more narrowly distributed ($M$ = 37.36, $SD$ = 3.73, median = 37.50, range = 27.50–47.50 good air quality days; Supplementary Table S8 and Supplementary Figure S1).

This distribution should be interpreted in light of the calibration procedure. The matching task first elicited each participant's monetary value $Y$ for 35 good air quality days, and the subsequent titration task used that individualized $Y$ value to estimate a day-based reversal point $Z$. Thus, the procedure allowed the monetary scale to vary substantially across participants while using the titration task to locate each participant's economic–environmental equivalence point. The subsequent objective value alignment manipulation was calibrated relative to each participant's own $Z$ value, so the AI-attributed decision patterns remained participant-specific.

*Phase 2: Participants Observe AI-Attributed Choices*

We randomly assigned participants to one of eight experimental conditions. In each condition, participants observed five pre-programmed economic–environmental trade-off choices attributed to a simulated AI advisor in the same

context as Phase 1. The choices appeared sequentially on the computer screen and were generated by the experimental program according to the condition-specific parameters. The advisor was labeled “Artificial intelligence E” throughout the task, and the same label was used in the post-task questionnaire.

To manipulate the AI decision orientation and objective value alignment level, we established decision parameters for the simulated AI advisor across experimental conditions as follows:

First, we manipulated AI decision orientation by adjusting the number of good air quality days $X$ in plan A. Across the five decision tasks, the economic benefit in Plan B remained fixed at value $Y$ (provided by the participant in the matching method), while $X$ was set to five distinct values either above or below the participant’s subjective equivalence point $Z$.

In the environment-leaning condition, $X$ was set below $Z$, indicating that the AI’s equivalence point favored environmental values more strongly than the participant’s (i.e., the AI considered fewer good air quality days equivalent to economic benefit $Y$). Conversely, in the economy-leaning condition, $X$ was set above $Z$, signifying that the AI’s equivalence point favored economic values more prominently than the participant’s (i.e., the AI required more good air quality days to be equivalent to economic benefit $Y$).

Second, we manipulated objective value alignment by varying the distance between the AI’s inflection point and the participant’s inflection point across the five decisions. The magnitude of this difference inversely correlated with value alignment—larger differences represented lower alignment between human and AI values. Operationally, higher alignment levels placed the inferred advisor threshold progressively closer to the participant's threshold and, in the examples

shown in Tables 2 and 3, corresponded to 1, 2, 3, and 4 advisor choices matching the participant's implied preferences across the five displayed decisions. Thus, Level 4 represented the highest tested alignment level rather than complete agreement across all five choices.

To illustrate, consider a participant with a subjective equivalence point $Z$ of 42.5 days (inflection point between 40 and 45 days). In the environment-leaning AI condition, we set the number of good air quality days $X$ in Plan A to 20, 25, 30, 35, and 40 days across the five decisions. Under these parameters, participants would consistently select Plan B, demonstrating preference for increased annual economic benefits. Table 2 presents the systematic manipulation of the AI's choice inflection point across incremental objective value alignment levels for this condition.

Conversely, in the economy-leaning AI condition, we set the number of good air quality days $X$ in Plan A to 45, 50, 55, 60, and 65 days across the five decisions. Given these parameters, participants would select Plan A in all instances, indicating preference for increased annual good air quality days. Table 3 illustrates the manipulation of the AI's choice inflection point across incremental objective value alignment levels for this condition.



*Phase 3: Participants Complete Questionnaire Measures*

We measured participants' perceived value alignment with the AI advisor, trust in the AI advisor, and perceived familiarity through questionnaires (see Supplementary Table S9). All questionnaires used 5-point scoring. The perceived value alignment questionnaire (Cronbach's $\alpha = 0.93$) was adapted from Yokoi and Nakayachi [33]. The trust questionnaire (Cronbach's $\alpha = 0.89$) was adapted from

Verberne et al. [49] and Yokoi and Nakayachi [33]. The reverse-coded item referred specifically to the experimental advisor, labeled Artificial intelligence E. Additionally, attention tests concerning experimental scenario details were included to verify participants' engagement. These tests selected two economic–environmental trade-off decision pairs from Phase 2 and required participants to recall the AI's choices.

