electronics

MDPI

Article

Collaborative Decision Making with Responsible AI: Establishing Trust and Load Models for Probabilistic Transparency

Xinyue Wang, Yaxin Li and Chengqi Xue *

School of Mechanical Engineering, Southeast University, Nanjing 211189, China; 230189401@seu.edu.cn (X.W.); 202018304@seu.edu.cn (Y.L.)
* Correspondence: ipd_xcq@seu.edu.cn; Tel.: +86-025-52090530

Abstract: In responsible AI development, the construction of AI systems with well-designed transparency and the capability to achieve transparency-adaptive adjustments necessitates a clear and quantified understanding of user states during the interaction process. Among these, trust and load are two important states of the user's internal psychology, albeit often challenging to directly ascertain. Thus, this study employs transparency experiments involving multiple probabilistic indicators to capture users' compliance and reaction times during the interactive collaboration process of receiving real-time feedback. Subsequently, estimations of trust and load states are established, leading to the further development of a state transition matrix. Through the establishment of a trust–workload model, probabilistic estimations of user states under varying levels of transparency are obtained, quantitatively delineating the evolution of states and transparency within interaction sequences. This research lays the groundwork for subsequent endeavors in optimal strategy formulation and the development of transparency dynamically adaptive adjustment strategies based on the trust–workload state model constraints.

Keywords: responsible AI; human computer interaction; transparency design; collaborative decision making; human computer trust; cognitive modeling

Citation: Wang, X.; Li, Y.; Xue, C. Collaborative Decision Making with Responsible AI: Establishing Trust and Load Models for Probabilistic Transparency. *Electronics* **2024**, *13*, 3004. https://doi.org/10.3390/electronics13153004

Academic Editors: Niusha Shafiabady and Jianlong Zhou

Received: 27 May 2024
Revised: 18 July 2024
Accepted: 29 July 2024
Published: 30 July 2024

1. Introduction

The transparency of systems is regarded as the most pressing issue in the practical application of AI [1], thus posing a significant challenge in constructing responsible AI. Studies indicate that even state-of-the-art AI systems, including various intelligent electronic devices or agents, cannot eliminate unreliability in real-world applications [2]. Therefore, in human–computer collaborative tasks, the design and investigation of transparency present crucial opportunities for implementing responsible AI. Transparency design methodologies focus on conveying clues about the probabilistic features of AI, involving information pertaining to uncertainty, dependency, and vulnerability [3]. Transparency design can positively impact user experience and behavior, serving as a pivotal starting point for the more comprehensive, reliable, and responsible application of AI [4].

However, current research on transparency often concentrates on algorithmic perspectives, striving to enhance the interpretability of AI itself [5], while overlooking human factors and ergonomics studies in AI application processes [6]. Blindly elevating transparency may not directly enhance user trust and could potentially degrade the performance of human–computer interaction and collaboration. In comparison to algorithm development, human-centric research bears the greatest responsibility for transparency design in AI [7], including leveraging theoretical research findings from cognitive psychology [5], visualizing probabilistic features [8], and thereby effectively enhancing AI's accountability through interaction. Hence, AI transparency should reconcile the tension between deficiencies in displaying probabilistic information in the interaction interface and user cognition [2]. Relative to the rapid advancement of algorithms, significant gaps persist in

AI transparency research, particularly in complex electronic technologies and systems [9]. How to present system transparency in a manner consistent with user cognition remains a major challenge [10], representing a crucial yet insufficiently explored research dimension and signifying an open frontier in responsible AI and interaction research.

From the perspective of the human–computer interaction (HCI), AI transparency should be designed based on user trust and workload states, further achieving dynamic adjustments based on these two states. The trust state is a complex and multidimensional concept. Establishing a trust state stems partly from the feedback and explanations provided by AI, aiding users in comprehending its operational principles and decision-making logic [11]. Users exhibit strong interest and concern about the concepts and logic behind AI, extending beyond simple acceptance of its computational results. Conversely, the workload state involves the magnitude of psychological effort users endure during the process of understanding transparent information. The quantity of transparency information often directly impacts cognitive resource consumption, thereby influencing user decisions and the efficiency of human–computer collaboration.

In the process of transparency design, the requirements of trust states and the constraints of cognitive load often pose a dilemma. For trust states, as transparency increases, it can assist users in understanding and constructing meaning around AI [12], which has been demonstrated to enhance trust states [3]. However, increasing transparency requires conveying more information and conveying excessive details significantly alters the cognitive load. Excessive cognitive load can lead to attentional distraction, cognitive fatigue, and even psychological discomfort, not only prolonging the time users expend on decision-making in human–computer collaboration but also increasing the likelihood of errors.

Therefore, if an AI system lacks transparency or has a low level of transparency, users cannot perceive the logic behind the AI system's decisions, making effective collaborative decision making difficult and hindering the AI system's accountability. Conversely, if the transparency level is too high, it may significantly increase the response time for collaborative decisions without necessarily improving the accuracy of the outcomes. Addressing these issues, this study extends the series of works conducted by Akash et al. [13–17] and proposes the following human–AI collaborative decision-making loop, as illustrated in Figure 1.

1. The AI system first makes a decision and generates "Actions". The AI system consolidates and analyzes various data to make decisions, such as determining whether an object is a threat or not. As a responsible AI, the "Actions" presented to users include three aspects: the current decision result, the correctness of the previous decision result, and the system's transparency. The system's transparency describes the reliability of the current decision result, reflected through a probabilistic indicator system. This probabilistic indicator system, being an abstract expression, can have different meanings in various application environments, such as reflecting the accuracy of data from various sources using measurement error, natural variation, and prediction error, or indicating the evaluation and measurement of the AI system itself using feature vector similarity, probability scores, and model confidence. The number of probabilistic indicators presented to the user reflects the transparency level of the AI system. For example, level 1 transparency indicates only providing the AI's decision result to the user, while level 4 transparency includes the decision result plus three probabilistic indicators;
2. "Actions" and "Feedback" are presented to the user in real-time, prompting changes in the user's "States". In the "Actions", the correctness of the previous decision result pertains to the AI system's decision result, while "Feedback" pertains to the correctness of the user's previous decision. For instance, if in the previous collaborative decision, the user judged an object as a threat and it indeed was a threat, the "Feedback" will inform the user in real time that their decision was correct. Upon perceiving the "Actions" and "Feedback", the user's cognitive system is stimulated,

which then responds to control interactive behaviors. This process affects the user's trust and cognitive load states;

3. The AI system generates "Observations" about the user and estimates the "States". Influenced by trust and cognitive load states, the user exhibits two types of behavioral responses. One aspect is the user's decision outcome, i.e., agreeing or disagreeing with the AI system's decision, reflecting the user's compliance with the AI system. The other aspect is the user's response time in collaborative decision making. By observing these two behavioral responses, the AI system estimates the trust and cognitive load states;
4. The AI system calculates the "Rewards" of the collaborative decision. The "Rewards" of the collaborative decision are composed of two aspects involving trust and cognitive load. One aspect is the ultimate correctness of the collaborative decision. Blind trust and compliance with the AI system are not always beneficial. For example, if the AI system judges an object as a threat when it is not and the user complies with the AI's result, this should be considered a negative reward. The other aspect is the shorter the decision time, the higher the reward. In summary, the reward setting provides an optimization direction for the AI system's dynamic adjustment, aiming to improve decision accuracy while reducing response time;
5. The AI system dynamically adjusts the system's transparency. Based on the estimation of "States" and the real-time calculation of "Rewards", the AI system can determine the "Actions" strategy that maximizes "Rewards" under the current "States". Among the three elements of "Actions", the current decision result and the correctness of the previous decision result are directly tied to system reliability and cannot be adjusted through design. In contrast, system transparency is the object of dynamic adjustment. Therefore, the AI system can dynamically adjust the number of probabilistic indicators presented to the user to build a responsible AI system. This constitutes a cycle in the human–AI collaboration process, wherein the sequence of decisions (e.g., continuously judging whether multiple targets are threats), the user's "States" dynamically change, and the number of probabilistic indicators displayed by the AI system also dynamically changes, ensuring that transparency always follows the optimal strategy for maximizing decision accuracy and minimizing time.

