Estimating Trust in Human-Robot Collaboration through Behavioral Indicators and Explainability

Campagna, Giulio; Lagomarsino, Marta; Lorenzini, Marta; Chrysostomou, Dimitrios; Rehm, Matthias; Ajoudani, Arash

Published in:
IEEE Robotics and Automation Letters

DOI (link to publication from Publisher):
10.1109/LRA.2025.3600170

Publication date:
2025

Document Version
Early version, also known as pre-print

Estimating Trust in Human-Robot Collaboration through Behavioral Indicators and Explainability

Giulio Campagna[1], Marta Lagomarsino[2], Marta Lorenzini[2], Dimitrios Chrysostomou[3], Matthias Rehm[1], Arash Ajoudani[2]

***Abstract*—Industry 5.0 focuses on human-centric collaboration between humans and robots, prioritizing safety, comfort, and trust. This study introduces a data-driven framework to assess trust using behavioral indicators. The framework employs a Preference-Based Optimization algorithm to generate trust-enhancing trajectories based on operator feedback. This feedback serves as ground truth for training machine learning models to predict trust levels from behavioral indicators. The framework was tested in a chemical industry scenario where a robot assisted a human operator in mixing chemicals. Machine learning models classified trust with over 80% accuracy, with the Voting Classifier achieving 84.07% accuracy and an AUC-ROC score of 0.90. These findings underscore the effectiveness of data-driven methods in assessing trust within human-robot collaboration, emphasizing the valuable role behavioral indicators play in predicting the dynamics of human trust.**

***Index Terms*—Human Factors and Human-in-the-Loop; Acceptability and Trust; Human-Robot Collaboration**

I. INTRODUCTION

Industry 5.0 shifts to human-centered manufacturing, where collaborative robots work alongside human workers to enhance productivity and efficiency. By integrating technologies like sensors and machine learning, robots adapt to human actions and dynamic environments. Focusing on safety, adaptability, and Human-Robot Collaboration (HRC), it creates a resilient, sustainable industrial ecosystem that combines human decision-making with robotic precision for improved performance [1]. Trust is key for ensuring safe, comfortable, and seamless human-robot interaction.

Lee and See [2] define trust as the belief that an agent will support a person's goals, particularly in uncertain or vulnerable situations. Confidence in a robot's ability to perform tasks reliably is critical to effective human-robot interaction. Under-trust may lead to operator overload by undervaluing the robot's capabilities, while over-trust can cause safety risks such as equipment damage or collisions [3]. Trust is often assessed via post-interaction surveys (e.g., [4]), which fail to capture real-time fluctuations and may not reflect actual user behavior [5]. This highlights the need for online trust estimation to enable adaptive robot behavior that ensures safety and supports ergonomic collaboration.

Recently, data-driven approaches have emerged as effective solutions for online trust estimation. Xu and Dudek [6] introduced a dynamic Bayesian network to continuously estimate human trust in robots based on task performance and factors like failure rates, human interventions, and task outcomes. Shayesteh et al. [7] developed a method to evaluate trust in construction robots during collaborative tasks using EEG signals as input to machine learning models, proving highly effective. In our recent studies, body motion data [8] and facial expressions [9] were independently utilized as inputs to machine learning models for estimating trust levels in a chemical industry scenario, where a robot assisted a human operator with chemical delivery and mixing tasks. Finally, Lagomarsino et al. [10] proposed a reinforcement learning framework that adjusts interaction parameters based on a human-robot coefficiency metric, integrating human physical and cognitive factors, and robot operational costs to optimize joint efficiency, align with user preferences, enhance comfort, and foster trust.

Despite recent advancements, data-driven trust estimation models remain underexplored and struggle to capture the full complexity of trust dynamics [5]. Many rely on single-factor indicators, overlooking the integration of behavioral, cognitive, and contextual features that are critical for accurate modeling. Moreover, most existing models lack personalization, failing to account for individual differences—such as personality traits, emotional responses, or past experiences—which are essential for adapting robot behavior to the unique trust-building needs of each user. This limits the potential for fostering long-term, meaningful interactions. In addition, many models are not seamlessly integrated into robotic systems, preventing online behavioral adaptation in response to fluctuating trust levels. Leveraging behavioral trust indicators offers a promising path forward by enabling deeper insight into trust dynamics and supporting more adaptive, context-aware collaboration.

This work introduces a data-driven framework that extends beyond our previous study [11] by integrating human- and robot-related behavioral trust indicators with machine learning models to estimate human trust preferences online. While the *Preference-Based Optimization* (PBO) algorithm from our prior work is employed solely to generate robot trajectories based on explicit human feedback, the key innovation here lies in leveraging these behavioral indicators to train a machine learning model. The model predicts trust preferences by analyzing the dynamic relationship between interaction pa-

Manuscript received: February, 18, 2025; Revised May, 30, 2025; Accepted August, 4, 2025. This paper was recommended for publication by Editor Angelika Peer upon evaluation of the Associate Editor and Reviewers' comments. The work was supported in part by the European Union Horizon Project TORNADO (Grant No. 101189557) and in part by the Independent Research Fund Denmark (Grant No. 1032-00311B).