## Data Analysis

We conducted statistical analyses using JASP [50] and R [51] software.

First, we performed a 4 × 2 ANOVA on perceived value alignment with objective value alignment level and decision orientation as between-subjects factors. This analysis treated perceived value alignment as the post-task manipulation-check measure. Tukey-adjusted comparisons of estimated marginal means were used to compare objective value alignment levels collapsed across decision orientation.

Second, we performed two-factor analysis of covariance (ANCOVA) to examine differences in self-reported trust in the AI advisor under objective value alignment levels and AI decision orientations, while controlling for participants' demographic characteristics (gender and age). The primary model retained the objective value alignment × decision orientation factorial structure and tested whether the interaction was supported before interpreting the overall alignment and orientation patterns. We evaluated the homogeneity-of-regression-slopes assumption for the primary continuous covariate by adding condition-by-age interactions. Because perceived familiarity with AI was measured after the

experimental task, it was not included in the primary ANCOVA and was instead examined in a sensitivity analysis.

Third, we employed sequential polynomial regression analysis to examine whether the alignment-trust pattern was consistent with linearity within the tested range, controlling for decision orientation, gender, and age. Since objective value alignment had ordered levels from low to high, we coded it as a continuous variable from 1 to 4. We then constructed linear (first-order), quadratic (second-order), and cubic (third-order) models and evaluated whether higher-order terms improved fit using changes in explained variance, information criteria, nested model comparisons, and BIC-approximated Bayes factors. Because the design included only four ordered levels, these analyses described the observed monotonic pattern rather than definitively establishing the functional form.

Finally, we conducted an exploratory indirect-association analysis in R using the lavaan package [52] with bootstrap confidence intervals to describe how post-task perceived value alignment was related to objective value alignment and trust. Objective value alignment was entered as an ordered numeric predictor coded 1–4, so coefficients represent the expected change associated with a one-level increase in objective value alignment. The model used 5,000 bootstrap resamples while controlling for decision orientation, gender, and age. Because perceived value alignment also served as the manipulation-check measure and was measured in the same post-task questionnaire as trust, this analysis describes associations rather than temporal or causal ordering.

For ANOVA and ANCOVA models, we reported $F$ tests, $p$ values, and partial eta-squared ($\eta^2_p$). For selected pairwise comparisons, we reported Cohen's $d$ using the relevant model residual standard deviation. Regression and indirect-

association analyses included regression coefficients, model explained variance ($R^2$), and bootstrap confidence intervals for indirect associations. Bayes factors were reported as complementary evidence indices: $BF_{10}$ indicates evidence favoring inclusion of the tested effect or model, whereas $BF_{01}$ indicates evidence favoring the simpler model. Bayes factors for ANOVA and ANCOVA model comparisons were computed in R using JZS priors with the BayesFactor package, and Bayes factors for polynomial model comparisons were approximated from BIC values. BF values of 1–3 indicate anecdotal evidence, 3–10 moderate evidence, 10–30 strong evidence, 30–100 very strong evidence, and >100 extremely strong evidence. Unlike traditional p-values, which cannot quantify evidence favoring the null hypothesis, Bayes Factors provide direct quantification of relative evidence strength for competing hypotheses.

We also conducted sensitivity analyses to evaluate robustness. These included excluding the reverse-coded trust item, adding post-task perceived familiarity with AI as a covariate, adding log-transformed $Y$ as a covariate, and excluding participants with $Y$ values above the 95th percentile. The latter two checks were included because the matching value Y was highly right-skewed. Sensitivity-analysis results are reported in Supplementary Table S3 and were used to assess robustness rather than primary inference.