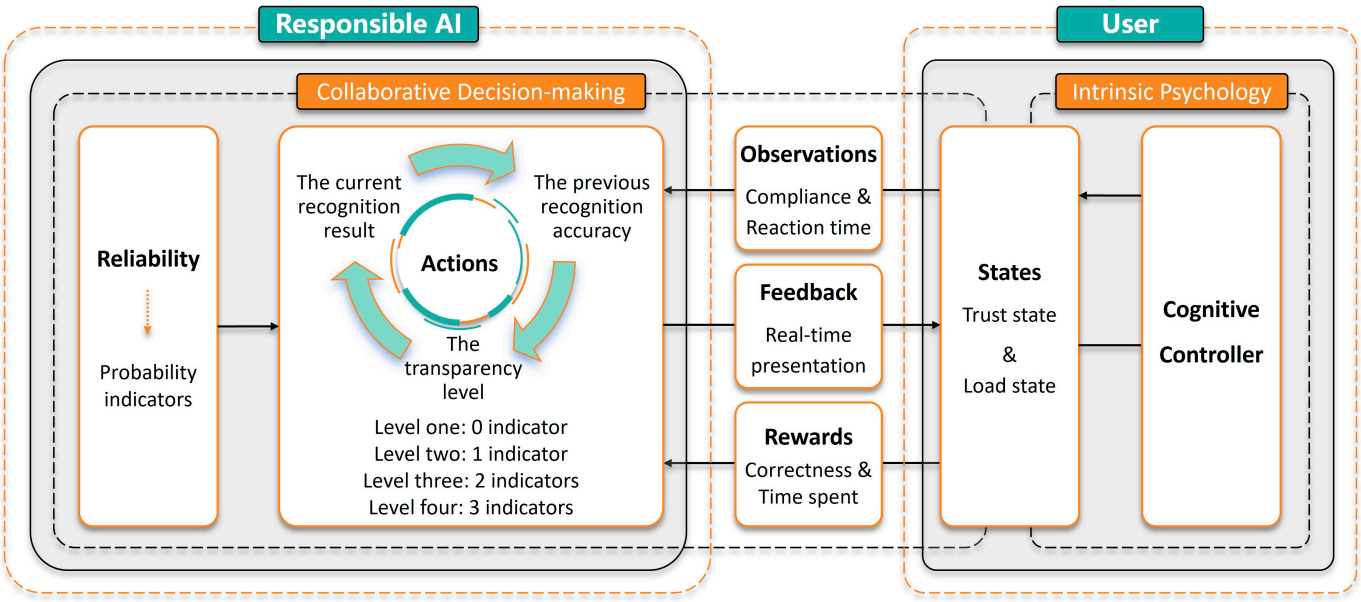

Figure 1. The framework of collaborative decision making with responsible AI.

To achieve the above cycle, we conducted a series of studies, constructing a Partially Observable Markov Decision Process (POMDP) and using reinforcement learning to de-

rive the optimal strategy for transparency adjustment. This study is part of a series of investigations, focusing primarily on points 1 to 3 of the above cycle. Specifically, for the optimization and updating of reinforcement learning strategies, it is first necessary to initialize the strategy, which includes defining the interaction process, conducting collaborative decision-making experiments, and observing users to construct trust and load models. The core significance of this process lies in the following two aspects:

- State estimation through observations: Trust and load states are not directly accessible and need to be estimated through partial observations. Quantifying trust and load states first requires obtaining their current status; however, these states are difficult to observe directly and are hidden variables within cognitive processes. In the field of human–computer interaction research, the evaluation of trust and load states for electronic systems largely relies on self-reported survey methods. For example, the Likert Scale is used to assess trust states and the NASA Task Load Index is used to assess load states. However, in the context of real-time feedback algorithms, continuously querying the cognitive domain is usually infeasible. Therefore, this study constructs observation probability functions based on user observations (compliance and response time) to quantitatively estimate users' trust and load states;
- Dynamic Characteristic Construction: Trust and load states are not fixed characteristics; they have a complex relationship with transparency and mutually influence each other. These states are highly sensitive to transparency and adaptively adjust during interactions. As the sequence of interactions unfolds, the dynamic interplay between these states and transparency evolves. Therefore, this study constructs transition probability functions to model the dynamic properties of users' trust and load states.

In summary, as part of a series of studies, this research constructs trust and load models for the human–AI collaborative process. Specifically, through collaborative decision-making experiments at different transparency levels, we observe users' compliance and response times to construct observation probability functions, forming estimates of trust and load states. Additionally, this study develops transition probability functions to quantify the dynamic evolution of states and transparency throughout the interaction sequence. These methods provide a deep understanding of the collaborative decision-making process with responsible AI and lay the foundation for further optimal strategy derivation and the development of transparency dynamic adaptive adjustment strategies constrained by the trust–load state model.

The remaining sections of this study are organized as follows: Section 2 introduces the relevant research background. Section 3 sets the parameters involved in the model. Section 4 conducts transparency experiments, mainly obtaining observations on user compliance and reaction times. Section 5 constructs the trust–load model. Section 6 discusses the experimental and modeling results and Section 7 summarizes the study.

2. Related Works

As electronic devices and intelligent agents become increasingly complex, transparency has become a prominent topic in recent years. There is a growing commitment to making AI outputs more transparent to maximize the joint performance of human–machine teams [18]. Throughout the application of modern AI systems, users consistently express strong concerns and interests regarding the underlying principles and logic behind system outputs. This demand has formed a humanizing cycle of AI ethical assurance [19]. Research by Brasse et al. indicates that user demands extend far beyond merely accepting system suggestions or ratings, as good transparency can positively impact domain experts' user experiences and behaviors [4]. Visser et al. point out that trust between humans and machines has been repeatedly shown to be a key factor influencing decision effectiveness, with empirical studies demonstrating the benefits of increased transparency [3]. Alexander et al., through neurophysiological measurements, demonstrate that information about others' prior use of algorithms has a greater impact on algorithm adoption than the accuracy of the algorithms themselves, resulting in lower cognitive load. Conversely, adopting algorithms

without any information leads to low cognitive engagement during task processes and compromised task performance [20]. As shown in Figure 2, in situations where AI reliability cannot be guaranteed perfectly, transparency design is significant for constructing and developing responsible AI for the following reasons.

1. Avoidance of abandonment. When AI operates in a "black box" manner and its output results do not match user expectations, it leads to biases in users' perceptions of AI capabilities [21]. This ongoing cognitive dilemma and negative emotions affect users' attitudes and behavioral intentions, gradually eroding and inhibiting their confidence in the system [22], thereby reducing their willingness to use the technology. More critically, this loss of confidence may introduce potential risks in critical real-time decision-making scenarios, affecting decision quality and implementation effectiveness [2]. This scenario is particularly evident in significant decision choices involving military recognition, disaster relief, and medical diagnosis [4]. Good transparency can present fundamental information, action reasons, and uncertain predictions, which aid users in understanding AI and making necessary adjustments, including providing missing instructions or information to AI and correcting its understanding [23];
2. Avoidance of misuse. When AI outputs are not sufficiently understood and validated, misuse of its results may lead to adverse consequences or poor decisions [24]. The lack of necessary questioning of AI results and critical thinking will lead to the cognitive domain's unreasonable confidence in recognition results [25]. Enhancing AI transparency displays can effectively manage uncertainty [2], help users identify when AI may operate beyond its limits, and determine when AI results should not be used [7];
3. Facilitation of assessment. Enhanced transparency assists in more accurately assessing computational domain capabilities and limitations for trust calibration. Visualization of transparency and probability indicators is a standard tool for assessing and communicating risks [26]. Responsible AI should enhance the visibility of underlying processes to enable users to understand current states [2]. Precise probability indicators should be implemented in the computational domain, encompassing all sources of decision uncertainty (e.g., model performance, prior knowledge about training data distributions, and input data noise) [27], to emphasize AI limitations [28]. Displaying this information aids in evaluating whether AI logic and decisions are reasonable, judging their consistency with domain prior knowledge and practical experience, thereby calibrating trust [29]. Additionally, providing and highlighting this metadata to users [3] increases their awareness of probabilistic features [30], improves their perception of cognitive decision risks, and promotes greater caution [31];
4. Bias correction. Responsible AI can complement users' ideas through transparency information and probabilistic feature inference [32], for example, Zhou et al. introduce the uncertainty of training data and model represented by knowledge graphs into AI-informed decision making [33]. This additional information will correct users' cognitive predictions and expectations, thereby improving the quality of recognition decisions [34]. Transparency design assists in constructing and testing causal relationship hypotheses related to recognition decisions [28], including forming reasons for decision outcomes and the association between causes and results [22]. This correction of decision biases is crucial for ensuring accuracy and scientific credibility [24].

Consequently, the establishment, maintenance, and calibration of such trust have become focal points of research [30]. In this context, researchers and designers in HCI bear significant responsibility for trust calibration and system transparency design [7]. Therefore, HCI research should strive to integrate transparency into the data sets, algorithms, and data models within the computational domain, as well as into the intent, behavior, and prediction uncertainties, enabling the cognitive domain to gain a deeper understanding of how the computational domain interprets and acts upon the received data [21].

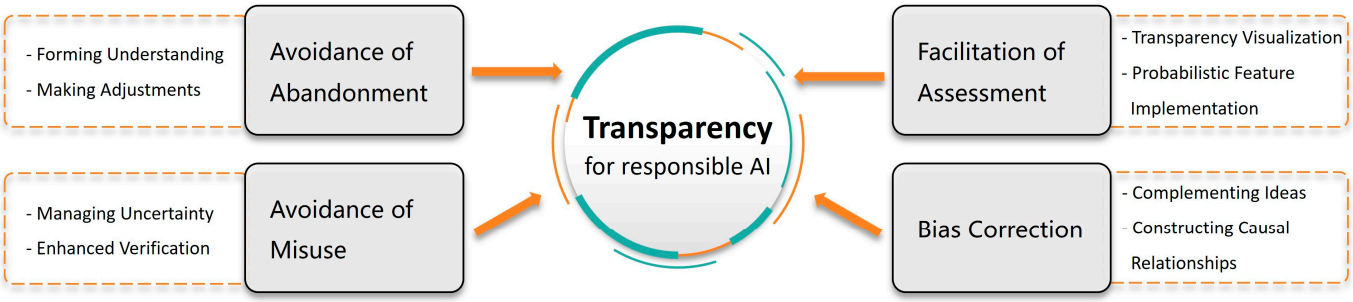

Figure 2. Transparency is significant for constructing and developing responsible AI.