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by ASL3 Genovese under Application No. IIT_HRII_ERGOLEAN 156/2020, and performed in line with the Helsinki Declaration.

Corresponding author's email: gica@create.aau.dk

[1] Human-Robot Interaction Lab., Technical Faculty of IT and Design, Aalborg University, Denmark.

[2] Human-Robot Interfaces and Interaction Lab., Istituto Italiano di Tecnologia (IIT), Italy.

[3] Smart Production Lab., Faculty of Engineering and Natural Sciences, Aalborg University, Denmark.

Digital Object Identifier (DOI): see top of this page.

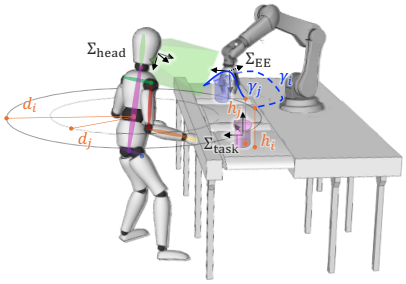


Fig. 1: Illustration of HRC in an industrial setting, highlighting operator whole-body tracking (head and upper body) and interaction parameters guiding the robotic manipulator's behavior.

rameters and trust indicators, enabling a more personalized and adaptive approach to HRC. Trust indicators are derived from human motion capture and robot motion characteristics during collaboration. To enhance interpretability, SHAP (SHapley Additive exPlanations) values are used to explain model predictions, highlighting how each indicator influences overall trust estimation and how these effects vary across individual human operators. The framework's effectiveness is evaluated in a chemical industry scenario, where a robotic arm assists a human in performing the potentially hazardous task of pouring chemicals.

The main contribution of this work is twofold:

- The development of a novel data-driven model that leverages both human and robot behavior to predict human trust preferences, enabling the optimization of key interaction parameters—specifically, the robot's execution time, the separation distance between human and robot, and the vertical proximity of the end-effector to the user's head—during collaborative tasks.
- A comprehensive analysis of behavioral trust indicators, with a focus on explainability and model-driven insights into which indicators most influence trust and how their effects differ across individuals.

The remainder of this paper is structured as follows: Section II details the methodology, followed by the experimental procedure in Section III. Section IV presents the results, and Section V provides the discussion and conclusions of the work.

II. METHODOLOGY

This section defines behavioral indicators from human movement sensor data during human-robot interactions to study human trust dynamics in robots. We then describe a machine-learning pipeline that uses these indicators to estimate human trust preferences for robot trajectories. Using PBO, we collected explicit user feedback comparing two interaction parameter sets shaping robot trajectories, providing ground-truth labels for model training. Finally, we evaluate model explainability to reveal how each behavioral indicator influences trust, identifying key factors shaping user preferences and highlighting individual variations. Understanding these trends enables adaptive, personalized robot behavior to foster trust effectively in human-robot interactions.

This paper proposes a rigorous pipeline to address the following research questions (RQs):

RQ1. *Can behavioral indicators and machine learning models predict user preferences for robot interaction parameters that promote trustworthy HRC?*

RQ2. *What is the contribution of each indicator to the estimation of trust levels?*

A. Definition of the Trust Indicators

Several indicators are proposed to capture trust dynamics in industrial HRC settings. These indicators are categorized into *human-related factors* and *robot-related factors*, following the categorization scheme used in [12]. By combining these parameters, the framework aims to provide a holistic view of trust, analyzing both human body language and perception of robot behaviors. The framework operates without assuming the relevance or direction of correlations between behavioral indicators and perceived trust. Instead, these relationships are determined by the machine learning model and explained through detailed analysis, forming the foundation for online trust modeling and effective collaboration.

1) Human-Related Trust Indicators: Human-related trust indicators reflect the cognitive and behavioral responses of the human operator during robot interaction, offering insights into their trust and comfort levels. The human body is modeled as a kinematic chain with N joints and $N+1$ rigid body segments, each associated with a frame Σ_i, where $i \in \{1,\dots,N\}$, relative to a global reference frame Σ_W. The global frame's position is calibrated initially using the motion tracking system, ensuring accurate tracking over time.

Human Attention to End-Effector - The proposed trust indicator examines the attention an individual directs toward the robot during a collaborative task. Insights from contemporary psychology suggest that average dwell time—defined as the mean duration a person's gaze remains fixed on a specific area, such as the robot's end-effector—decreases with increasing confidence and expertise [13]. Based on this premise, attention is a relevant parameter for assessing trust.