The study was not preregistered. Accordingly, the indirect-association analysis, polynomial model comparisons, and sensitivity checks were treated as exploratory or robustness analyses rather than confirmatory tests.

## Data Availability

The data are available on the Open Science Framework (OSF): https://osf.io/h4u2n/?view_only=4a3e60b8e507439587d3197cb6caee54.

## References


1. Yigitcanlar, T., Desouza, K. C., Butler, L. & Roozkhosh, F. Contributions and risks of artificial intelligence in building smarter cities: Insights from a systematic review of the literature. *Energies* **13**, 1473. https://doi.org/10.3390/en13061473 (2020).
2. Gao, J. & Wang, D. Quantifying the use and potential benefits of artificial intelligence in scientific research. *Nat. Hum. Behav.* **8**, 2281–2292. https://doi.org/10.1038/s41562-024-02020-5 (2024).
3. Mhlanga, D. Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *Int. J. Financ. Stud.* **9**, 39. https://doi.org/10.3390/ijfs9030039 (2021).
4. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56. https://doi.org/10.1038/s41591-018-0300-7 (2019).
5. Waldrop, M. M. Autonomous vehicles: No drivers required. *Nature* **518**, 20–23. https://doi.org/10.1038/518020a (2015).
6. Langer, M. & Landers, R. N. The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Comput. Human Behav.* **123**, 106878. https://doi.org/10.1016/j.chb.2021.106878 (2021).
7. Afroogh, S., Akbari, A., Malone, E., Kargar, M. & Alambeigi, H. Trust in AI: progress, challenges, and future directions. *Humanit. Soc. Sci. Commun.* **11**, 1568. https://doi.org/10.1057/s41599-024-04044-8 (2024).

8. Glikson, E. & Woolley, A. W. Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* **14**, 627–660. https://doi.org/10.5465/annals.2018.0057 (2020).
9. Habbal, A., Ali, M. K. & Abuzaraida, M. A. Artificial intelligence trust, risk and security management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Syst. Appl.* **240**, 122442. https://doi.org/10.1016/j.eswa.2023.122442 (2024).
10. Vanneste, B. S. & Puranam, P. Artificial intelligence, trust, and perceptions of agency. *Acad. Manag. Rev.* **50**, 726–744. https://doi.org/10.5465/amr.2022.0041 (2025).
11. Lee, M. K. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data Soc.* **5**, 1–16. https://doi.org/10.1177/2053951718756684 (2018).
12. Hoff, K. A. & Bashir, M. Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Factors* **57**, 407–434. https://doi.org/10.1177/0018720814547570 (2015).
13. Lee, J. D. & See, K. A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **46**, 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 (2004).
14. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114–126. https://doi.org/10.1037/xge0000033 (2015).
15. Parasuraman, R. & Riley, V. Humans and automation: Use, misuse, disuse, abuse. *Hum. Factors* **39**, 230–253. https://doi.org/10.1518/001872097778543886 (1997).

16. Lahusen, C., Maggetti, M. & Slavkovik, M. Trust, trustworthiness and AI governance. *Sci. Rep.* **14**, 20752. https://doi.org/10.1038/s41598-024-71761-0 (2024).
17. Burrell, J. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data Soc.* **3**, 2053951715622512. https://doi.org/10.1177/2053951715622512 (2016).
18. Von Eschenbach, W. J. Transparency and the black box problem: Why we do not trust AI. *Philos. Technol.* **34**, 1607–1622. https://doi.org/10.1007/s13347-021-00477-0 (2021).
19. Ribeiro, M. T., Singh, S. & Guestrin, C. ‘Why should I trust you?’: Explaining the predictions of any classifier. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144. https://doi.org/10.1145/2939672.2939778 (ACM, 2016).
20. Mayer, R. C., Davis, J. H. & Schoorman, F. D. An integrative model of organizational trust. *Acad. Manag. Rev.* **20**, 709–734. https://doi.org/10.5465/amr.1995.9508080335 (1995).
21. Earle, T. C. & Cvetkovich, G. T. *Social Trust: Toward a Cosmopolitan Society*. (Praeger Press, 1995).
22. Siegrist, M., Cvetkovich, G. & Roth, C. Salient value similarity, social trust, and risk/benefit perception. *Risk Anal.* **20**, 353–362. https://doi.org/10.1111/0272-4332.203034 (2000).
23. Prahl, A. Mortal vs. Machine: A compact two-factor model for comparing trust in humans and robots. *Robotics* **14**, 112. https://doi.org/10.3390/robotics14080112 (2025).