Firstly, HCI researchers must ensure the clear and comprehensible communication of uncertainties. Andrienko posits that merely providing descriptions of data and uncertainties is insufficient to enhance system transparency; the key lies in how this information is presented. The primary task of information visualization is to represent information visually, enabling users to perceive it accurately and efficiently [32]. Bles et al. stress the need to combine statistical methods for quantifying uncertainty with psychological perspectives that highlight the importance of communication's impact on the audience [34]. Hullman emphasizes that designers must identify effective communication methods to successfully convey uncertainty [35]. Jiang et al. point out that, particularly in AI system collaborations, ensuring that the system clearly communicates associated uncertainties when presenting its outputs is crucial for achieving more efficient and reliable decision support [2]. Thus, comprehensively understanding and effectively managing these uncertainties are not only cutting-edge issues in research and design but also essential for ensuring the robustness, reliability, and interpretability of AI technologies.

Simultaneously, HCI researchers should manage probabilistic characteristics through visualization to enhance human–machine trust. Sterzik demonstrates that skillfully conveying probabilistic characteristics profoundly impacts the interpretation of data spaces, with the degree of this interpretation's alignment with cognitive spaces playing a crucial role in building and maintaining human–machine trust [36]. Shin's research elucidates the specific cognitive processes involved in intelligent recommendation algorithms' characteristics (fairness, accountability, transparency, and explainability) and their fundamental connections to trust and subsequent behaviors. Shin asserts that users employ a dual-process model, wherein trust is built upon the combination of the algorithm's normative values and performance-related qualities [37]. Ferrario et al. proposed an incremental trust model applicable to both human–human and human–AI interactions, describing simple, reflective, and paradigmatic forms of trust. Simple trust, characterized by a willingness to depend in the absence of control, demands low cognitive effort. Reflective trust involves a belief in the AI's credibility, while paradigmatic trust combines simple and reflective trust [38]. Cassenti and Kaplan note that probabilistic characteristics are core factors influencing decision confidence, with an inverse relationship between uncertainty and confidence [39]. Panagiotidou et al. emphasize that due to overtrust and lack of criticality, visualization developers must exert effort to make users aware of the inherent errors in visualizations and consciously correct them [40]. To achieve this goal, the system's information representation and interaction design should fully consider users' cognitive needs and expectations, ensuring the provision of clear consistent information that aligns with the users' cognitive space. HCI research has the opportunity to address how to reconcile users' mental models of agent capabilities with the agents' actual constraints [21].

3. Model Parameter Settings

From the previous analysis, it is evident that there is a need to construct a trust–load model to facilitate the observation and estimation of states. This study employs a POMDP to achieve this. POMDPs are used to describe dynamic environments under conditions of uncertainty and partial observability by mathematically modeling the environment's

states, the effects of actions, the probabilities of observations, and the reward of actions, thereby enabling the selection of optimal or near-optimal actions in environments with limited information. Therefore, POMDPs are particularly suitable for modeling scenarios where designers cannot directly and fully observe the states of trust and load.

The specific task involves a human computer collaboration recognition task, where the AI first performs threat identification on a given target, presenting the recognition results and transparency information to the user, who then determines whether the target poses a threat based on this information. The specific parameters for each aspect are described below, with the detailed parameters for the state set $\mathcal{S}$, action set $\mathcal{A}$, and observation set $\mathcal{O}$ shown in Table 1.

Table 1. Parameters for the state set $\mathcal{S}$, action set $\mathcal{A}$, and observation set $\mathcal{O}$.

Model Sets	Inclusive Tuples	Tuple Parameters
State Set $s \in \mathcal{S}$	$s = \begin{bmatrix} \text{trust state } s_T \\ \text{load state } s_W \end{bmatrix}$	$s_T \in T := \begin{Bmatrix} \text{low state } T_\downarrow \\ \text{high state } T_\uparrow \end{Bmatrix}$ $s_W \in W := \begin{Bmatrix} \text{low state } W_\downarrow \\ \text{high state } W_\uparrow \end{Bmatrix}$
Action Set $a \in \mathcal{A}$	$a = \begin{bmatrix} \text{current recognition result } a_{S_A} \\ \text{correctness of the previous recognition } a_E \\ \text{transparency level } a_\tau \end{bmatrix}$	$a_{S_A} \in S_A := \begin{Bmatrix} \text{non} - \text{threat } S_A^- \\ \text{threat} S_A^+ \end{Bmatrix}$ $a_E \in E := \begin{Bmatrix} \text{incorrect } E^- \\ \text{correct } E^+ \end{Bmatrix}$ $a_\tau \in \tau := \begin{Bmatrix} \text{level one } \tau_1 \\ \text{level two } \tau_2 \\ \text{level three } \tau_3 \\ \text{level four } \tau_4 \end{Bmatrix}$
Observation Set $o \in \mathcal{O}$	$o = \begin{bmatrix} \text{compliance } o_C \\ \text{reaction time } o_{RT} \end{bmatrix}$	$o_C \in C := \begin{Bmatrix} \text{rejection } C^- \\ \text{acceptance } C^+ \end{Bmatrix}$ $o_{RT} \in \mathbb{R}^+$

3.1. State Set

The state set is the collection of all possible user states. Each state represents a specific configuration or condition of the user, encompassing relevant information that needs to be considered when presenting transparency. Since the study assumes that the AI cannot directly observe the user's true state, the elements in the state set are not directly accessible. The properties of the state set (such as its size and complexity) directly affect the complexity and solvability of the model.

In this specific study, the user states s are defined as a finite state set $\mathcal{S}$, consisting of tuples that include both the trust state s_T and the workload state s_W. To manage the complexity of subsequent modeling, the trust state $s_T \in T$ and the workload state $s_W \in W$ are each set to two levels: the low state$\downarrow$ and high state$\uparrow$.

3.2. Action Set

The action set encompasses all possible actions that the AI can select. At each time step, the AI chooses an action based on its current state estimate. These actions aim to alter the user's state or obtain new information regarding the state.

In this specific study, the set of AI actions a is defined as a finite action set $\mathcal{A}$, composed of the current recognition result a_{S_A}, the correctness of the previous recognition a_E, and the transparency level a_τ. Since the task involves determining whether a specific target is a threat, the current recognition result $a_{S_A} \in S_A$ includes non-threat (S_A^-) and threat (S_A^+). Additionally, due to the iterative nature of sequential recognition tasks, the correctness of historical recognitions can be replaced by the correctness of the previous recognition $a_E \in E$, which includes incorrect (E^-) and correct (E^+) parameters. Furthermore, the transparency level $a_\tau \in \tau$ is set at four levels, ranging from τ_1 to τ_4, representing levels one to four of transparency, respectively.

3.3. Observation Set

The observation set defines all possible observations that the computational domain can receive. Since states in a POMDP are not directly observable, observations provide indirect information about the current state of the environment. Each time the computational domain performs an action, it receives an observation that depends on the post-action environmental state and the observation probability. The size and nature of the observation set depend on the specific problem's observational capabilities and environmental uncertainty.

In this specific study, the set of user behavior data o that can be obtained directly and in real-time is defined as a finite observation set $\mathcal{O}$, consisting of tuples representing compliance with the AI's recognition result (o_C) and reaction time for decision making (o_{RT}). Compliance $o_C \in C$ includes rejection (C^-) and acceptance (C^+). Additionally, reaction time $o_{RT} \in \mathbb{R}^+$ represents the time required by the user to respond after receiving the recognition result from the AI.

3.4. Transition Probability Function

The transition probability function defines the probability of the user transitioning from the current state to a new state given a particular action. This function reflects the dynamic nature of states, accurately mapping how user states respond to actions taken by the AI and elucidating the impact of each potential action on state changes.

In this specific study, the transition probability function $\mathcal{T}$ is defined to describe the probability of transitioning to subsequent states s_T and s_W given the current trust state s'_T and workload state s'_W, following the action a. The research postulates the conditional autonomy between trust state and workload state with respect to their impact on observations given specific actions. To clarify, the trust state exclusively influences compliance, while the workload state primarily affects reaction time. This postulation facilitates the independent identification of trust and workload models, leading to a marked reduction in the number of parameters within each model, thereby streamlining the requisite subject data for model training. Consequently, the transition probability functions for the trust model $\mathcal{T}_T := T \times T \times \mathcal{A} \to [0,1]$ and the workload model $\mathcal{T}_W := W \times W \times \mathcal{A} \to [0,1]$ can each be represented by a $2 \times 2 \times 16$ matrix.

3.5. Observation Probability Function

The observation probability function, also known as the emission probability function, describes the probability of observing each possible observation given a particular state. This function bridges the relationship between the user states and the observations that the AI can receive.

In this specific study, the observation probability function ε is defined to describe the probability of observing o_C and o_{RT} after taking action a and resulting in state transitions to s'_T and s'_W. For the trust model, the observation probability function $\mathcal{E}_T := C \times T \to [0,1]$ is represented by a 2×2 matrix. For the workload model, the observation probability function $\mathcal{E}_W := \mathbb{R}^+ \times W \to [0,1]$ is represented by two probability density functions. Research findings suggest that the distribution of human reaction times follows an ex-Gaussian distribution [41,42]. Consequently, this study postulates that each workload state exhibits a distinct reaction time pattern, characterized by an ex-Gaussian distribution.