24. Gehman, S., Gururangan, S., Sap, M., Choi, Y. & Smith, N. A. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. in *Findings of the Association for Computational Linguistics: EMNLP 2020* 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301 (Association for Computational Linguistics, 2020).
25. Sheng, E., Chang, K.-W., Natarajan, P. & Peng, N. The woman worked as a babysitter: On biases in language generation. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* 3405–3410. https://doi.org/10.18653/v1/D19-1339 (Association for Computational Linguistics, 2019).
26. Si, W. M. *et al.* Why so toxic? Measuring and triggering toxic behavior in open-domain chatbots. in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* 2659–2673. https://doi.org/10.1145/3548606.3560599 (ACM, 2022).
27. Weidinger, L. *et al.* Ethical and social risks of harm from language models. Preprint at https://doi.org/10.48550/arXiv.2112.04359 (2021).
28. Gabriel, I. Artificial intelligence, values, and alignment. *Minds Mach.* **30**, 411–437. https://doi.org/10.1007/s11023-020-09539-2 (2020).
29. Gabriel, I. & Ghazavi, V. The challenge of value alignment: from fairer algorithms to AI safety. In *The Oxford Handbook of Digital Ethics* (ed. Véliz, C.). https://doi.org/10.1093/oxfordhb/9780198857815.013.18 (Oxford University Press, 2021).

30. Prasad, M. Social choice and the value alignment problem. In *Artificial Intelligence Safety and Security* (ed. Yampolskiy, R. V.) 291–314. https://doi.org/10.1201/9781351251389-21 (CRC Press, 2018).
31. Schmager, S., Pappas, I. O. & Vassilakopoulou, P. Understanding Human-Centred AI: a review of its defining elements and a research agenda. *Behav. Inf. Technol.* **44**, 3771–3810. https://doi.org/10.1080/0144929X.2024.2448719 (2025).
32. Mehrotra, S., Jonker, C. M. & Tielman, M. L. More similar values, more trust? - the effect of value similarity on trust in human-agent interaction. in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* 777–783. https://doi.org/10.1145/3461702.3462576 (ACM, 2021).
33. Yokoi, R. & Nakayachi, K. The effect of value similarity on trust in the automation systems: A case of transportation and medical care. *Int. J. Hum. Comput. Interact.* **37**, 1269–1282. https://doi.org/10.1080/10447318.2021.1876360 (2021).
34. Leiserowitz, A. A., Kates, R. W. & Parris, T. M. Sustainability values, attitudes, and behaviors: A review of multinational and global trends. *Annu. Rev. Environ. Resour.* **31**, 413–444. https://doi.org/10.1146/annurev.energy.31.102505.133552 (2006).
35. Pew Research Center. What the world thinks in 2002. https://www.pewresearch.org/global/2002/12/04/what-the-world-thinks-in-2002/ (2002).
36. Pew Research Center. Views of a changing world 2003. https://www.pewresearch.org/global/2003/06/03/views-of-a-changing-world-2003/ (2003).