3.6. Methodological Guide

The above content defines the main parameters required for constructing the trust and load models. The specific forms of the action set will be elaborated in Section 4, while the contents of the observation set will be directly derived from the experiments detailed in Section 4. Based on the experimental results, the transition probability functions and observation probability functions can be solved. The establishment of the trust model requires the use of an extended Baum–Welch algorithm. The Baum–Welch algorithm, an expectation–maximization algorithm for parameter estimation, iteratively optimizes the model parameters (transition probability functions, observation probability functions, and

initial state probabilities) to maximize the likelihood of the given observation sequence. In contrast, for the parameter estimation of the load model, the ex-Gaussian distribution of human reaction times renders the Baum–Welch algorithm infeasible. Therefore, Matlab's genetic algorithm is employed to estimate the parameters of the load model.

4. Transparency Experiment

4.1. Experiment Objective

The primary objective of this experiment is to observe user behavior under different AI actions a (current recognition result a_{S_A}, correctness of the previous recognition a_E, and transparency a_τ), as well as the influence of real-time feedback $\mathcal{R}$. Specifically, we aim to gather observations o about user compliance o_C and reaction time o_{RT}.

4.2. Experimental Method

4.2.1. Experimental Scenario

The specific scenario for the human–AI collaborative decision-making experiment is as follows. The AI system first makes a target identification decision (determining whether the target is a threat or non-threat) based on data provided by sensors and other electronic devices. The identification result and the basis for the identification (probability indicators) are presented on the experimental interface. The participants make their identification decisions based on the display on the experimental interface.

4.2.2. Transparency

The experiment employs experimental interfaces with four different levels of transparency. The interfaces for each transparency level are shown in Figure 3. The level-one transparency τ_1 interface provides only the system's recognition result (see Figure 3a); the level-two transparency τ_2 interface adds a probability indicator to the level-one information (see Figure 3b); and the level-three τ_3 and level-four τ_4 transparency interfaces each add an additional probability indicator (see Figure 3c,d).

The different probability indicators represent various probabilistic characteristics of the AI's decision-making process, such as feature vector similarity, probability scores, or model confidence. To reduce the participants' comprehension difficulty, it was explained to them that the three probability indicators are independent abstract bases for the AI's decision making, with each equally influencing the AI's decision outcome. In other words, the red lines and values on each indicator represent the probability that the AI system, based on that indicator, believes the target object is a threat.

The shading in the probability indicators is included only to maintain consistency with our series of studies and does not serve as a variable in this study. In our series of studies, due to the presence of uncertainty, the threshold for the AI system's decision-making is not a fixed value: rather, it fluctuates with specific contexts and tasks, forming a probability distribution and thus being displayed using density bars. However, in this study, to control the complexity of the experiment, the distribution of thresholds is not treated as a variable but merely as an accurate representation of the experimental interface. Participants primarily use the red lines and annotated probability values to make their decision judgments.

4.2.3. Probability Characteristics Setting

First, the true situation of whether the target is a threat is defined as $\overline{a}_S \in S := S^-, S^+$, where S^- represents a true non-threat and S^+ represents a true threat. In each trial, the probability of the target being a true threat is equal, i.e., the prior probability of the true situation is $p(S^-) = p(S^+) = 0.5$.

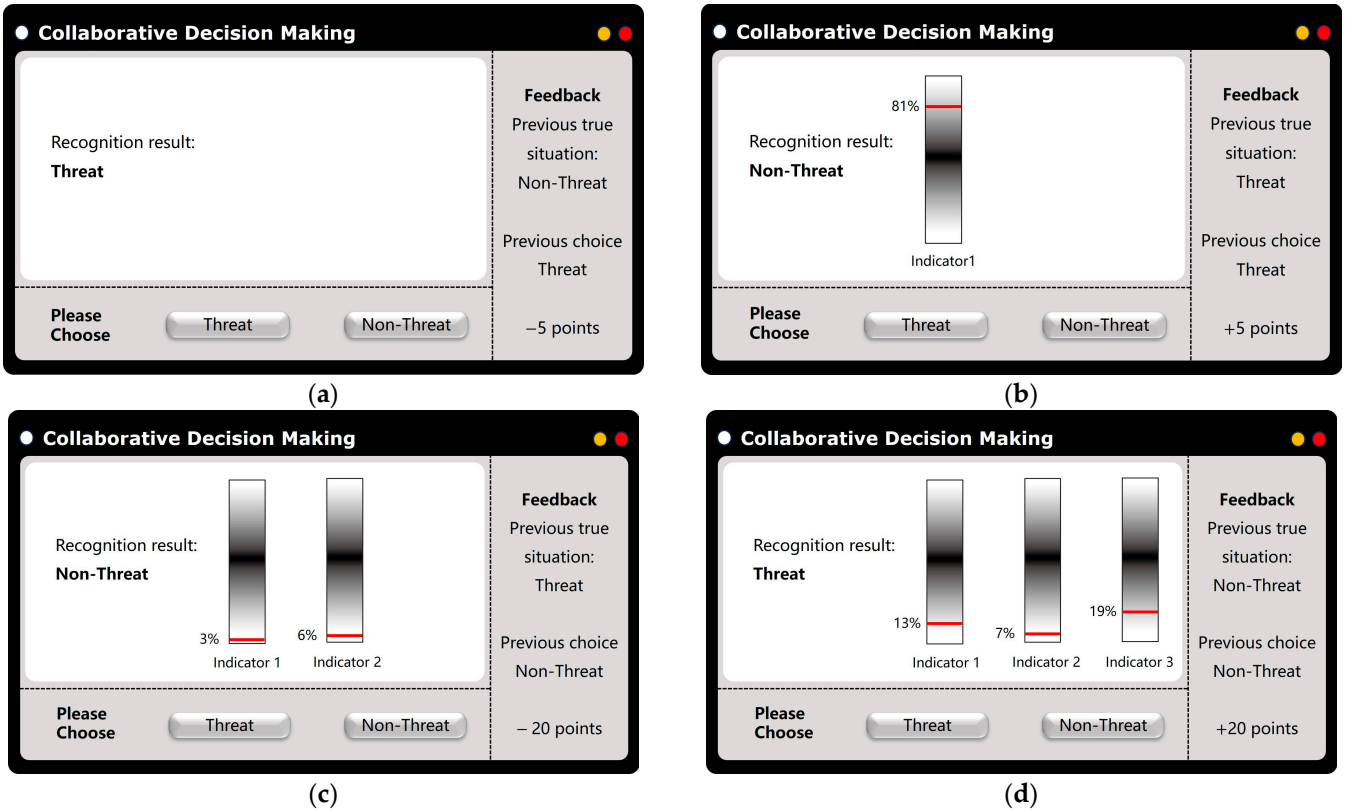

Figure 3. Experimental interfaces for different transparency levels: (**a**) The level-one transparency τ_1 interface provides only the system recognition result; (**b**) The level-two transparency τ_2 interface provides the system recognition result and one indicator; (**c**) The level-three transparency τ_3 interface provides the system recognition result and two indicators; (**d**) The level-four transparency τ_4 interface provides the system recognition result and three indicators.

To determine the specific values displayed for the probability indicators on the interface, this study conducted actual measurements on a particular AI system. The reliability of this AI system is 80%, with three indicators equally influencing the judgment outcome. By statistically analyzing the AI system's current judgment result a_{S_A}, the corresponding actual situation $\bar{a}_S$ and the corresponding indicator display results, such that the following relationships can be established in Table 2.

Table 2. Confusion matrix of the true situation $\bar{a}_S$, current recognition result a_{S_A}, and interface display (Non-τ_1).

True Situation	AI Recognizes Correctly	AI Recognizes Incorrectly
Threat	$p(S_A^+ \mid S^+) = 0.8$	$p(S_A^- \mid S^+) = 0.2$
Non-threat	$p(S_A^- \mid S^-) = 0.8$	$p(S_A^+ \mid S^-) = 0.2$

1. When the system correctly recognizes a true threat (S^+) as "threat" (S_A^+), it is considered a True Positive. The conditional probability is $p(S_A^+ \mid S^+) = 0.8$. In this case, if the experimental interface is not τ_1, at least one probability indicator displayed will have a value between 93% and 97%, while the other probability indicators will randomly distribute between the ranges of 93 to 97% and 80 to 90%;
2. When the system incorrectly recognizes a true threat (S^+) as "non-threat" (S_A^-), it is considered a False Negative. The conditional probability is $p(S_A^- \mid S^+) = 0.2$. In this case, if the experimental interface is not τ_1, at least one probability indicator displayed will have a value between 80% and 90%, while the other probability indicators will randomly distribute between the ranges of 80 to 90% and 10 to 20%;

3. When the system correctly recognizes a true non-threat (S^-) as "non-threat" (S_A^-), it is considered a True Negative. The conditional probability is $p(S_A^-|S^-) = 0.8$. In this case, if the experimental interface is not $\tau 1 \tau 1$, the values of all displayed probability indicators will randomly distribute between 3% and 7%;
4. When the system incorrectly recognizes a true non-threat (S^-) as "threat" (S_A^+), it is considered a False Positive. The conditional probability is $p(S_A^+|S^-) = 0.2$. In this case, if the experimental interface is not τ_1, at least one probability indicator displayed will have a value between 10% and 20%, while the other probability indicators will randomly distribute between the ranges of 10 to 20% and 3 to 7%.

4.2.4. Real-Time Feedback Setting

Real-time feedback is derived from comparing the participant's final decision result with the true situation. The participant's decision result is defined as $\overline{a}_{S_H} \in S_H := S_H^-, S_H^+$. Scores can be assigned based on the relationship between the participant's decision result $\overline{a}_{S_H}$ and the true situation $\overline{a}_S$, defining a decision feedback function $\mathcal{R}_D : S_H \times S \to \mathbb{R}$. Therefore, $\overline{a}_S$, $\overline{a}_{S_H}$, and $\mathcal{R}_D$ can also form a confusion matrix. The notation, description, and specific score values are as follows and are summarized in Table 3.

1. When the participant correctly identifies a true threat (S^+) as a "threat" (S_H^+), it means that after correctly recognizing the threat, resources must be expended to address the threat, ensuring the completion of the task. Therefore, the decision feedback is set at +5 points, $\mathcal{R}_D(S_H^+|S^+) = +5$;
2. When the participant incorrectly identifies a true threat (S^+) as a "non-threat" (S_H^-), it means that after incorrectly ignoring the threat, the threat is not addressed, leading to punitive consequences and failure to complete the task. Therefore, the decision feedback is set at -20 points, $\mathcal{R}_D(S_H^-|S^+) = -20$;
3. When the participant correctly identifies a true non-threat (S^-) as a "non-threat" (S_H^-), it means that after correctly recognizing the non-threat, resources are not expended to address the threat, ensuring the completion of the task. Therefore, the decision feedback is set at +20 points, $\mathcal{R}_D(S_H^-|S^-) = +20$;
4. When the participant incorrectly identifies a true non-threat (S^-) as a "threat" (S_H^+), it means that after incorrectly recognizing the threat, resources are wasted to address the non-threat, ensuring the completion of the task. Therefore, the decision feedback is set at -5 points, $\mathcal{R}_D(S_H^+|S^-) = -5$.

Table 3. Confusion matrix of the true situation $\overline{a}_S$, participant's recognition result $\overline{a}_{S_H}$, and decision feedback $\mathcal{R}_D$.

True Situation	Participant Recognizes Correctly	Participant Recognizes Incorrectly
Threat	$\mathcal{R}_D(S_H^+\|S^+) = +5$	$\mathcal{R}_D(S_H^-\|S^+) = -20$
Non-threat	$\mathcal{R}_D(S_H^-\|S^-) = +20$	$\mathcal{R}_D(S_H^+\|S^-) = -5$

4.3. Experimental Procedure

The experimenter explains to the participants that the task is to make a rapid yet accurate decision on whether the target is a threat based on the displayed results on the experimental interface. The system's recognition accuracy is set to a value below 100%, with each probability indicator equally influencing the system's recognition decision. It is important to note that the specific accuracy value of the system is not disclosed to the participants. The sequence of events for each trial is as follows, as illustrated in Figure 4.

1. The interface shows that the system is performing recognition, requiring no action from the participant. This display lasts for 1 s before disappearing. The purpose of this interface is to separate consecutive judgments and prevent participant fatigue, which could affect reaction times;

2. The interface then presents the interactive recognition screen, where the participant must decide whether the target is a "threat" or "non-threat" based on the displayed content;
3. The interface immediately shows feedback, indicating the correctness of the participant's judgment and the corresponding score adjustment, which remains until being updated by the next feedback.

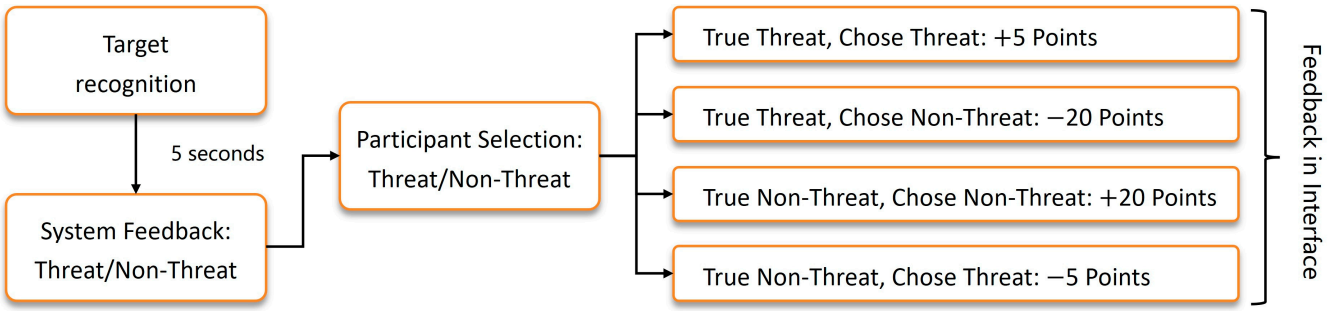

Figure 4. Sequence of events in a single trial.

Before the formal experiment begins, participants complete 20 trials of a pre-experiment to familiarize themselves with the experimental interface and the four different transparency levels. The formal experiment consists of four sets of recognition tasks, each set containing 25 trials, with a 1-min rest period between sets. The interface for each transparency level is randomly assigned to participants to minimize order effects.

A total of 30 participants were recruited for this experiment, all of whom were undergraduate or graduate students aged between 18 and 25 years. Among them, 18 were male and 12 were female, with a male-to-female ratio of 3:2. The experimental platform was built using E-Prime 2.0.

5. Model Construction

To establish a POMDP model that matches the probabilities between the computational domain and the cognitive domain, parameter estimation is necessary. This involves estimating the initial state probability p_0, the observation probability function $\mathcal{E}$ and the state transition probability function $\mathcal{T}$. Therefore, data from all participants are aggregated to estimate the parameters for both the trust model and the workload model.

5.1. Trust Model

The parameter estimation for the trust probability matching process involves finding the optimal parameters that maximize the likelihood of the observed sequence given a specific action sequence.

First, the initial state probability $p_0(s_T)$ for the trust model is estimated. The $p_0(s_T)$ reflects the probability with which participants start interacting with the system in a given trust state. Parameter estimation reveals that the initial state probability $p_0(T_\downarrow)$ for low trust $T_\downarrow$ is 0.1323, while the initial state probability $p_0(T_\uparrow)$ for high trust $T_\uparrow$ is 0.8677. This indicates that participants tend to trust the system at the beginning of the interaction sequence.

Next, the observation probability function $\mathcal{E}_T(o_C|s_T)$ for the trust model is estimated. The $\mathcal{E}_T(o_C|s_T)$ represents the probability of observing participants accepting or rejecting the system's recognition result in each trust state. As shown in Figure 5, the probabilities of observing acceptance or rejection in both trust states are indicated by the arrows. The probabilities of acceptance and rejection are both over 0.93 in both trust states. However, there remains a 0.0655 probability of rejection in the high trust state and a 0.0107 probability of acceptance in the low trust state.

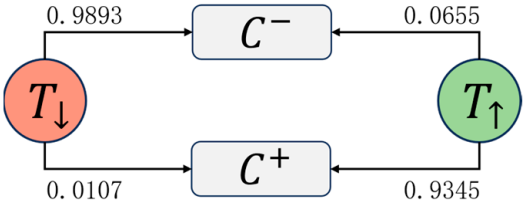

Figure 5. The observation probability function $\mathcal{E}_T(o_C|s_T)$ for the trust model.

Finally, the transition probability function $\mathcal{T}_T(s'_T|s_T, a)$ for the trust model is estimated. The $\mathcal{T}_T(s'_T|s_T, a)$ reflects the probability of transitioning from the current trust state s_T to the next trust state s'_T given an action $a \in \mathcal{A}$. Figure 6 depicts the transition probability graph based on $\mathcal{T}_T(s'_T|s_T, a)$. The numbers next to the arrows in each graph indicate the probabilities of state transitions.