37. Colquitt, J. A., Scott, B. A. & LePine, J. A. Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *J. Appl. Psychol.* **92**, 909–927. https://doi.org/10.1037/0021-9010.92.4.909 (2007).
38. Hoffman, R. R., Mueller, S. T., Klein, G. & Litman, J. Metrics for Explainable AI: Challenges and Prospects. Preprint at https://doi.org/10.48550/arXiv.1812.04608 (2019).
39. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267**, 1–38. https://doi.org/10.1016/j.artint.2018.07.007 (2019).
40. Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S. & Netland, T. Explainable AI improves task performance in human–AI collaboration. *Sci. Rep.* **14**, 31150. https://doi.org/10.1038/s41598-024-82501-9 (2024).
41. Baron, J. & Leshner, S. How serious are expressions of protected values? *J. Exp. Psychol. Appl.* **6**, 183–194. https://doi.org/10.1037/1076-898X.6.3.183 (2000).
42. Baron, J. & Spranca, M. Protected values. *Organ. Behav. Hum. Decis. Process.* **70**, 1–16. https://doi.org/10.1006/obhd.1997.2690 (1997).
43. Tetlock, P. E. Thinking the unthinkable: sacred values and taboo cognitions. *Trends Cogn. Sci.* **7**, 320–324. https://doi.org/10.1016/S1364-6613(03)00135-9 (2003).
44. UNDP Climate Promise. The world's largest survey on climate change is out – here's what the results show. https://climatepromise.undp.org/news-and-stories/worlds-largest-survey-climate-change-out-heres-what-results-show (2024).

45. Aguinis, H. & Bradley, K. J. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organ. Res. Methods* **17**, 351–371. https://doi.org/10.1177/1094428114547952 (2014).
46. Globig, L. K., Xu, R., Rathje, S. & Van Bavel, J. J. Perceived (mis)alignment in generative artificial intelligence varies across cultures. Preprint at https://doi.org/10.31234/osf.io/suqa2 (2024).
47. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191. https://doi.org/10.3758/BF03193146 (2007).
48. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. https://doi.org/10.4324/9780203771587 (Lawrence Erlbaum Associates, 1988).
49. Verberne, F. M. F., Ham, J. & Midden, C. J. H. Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Hum. Factors* **54**, 799–810. https://doi.org/10.1177/0018720812443825 (2012).
50. JASP Team. JASP (Version 0.19.1) [Computer software]. https://jasp-stats.org/ (2024).
51. R Core Team. R: A Language and Environment for Statistical Computing [Computer software]. https://www.R-project.org/ (2024).
52. Rosseel, Y. lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **48**, 1–36. https://doi.org/10.18637/jss.v048.i02 (2012).

## Acknowledgements

The authors would like to thank all participants for their contribution to this research.

## Funding

This work was supported by the National Natural Science Foundation of China (72571243) and the National Key Research and Development Program (2021ZD0200409).

## Author Contributions

L.C. contributed to the study design, performed the investigation and data collection, conducted the formal analysis, and prepared the visualization. L.S. provided supervision and contributed to the conceptualization and manuscript revision. G.H. conceptualized the study, oversaw the project administration, validated the data analysis, and contributed to the methodology. L.C. wrote the original draft, and all authors reviewed and edited the final manuscript.



## Additional Information

### Competing Interests

The authors declare no competing interests.