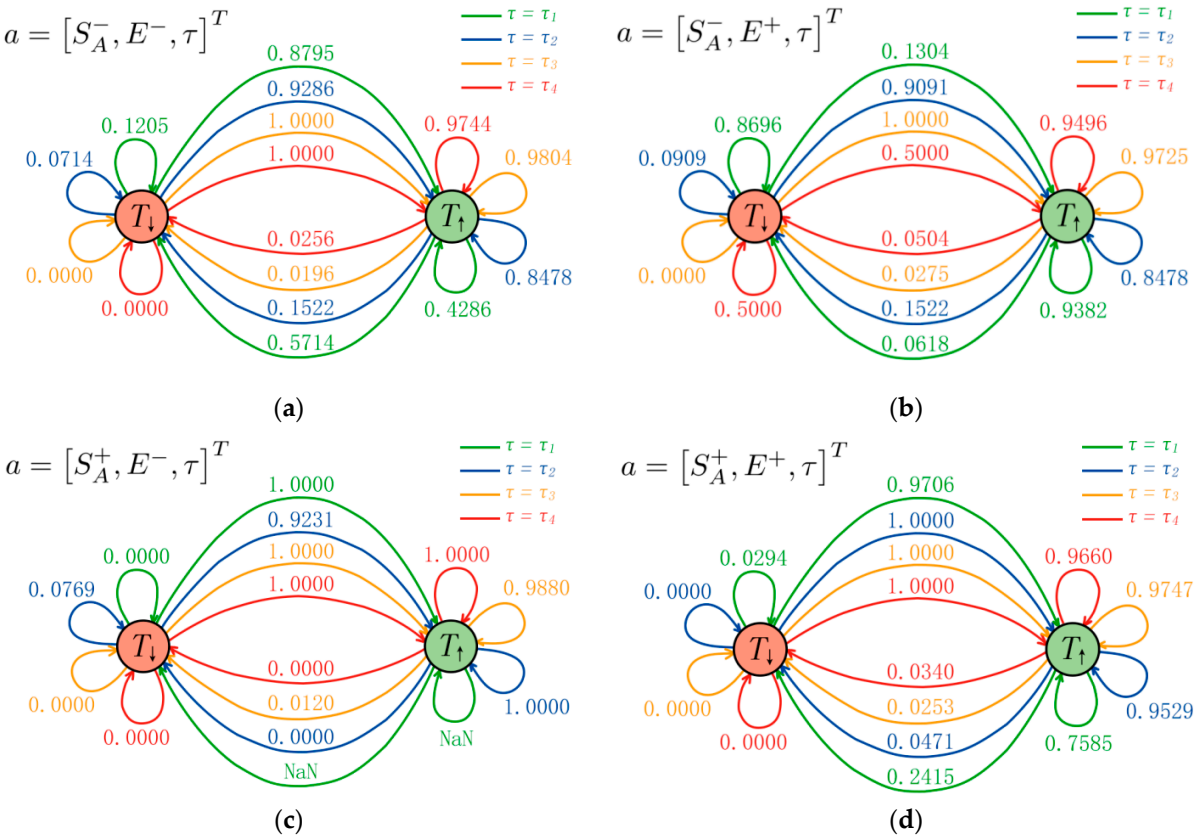

Figure 6. Transition probability function $\mathcal{T}_T(s'_T|s_T, a)$ for the trust model: (**a**) Recognition: non-threat. Previous recognition incorrect. (**b**) Recognition: non-threat. Previous recognition correct. (**c**) Recognition: threat. Previous recognition incorrect. (**d**) Recognition: threat. Previous recognition correct.

After completing the parameter estimation, we can analyze the factors influencing the trust state based on the transition probability graph. First, consider the effects of transparency a_τ and the current trust state s_T on the next trust state s'_T. The transition probabilities for the τ_1 interface varies significantly in each scenario. This indicates that when participants are presented only with the system's recognition result, without additional information to aid their decision making, their trust state is heavily influenced by the system's current recognition result a_{S_A} and the correctness of the system's previous recognition a_E. According to the current experimental data, this influence does not follow a fixed pattern.

For other levels of transparency, the following conclusions can be drawn by comparison. For the probability of transitioning from low trust $T_{\downarrow}$ to high trust $T_{\uparrow}$ and for remaining in $T_{\uparrow}$, the τ_3 and τ_4 interfaces have similar effects. In contrast, the τ_2 interface slightly reduces these probabilities, indicating that using the τ_2 interface results in a relatively higher probability of transitioning from $T_{\uparrow}$ to $T_{\downarrow}$ or remaining in $T_{\downarrow}$. This may be because the τ_2 interface makes participants aware of system errors; but, without additional indicators, they cannot understand why the system made an error, leading to a decrease in trust levels.

Next, consider the effects of the current recognition result a_{S_A} and the correctness of the previous recognition a_E on the next trust state s'_T. The a_{S_A} can create different risk scenarios. Figure 6a,b can be considered to represent high-risk scenarios. Whether the system's previous recognition was correct or not, accepting the system's S_A^- result means no countermeasures will be taken and, if the true state is a threat, a penalty (-20 points) will be incurred. In contrast, Figure 6c,d can be considered to represent low-risk scenarios. Regardless of whether the system's previous recognition was correct, accepting the system's S_A^+ result means preparing for a threat response and, even if the true state is not a threat, only a small penalty (-5 points) will be incurred.

The data show that in low-risk scenarios, participants are more likely to transition to and remain in high trust $T_{\uparrow}$. For example, in various transparency levels, the probability of transitioning from low trust $T_{\downarrow}$ to $T_{\uparrow}$ in low-risk scenarios is at least 92.31% and in some transparency levels, it even reaches 100%, which is significantly higher than the system's 80% accuracy rate. This indicates the inherent risk-averse behavior within the cognitive domain.

5.2. Workload Model

First, the initial state probability $p_0(s_W)$ for the workload model is estimated. The $p_0(s_W)$ reflects the probability with which participants start interacting with the system in a given workload state. Parameter estimation reveals that the initial state probability $p_0(W_{\downarrow})$ for low workload $W_{\downarrow}$ is 0.2689, while the initial state probability $p_0(W_{\uparrow})$ for high workload $W_{\uparrow}$ is 0.7311. This indicates that participants tend to start with a higher workload to familiarize themselves with the system.

Next, the observation probability function $\mathcal{E}_W(o_{RT}|s_W)$ for the workload model is estimated. The $\mathcal{E}_W(o_{RT}|s_W)$ represents the probability density function of the reaction time observed in each workload state, as shown in Figure 7. Both probability density functions can be represented by an ex-Gaussian distribution. For low workload $W_{\downarrow}$, $\mu_{W_{\downarrow}}$ is 0.3804, $\sigma_{W_{\downarrow}}$ is 0.2487, and $\tau_{W_{\downarrow}}$ is 0.5172. For high workload $W_{\uparrow}$, $\mu_{W_{\downarrow}}$ is 1.4347, $\sigma_{W_{\downarrow}}$ is 0.3249, and $\tau_{W_{\downarrow}}$ is 2.8436.

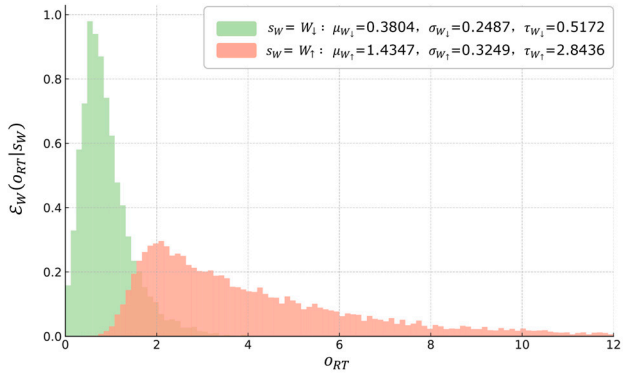

Figure 7. Observation probability function $\mathcal{E}_W(o_{RT}|s_W)$ for the workload model.

Finally, the transition probability function $\mathcal{T}_W(s'_W|s_W, a)$ for the workload model is estimated. The $\mathcal{T}_W(s'_W|s_W, a)$ reflects the probability of transitioning from the current workload state s_T to the next workload state s'_W given an action $a \in \mathcal{A}$. Figure 8 depicts the

transition probability graph based on $\mathcal{T}_W(s'_W|s_W, a)$. The numbers next to the arrows in each graph indicate the probabilities of state transitions.

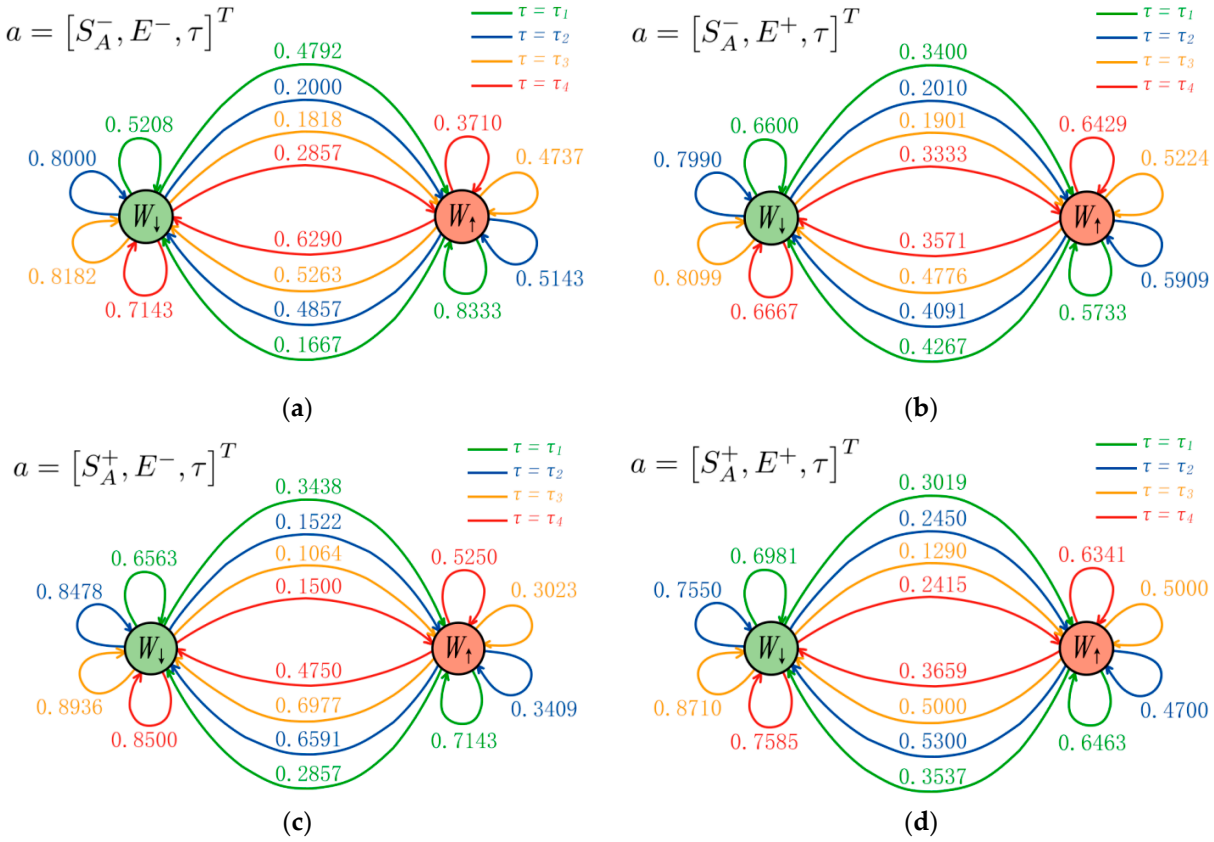

Figure 8. Transition probability function $\mathcal{T}_W(s'_W|s_W, a)$ for the workload model: (**a**) Recognition: non-threat. Previous recognition incorrect. (**b**) Recognition: non-threat. Previous recognition correct. (**c**) Recognition: threat. Previous recognition incorrect. (**d**) Recognition: threat. Previous recognition correct.

After completing the parameter estimation, the factors influencing the workload state can be analyzed based on the transition probability graph. Firstly, consider the effects of transparency a_τ and the current workload state s_W on the next workload state s'_W. The τ_1 interface leads to the highest probability of transitioning from low workload $W_\downarrow$ to high workload $W_\uparrow$ and remaining in $W_\uparrow$. This indicates that when participants are provided only with the system's recognition results, without additional information to aid their decision making, their workload is maximized. This is because participants feel uncertain and hesitate to follow the system's recommendation when they have no other information to judge the system's accuracy, leading to increased time and effort in decision making.

For other levels of transparency, the following conclusions can be drawn: given the current recognition result a_{S_A} and the correctness of the previous recognition a_E, the τ_4 interface is more likely to transition participants from low workload $W_\downarrow$ to high workload $W_\uparrow$ and to keep them in $W_\uparrow$. Therefore, under the $W_\downarrow$ state, higher transparency is more likely to increase participants' workload, as they need to process more information to make a decision. However, the interface is most likely to transition participants from $W_\uparrow$ to $W_\downarrow$ and to keep them in $W_\downarrow$ is τ_3. This may be because the τ_3 interface provides an optimal amount of information, which is sufficient for understanding the system without excessively taxing cognitive resources.

Secondly, consider the effects of the current recognition result a_{S_A} and the correctness of the previous recognition a_E on the next workload state s'_W. It can be seen that when using the $\tau_2 - \tau_4$ interfaces, the state transition probabilities are relatively stable across the four different scenarios. In contrast, when using the τ_1 interface, the probability of transitioning

to and remaining in low workload $W_{\downarrow}$ is higher in low-risk scenarios (Figure 8c,d) compared to corresponding high-risk scenarios (Figure 8a,b). Furthermore, compared to situations where the system's previous recognition was incorrect (E^{-}, Figure 8a,c), the probability of transitioning to and remaining in $W_{\downarrow}$ is higher when the previous recognition was correct (E^{+}, Figure 8b,d). This indicates that when participants have no other information to judge the system's recognition accuracy, their workload is significantly influenced by a_{S_A} and a_E. Low-risk and E^{+} scenarios make participants more decisive in their decision making. When additional information is available to help participants judge the system's recommendation accuracy, the influence of a_{S_A} and a_E on participants' workload is significantly reduced.

6. Discussion

6.1. Research Methods

In the development and application of AI, the design and study of transparency are critical opportunities for ensuring its responsibility. In human–computer interaction and collaboration, AI often cannot achieve perfect reliability and its explanations and the manner in which they are provided can easily be modified [22]. This is particularly true for the presentation of transparency information such as parameters and indicators in interactive interfaces [43]. Conducting user-centered research to develop responsible AI fully responds to the call by Abdul et al. for transparency design to integrate cognitive psychology and be quantified [5]. This is specifically reflected in three aspects:

- Through cognitive psychology experiments, this study explores how users utilize transparency for visualization, reasoning, and knowledge construction [32], capturing the dynamic impacts of transparency on trust and workload states;
- It combines the cognitive psychology perspective, which emphasizes the importance of communicating transparency, with quantitative statistical methods to manage uncertainty [34], ensuring that AI can clearly convey the decision basis and the probabilistic characteristics contained within its outputs [2];
- By precisely modeling functions such as state transition probability and observation probability, this research provides a method to detail the impact and evolution processes between transparency and trust–load states.

This study references a series of research contents by Akash et al. [13–17]. In their research, tasks were typically set to determine whether there was a shooter in a building, with the highest level of transparency only up to three levels, including identification results, probability indicators, and infrared images, with the system's recognition accuracy often being 50% or 70%. This chapter extends their research settings in several significant ways:

- The experimental tasks adopt more general scenarios, thus making the research results more generalizable;
- The transparency levels were increased to four and the system's recognition accuracy was improved to 80%, thereby exponentially increasing the complexity of modeling and analysis;
- The feedback given to participants was quantified and scored, constructing a feedback confusion matrix;
- The transparency levels were increased to four and the system's recognition accuracy was improved to 80%, thereby exponentially increasing the complexity of modeling and analysis.

In summary, this study introduces significant innovations and deepens existing research, thereby enhancing the extensibility and theoretical value of the research results.

6.2. State Estimation

Traditional methods for assessing user states, such as surveys and interviews, can provide direct user feedback that is often limited by their static data collection approach and the influence of user subjectivity. These methods may not accurately reflect the real-time experience and psychological state of users during interaction. The estimation methods

and modeling techniques employed in this study allow for real-time monitoring and analysis of user states without disrupting the execution of primary tasks. Among various reinforcement learning methods, a POMDP approach was utilized, estimating states using behavior data indicators that can be directly and continuously obtained in real time.

Compliance and response time in the cognitive domain are two key indicators that implicitly reflect users' basic trust in AI and their current workload state. By comprehensively analyzing these indicators, the data collection process becomes more efficient, reducing response delays and subjective biases that might be introduced by traditional evaluation methods. Moreover, this provides a deeper understanding of the dynamic changes in trust and workload levels during user interaction.

6.3. State Modeling

The modeling of states reveals that trust and workload states do not directly transition or remain constant with the change in transparency. In other words, higher transparency is not always beneficial or detrimental. Specifically, the next trust state s'_T does not directly transition or remain based solely on changes in transparency a_τ. Instead, it is influenced by a combination of factors including a_τ, the current trust state s_T, the current recognition result a_{S_A}, and the correctness of the previous recognition a_E. Similarly, the next workload state s'_W does not transition or remain directly based on changes in transparency a_τ. Instead, it is influenced by a combination of factors, including a_τ, the current workload state s_W, the current recognition result a_{S_A}, and the correctness of the previous recognition a_E. Among these, given the conditions of a_{S_A} and a_E, the highest or lowest levels of transparency are most likely to place participants in a high workload state.

Therefore, transparency should be dynamically adjusted based on the estimation of trust and workload states, as well as the system's current and previous recognition results. This deep understanding provides a scientific basis for designing transparent and responsible AI, contributing to the enhancement of overall efficiency in human–computer interaction and collaboration by precisely adjusting interface transparency.

6.4. Limitations

The limitations of this study lie in the fact that, relative to the ultimate goal of constructing an AI system that dynamically and adaptively adjusts transparency, this research represents a necessary but initial first step. The estimates and models of trust–load states obtained in this study are necessary conditions for the development of adaptive transparency adjustment but further solutions are still required. Our subsequent research will focus on the design space constrained by trust–load states, using reinforcement learning methods to find the optimal subset under dual constraints, ultimately obtaining a transparency adjustment strategy that maximizes cumulative reward.

7. Conclusions

This study constructs a trust–load model within the human–computer collaboration decision-making process based on the POMDP method. In scenarios where AI transparency is presented with multiple probabilistic indicators, the research focuses on observing compliance and response time through experiments, forming estimates and models of trust and load states based on these observations. Furthermore, this study establishes trust–load state transition matrices under different levels of transparency, quantitatively describing the evolutionary process between states and transparency in interaction sequences. Through the above methods, this study constructed the observation probability functions and transition probability functions for the trust and load models. This enables the acquisition of probabilistic estimates and dynamic characteristics of the user's internal states under different transparency levels. This study offers a deep understanding of the user's state during interactions with transparent AI, thereby laying the foundation for further optimal strategy solutions, the development of dynamically adaptive transparency adjustment strategies constrained by the trust–load state model, and the implementation of responsible AI.

Author Contributions: Conceptualization, X.W.; methodology, X.W. and Y.L.; software, X.W. and Y.L.; validation, X.W.; formal analysis, X.W. and Y.L.; investigation, X.W.; resources, X.W. and Y.L.; data curation, Y.L.; writing—original draft preparation, X.W.; writing—review and editing, X.W.; visualization, X.W. and Y.L.; supervision, C.X.; project administration, C.X.; funding acquisition, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 72271053 and 71871056.