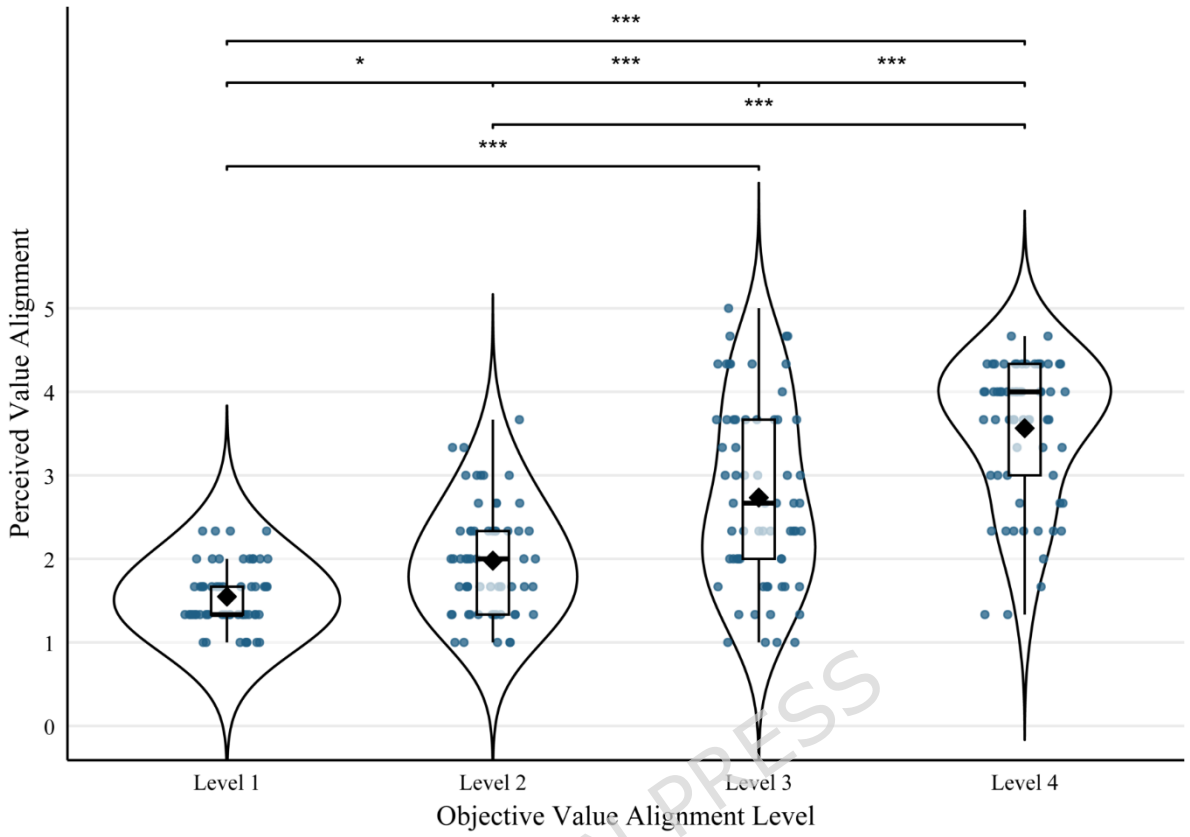


**Figure 1. Perceived value alignment across objective value alignment levels.** Violin and box plots show the distribution, with jittered points indicating individual observations. The central box indicates the interquartile range, and the horizontal line represents the median. Stars reflect selected Tukey-adjusted pairwise comparisons shown in the figure. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

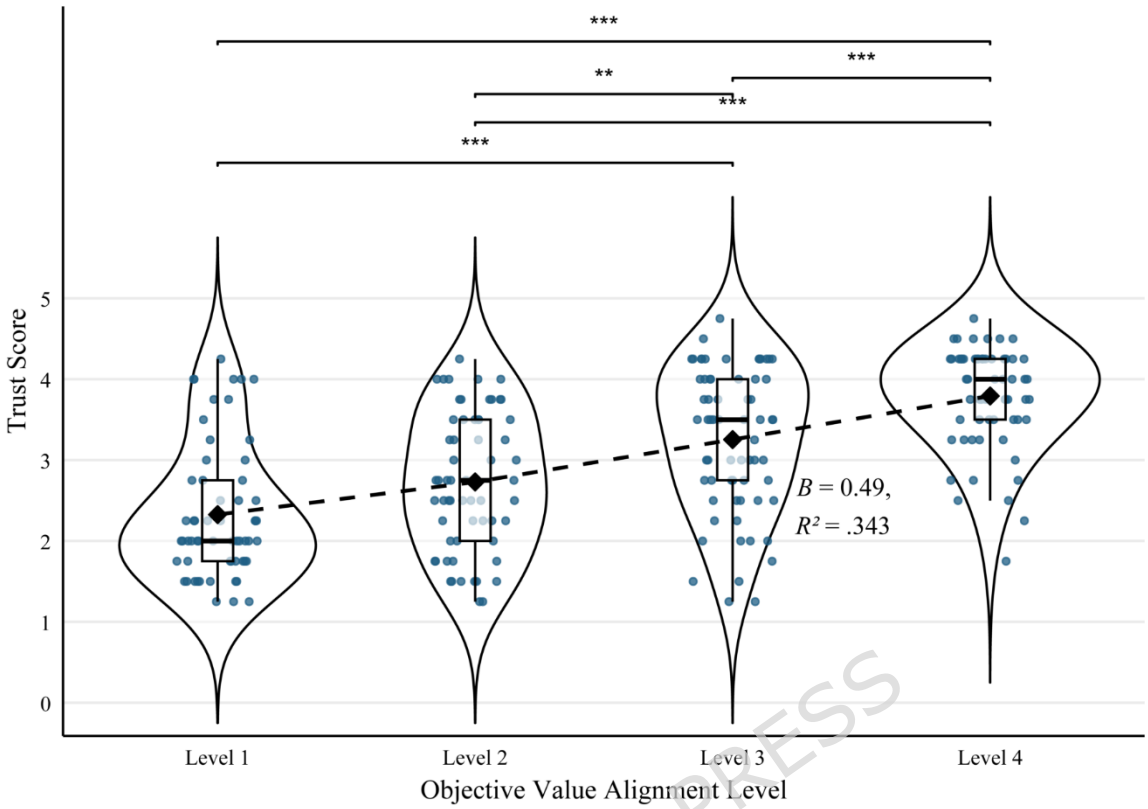


**Figure 2. Trust across objective value alignment levels.** Violin and box plots show the distribution of trust scores, with jittered points indicating individual observations. The central box indicates the interquartile range, and the horizontal line represents the median. The dashed line represents the fitted linear trend across the tested range, with the unstandardized regression coefficient ($B$) and explained variance ($R^2$) shown for reference. Stars reflect selected Tukey-adjusted pairwise comparisons shown in the figure. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