Institutional Review Board Statement: All subjects gave their informed consent for inclusion before they participated in the study. Ethics approval is not required for this type of study. The study has been granted an exemption by the Institutional Review Board of the School of Mechanical Engineering Southeast University.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We would like to thank the anonymous reviewers of this paper for their constructive suggestions and comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
2. Jiang, J.; Karran, A.J.; Coursaris, C.K.; Léger, P.-M.; Beringer, J. A Situation Awareness Perspective on Human-AI Interaction: Tensions and Opportunities. *Int. J. Hum.–Comput. Interact.* **2023**, *39*, 1789–1806. [CrossRef]
3. de Visser, E.J.; Peeters, M.M.M.; Jung, M.F.; Kohn, S.; Shaw, T.H.; Pak, R.; Neerincx, M.A. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *Int. J. Soc. Robot.* **2020**, *12*, 459–478. [CrossRef]
4. Brasse, J.; Broder, H.R.; Förster, M.; Klier, M.; Sigler, I. Explainable Artificial Intelligence in Information Systems: A Review of the Status Quo and Future Research Directions. *Electron. Mark.* **2023**, *33*, 26. [CrossRef]
5. Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B.Y.; Kankanhalli, M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–18.
6. Purificato, E.; Lorenzo, F.; Fallucchi, F.; De Luca, E.W. The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes. *Int. J. Hum.–Comput. Interact.* **2023**, *39*, 1543–1562. [CrossRef]
7. Vorm, E.S.; Combs, D.J.Y. Integrating Transparency, Trust, and Acceptance: The Intelligent Systems Technology Acceptance Model (ISTAM). *Int. J. Hum.–Comput. Interact.* **2022**, *38*, 1828–1845. [CrossRef]
8. Korporaal, M.; Ruginski, I.T.; Fabrikant, S.I. Effects of Uncertainty Visualization on Map-Based Decision Making Under Time Pressure. *Front. Comput. Sci.* **2020**, *2*. [CrossRef]
9. Franconeri, S.L.; Padilla, L.M.; Shah, P.; Zacks, J.M.; Hullman, J. The Science of Visual Data Communication: What Works. *Psychol. Sci. Public Interest* **2021**, *22*, 110–161. [CrossRef] [PubMed]
10. Huang, C.-L.; Haried, P. An Evaluation of Uncertainty and Anticipatory Anxiety Impacts on Technology Use. *Int. J. Hum.–Comput. Interact.* **2020**, *36*, 641–649. [CrossRef]
11. Angerschmid, A.; Zhou, J.; Theuermann, K.; Chen, F.; Holzinger, A. Fairness and Explanation in AI-Informed Decision Making. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 556–579. [CrossRef]
12. Sacha, D.; Senaratne, H.; Kwon, B.C.; Ellis, G.; Keim, D.A. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 240–249. [CrossRef] [PubMed]
13. Akash, K.; McMahon, G.; Reid, T.; Jain, N. Human Trust-Based Feedback Control: Dynamically Varying Automation Transparency to Optimize Human-Machine Interactions. *IEEE Control Syst. Mag.* **2020**, *40*, 98–116. [CrossRef]
14. Akash, K.; Reid, T.; Jain, N. Improving Human-Machine Collaboration Through Transparency-Based Feedback—Part II: Control Design and Synthesis. *IFAC-PapersOnLine* **2019**, *51*, 322–328. [CrossRef]
15. McMahon, G.; Akash, K.; Reid, T.; Jain, N. On Modeling Human Trust in Automation: Identifying Distinct Dynamics through Clustering of Markovian Models. *IFAC-PapersOnLine* **2020**, *53*, 356–363. [CrossRef]
16. Hu, W.-L.; Akash, K.; Reid, T.; Jain, N. Computational Modeling of the Dynamics of Human Trust During Human–Machine Interactions. *IEEE Trans. Hum.-Mach. Syst.* **2019**, *49*, 485–497. [CrossRef]
17. Akash, K.; Polson, K.; Reid, T.; Jain, N. Improving Human-Machine Collaboration Through Transparency-Based Feedback—Part I: Human Trust and Workload Model. *IFAC-PapersOnLine* **2019**, *51*, 315–321. [CrossRef]
18. Chen, J.Y.C. Transparent Human–Agent Communications. *Int. J. Hum.–Comput. Interact.* **2022**, *38*, 1737–1738. [CrossRef]
19. Zhou, J.; Chen, F. Towards Humanity-in-the-Loop in AI Lifecycle. In *Humanity Driven AI: Productivity, Well-Being, Sustainability and Partnership*; Chen, F., Zhou, J., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 3–13, ISBN 978-3-030-72188-6.

20. Alexander, V.; Blinder, C.; Zak, P.J. Why Trust an Algorithm? Performance, Cognition, and Neurophysiology. *Comput. Hum. Behav.* **2018**, *89*, 279–288. [CrossRef]
21. Cila, N. Designing Human-Agent Collaborations: Commitment, Responsiveness, and Support. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1–18.
22. Shulner-Tal, A.; Kuflik, T.; Kliger, D. Enhancing Fairness Perception—Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople's Fairness Perceptions of Algorithmic Decisions. *Int. J. Hum.–Comput. Interact.* **2023**, *39*, 1455–1482. [CrossRef]
23. Chen, J.Y.C.; Lakhmani, S.G.; Stowers, K.; Selkowitz, A.R.; Wright, J.L.; Barnes, M. Situation Awareness-Based Agent Transparency and Human-Autonomy Teaming Effectiveness. *Theor. Issues Ergon. Sci.* **2018**, *19*, 259–282. [CrossRef]
24. Padilla, L.M.K.; Castro, S.C.; Hosseinpour, H. Chapter Seven—A Review of Uncertainty Visualization Errors: Working Memory as an Explanatory Theory. In *Psychology of Learning and Motivation*; Federmeier, K.D., Ed.; The Psychology of Learning and Motivation; Academic Press: Cambridge, MA, USA, 2021; Volume 74, pp. 275–315.
25. Kale, A.; Nguyen, F.; Kay, M.; Hullman, J. Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 892–902. [CrossRef]
26. Bancilhon, M.; Liu, Z.; Ottley, A. Let's Gamble: How a Poor Visualization Can Elicit Risky Behavior. In Proceedings of the 2020 IEEE Visualization Conference (VIS), Salt Lake City, UT, USA, 25–30 October 2020; pp. 196–200.
27. Begoli, E.; Bhattacharya, T.; Kusnezov, D. The Need for Uncertainty Quantification in Machine-Assisted Medical Decision Making. *Nat. Mach. Intell.* **2019**, *1*, 20–23. [CrossRef]
28. Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–15.
29. Stone, P.; Jessup, S.A.; Ganapathy, S.; Harel, A. Design Thinking Framework for Integration of Transparency Measures in Time-Critical Decision Support. *Int. J. Hum.–Comput. Interact.* **2022**, *38*, 1874–1890. [CrossRef]
30. Heltne, A.; Frans, N.; Hummelen, B.; Falkum, E.; Germans Selvik, S.; Paap, M.C.S. A Systematic Review of Measurement Uncertainty Visualizations in the Context of Standardized Assessments. *Scand. J. Psychol.* **2023**, *64*, 595–608. [CrossRef] [PubMed]
31. Preston, A.; Ma, K.-L. Communicating Uncertainty and Risk in Air Quality Maps. *IEEE Trans. Vis. Comput. Graph.* **2023**, *29*, 3746–3757. [CrossRef]
32. Andrienko, N.; Andrienko, G.; Chen, S.; Fisher, B. Seeking Patterns of Visual Pattern Discovery for Knowledge Building. *Comput. Graph. Forum* **2022**, *41*, 124–148. [CrossRef]
33. Zhou, J.; Zheng, B.; Chen, F. Effects of Uncertainty and Knowledge Graph on Perception of Fairness. In Proceedings of the IUI'23 Companion: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, Sydney, NSW, Australia, 27–31 March 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 151–154.
34. van der Bles, A.M.; van der Linden, S.; Freeman, A.L.J.; Mitchell, J.; Galvao, A.B.; Zaval, L.; Spiegelhalter, D.J. Communicating Uncertainty about Facts, Numbers and Science. *R. Soc. Open Sci.* **2019**, *6*, 181870. [CrossRef]
35. Hullman, J. Why Authors Don't Visualize Uncertainty. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 130–139. [CrossRef] [PubMed]
36. Sterzik, A.; Lichtenberg, N.; Krone, M.; Baum, D.; Cunningham, D.W.; Lawonn, K. Enhancing Molecular Visualization: Perceptual Evaluation of Line Variables with Application to Uncertainty Visualization. *Comput. Graph.* **2023**, *114*, 401–413. [CrossRef]
37. Shin, D. Embodying Algorithms, Enactive Artificial Intelligence and the Extended Cognition: You Can See as Much as You Know about Algorithm. *J. Inf. Sci.* **2023**, *49*, 18–31. [CrossRef]
38. Ferrario, A.; Loi, M.; Viganò, E. In AI We Trust Incrementally: A Multi-Layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philos. Technol.* **2020**, *33*, 523–539. [CrossRef]
39. Cassenti, D.N.; Kaplan, L.M. Robust Uncertainty Representation in Human-AI Collaboration. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III, Online Only, 12–17 April 2021; SPIE: Philadelphia, PA, USA, 2021; Volume 11746, pp. 249–262.
40. Panagiotidou, G.; Vandam, R.; Poblome, J.; Moere, A.V. Implicit Error, Uncertainty and Confidence in Visualization: An Archaeological Case Study. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 4389–4402. [CrossRef] [PubMed]
41. Manjarrez, E.; DeLuna-Castruita, A.; Lizarraga-Cortes, V.; Flores, A. Similarity Index of Ex-Gaussian Reaction Time Signatures. *BioRxiv* **2023**. [CrossRef]
42. Castro-Palacio, J.C.; Fernández-de-Córdoba, P.; Isidro, J.M.; Sahu, S.; Navarro-Pardo, E. Human Reaction Times: Linking Individual and Collective Behaviour Through Physics Modeling. *Symmetry* **2021**, *13*, 451. [CrossRef]
43. Piccolotto, N.; Bögl, M.; Miksch, S. Visual Parameter Space Exploration in Time and Space. *Comput. Graph. Forum* **2023**, *42*, e14785. [CrossRef]