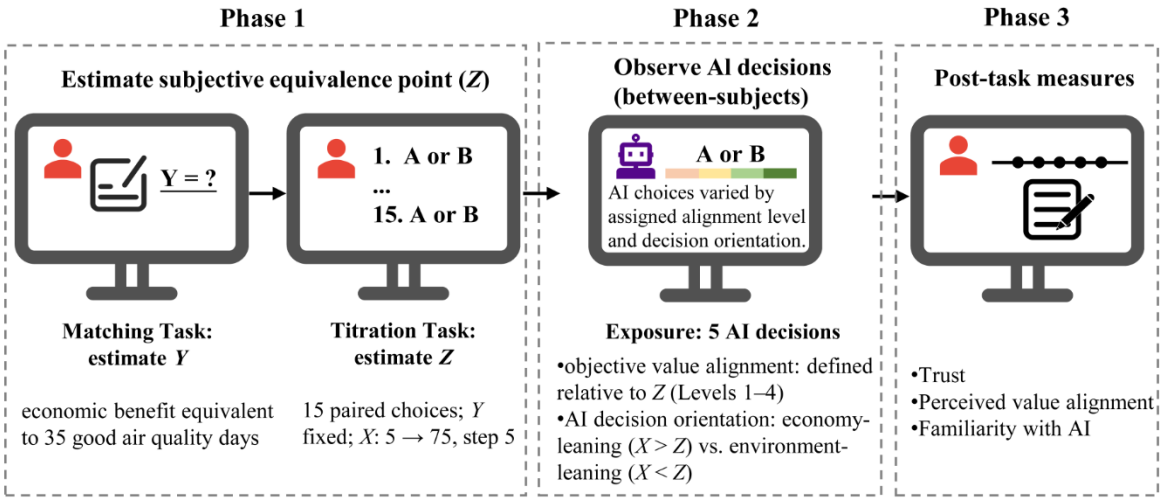


**Figure 3. Experimental procedure and design overview.** Participants completed three phases. In Phase 1, an individual equivalence point ($Z$) for the economic–environmental trade-off was estimated in two steps: a matching task to estimate $Y$ (the economic benefit judged equivalent to 35 good air quality days), followed by a titration task (15 paired choices; $Y$ fixed; $X$ [good air quality days] increasing from 5 to 75) to obtain $Z$. In Phase 2, participants observed five pre-programmed AI-attributed choices which were systematically varied by (i) objective value alignment level defined relative to $Z$ (Levels 1–4) and (ii) AI decision orientation (economy-leaning: X > Z vs environment-leaning: X < Z). In Phase 3, participants completed post-task measures, including trust, perceived value alignment, and perceived familiarity with AI.

**Table 1. Descriptive statistics for human–AI trust by objective value alignment levels and AI decision orientation**

| Objective Value Alignment Levels | AI Decision Orientation | *M* | *SD* | *n* |
|---|---|---|---|---|
| Alignment Level 1 | Economy-leaning | 2.18 | 0.74 | 31 |
| | Environment-leaning | 2.49 | 0.90 | 28 |
| Alignment Level 2 | Economy-leaning | 2.48 | 0.78 | 30 |
| | Environment-leaning | 2.98 | 0.84 | 31 |
| Alignment Level 3 | Economy-leaning | 3.15 | 0.83 | 32 |
| | Environment-leaning | 3.36 | 0.95 | 33 |
| Alignment Level 4 | Economy-leaning | 3.81 | 0.62 | 33 |

| | | | | |
|---|---|---|---|---|
| | Environment-leaning | 3.77 | 0.62 | 32 |

**Table 2. Example of manipulating choice inflection points for environment-leaning AI at incremental objective value alignment levels**

| Good Air Quality Days Setting | 20 days | 25 days | 30 days | 35 days | 40 days |
|---|---|---|---|---|---|
| Alignment Level 1 | Choose B \| | Choose A | Choose A | Choose A | Choose A |
| Alignment Level 2 | Choose B | Choose B \| | Choose A | Choose A | Choose A |
| Alignment Level 3 | Choose B | Choose B | Choose B \| | Choose A | Choose A |
| Alignment Level 4 | Choose B | Choose B | Choose B | Choose B \| | Choose A |

*Note.* "Choose A" indicates the AI selected "Increase annual good air quality days" in that decision pair, while "Choose B" indicates the AI selected "Increase annual economic benefits." The vertical line represents the location of the AI's inflection point. In this example, since the participant's inflection point is between 40 and 45 days, they would choose Plan B in all of the above decision pairs.

**Table 3. Example of manipulating choice inflection points for economy-leaning AI at incremental objective value alignment levels**

| Good Air Quality Days Setting | 45 days | 50 days | 55 days | 60 days | 65 days |
|---|---|---|---|---|---|
| Alignment Level 1 | Choose B | Choose B | Choose B | Choose B \| | Choose A |
| Alignment Level 2 | Choose B | Choose B | Choose B | Choose A \| | Choose A |
| Alignment Level 3 | Choose B | Choose B \| | Choose A | Choose A | Choose A |
| Alignment Level 4 | Choose B \| | Choose A | Choose A | Choose A | Choose A |

*Note.* "Choose A" indicates the AI selected "Increase annual good air quality days" in that decision pair, while "Choose B" indicates the AI selected "Increase annual economic benefits." The vertical line represents the location of the AI's inflection point. In this example, since the participant's inflection point is between 40 and 45 days, they would choose Plan A in all of the above decision pairs.