To evaluate the level of attention toward the end effector, we adopt the method proposed in [14]. A frame, namely Σ_{head}, is positioned at the center of the head and tilted by ten degrees to approximate the direction of the gaze [15]. The Cartesian vector representing the relative position between Σ_{gaze} (the estimated gaze frame) and Σ_{EE} (the end-effector frame) is converted into spherical coordinates, characterized by the azimuth angle $\theta_{EE} \in \mathbb{R}$, the elevation angle $\phi_{EE} \in \mathbb{R}$, and the radial distance $d_{\text{EE}} \in \mathbb{R}$. A fuzzy logic function that utilizes a Raised-Cosine Filter is then applied separately to normalize the attention angles θ_{EE} and ϕ_{EE}, which are denoted here as $\alpha(t)$:

$$\lambda(\alpha(t)) = \begin{cases} 1, & \text{if } |\alpha(t)| \le \alpha_{\min}(t), \\ \frac{1}{2}\left[1-\cos\left(\frac{|\alpha(t)|-\alpha_{\min}(t)}{\alpha_{\max}(t)-\alpha_{\min}(t)}\pi\right)\right], & \text{if } |\alpha(t)| > \alpha_{\min}(t) \\ & \;\&\; |\alpha(t)| \le \alpha_{\max}(t), \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The threshold values $\alpha_{\min}(t)$ and $\alpha_{\max}(t)$ depend on the current distance d_{EE} of the end-effector from the human

operator. These thresholds are calculated using the approach described in [16]:

$$\alpha_{\min}(t) = \tan^{-1}\left(\frac{(1-\delta)r_{\text{EE}}}{d_{\text{EE}}(t)}\right), \quad \alpha_{\max}(t) = \tan^{-1}\left(\frac{(1+\delta)r_{\text{EE}}}{d_{\text{EE}}(t)}\right), \tag{2}$$

where $r_{\text{EE}} \in \mathbb{R}$ represents a fixed predefined radius of the end-effector's containment area. The parameter $\delta \in \mathbb{R}$ is set to 0.4 to ensure smooth variation in the function $\lambda(\alpha(t))$.

In conclusion, the attention level $\Lambda_{\text{EE}}(t) \in [0,1]$ directed toward the end-effector is defined as the product of the normalized azimuth and elevation indicators:

$$\Lambda_{\text{EE}}(t) = \lambda(\theta_{\text{EE}}(t))\lambda(\phi_{\text{EE}}(t)). \tag{3}$$

When $\Lambda_{EE} \approx 1$, the human is actively monitoring the robot's motion; when $\Lambda_{EE} \approx 0$, the human shows no focus on the robot.

Human Attention to Task - This trust indicator monitors the level of attention the human directs toward the task itself. Similar to the *Human Attention to End-Effector* indicator, we compute the relative position between Σ_{gaze} and a fixed reference frame Σ_{task}, which defines the area where the task is performed and uses r_{task} to delimit the task area. Consequently, we apply Eq. (1) and (3) to evaluate the current level of attention $\Lambda_{\text{task}}(t) \in [0,1]$.

Spatial Displacement - This indicator evaluates the human operator's spatial deviation from the initial position. The spatial dynamics of human-robot interaction offer valuable insights into trust. Some studies suggest that increased human displacement may indicate discomfort, stress, or uncertainty about the robot's performance [8], [17]. Conversely, other research shows that a lack of confidence in the robot can cause humans to "freeze", waiting for the robot to finish its movement before continuing the task [18].

To calculate this indicator, the displacement $\Delta p \in \mathbb{R}$ is defined as the normalized distance between the current position of the human head, $\mathbf{p}_{\text{head}} \in \mathbb{R}^3$, and its initial position at the start of the task, $\mathbf{p}_{\text{init}} \in \mathbb{R}^3$:

$$\Delta p(t) = \frac{\|\mathbf{p}_{\text{head}}(t) - \mathbf{p}_{\text{init}}\|}{d_{\max}}, \tag{4}$$

where $d_{\max} \in \mathbb{R}$ is the maximum allowable distance between the head and the fixed point. This ensures that $\Delta p(t)$ is normalized to the range $[0,1]$, with $\Delta p(t) \approx 0$ corresponding to minimal head movement and $\Delta p(t) \approx 1$ indicating significant motion.

Human-Robot Speed Synchronization - Previous studies have shown that smooth coordination and synchronized movements between humans and robots are strongly correlated with a positive perception and trust in the robot. Conversely, significant speed discrepancies indicate misalignment and potential trust issues [19], [20]. To address this, we define an indicator that measures the alignment between the speed of the human hand and the robot's end-effector during the task.

The relative speed match, $\Delta \text{v} \in [0,1]$, is computed as:

$$\sigma(t) = 1 - \left| \frac{\text{v}_{\text{hand}}(t) - \text{v}_{\text{min,hand}}}{\text{v}_{\text{max,hand}} - \text{v}_{\text{min,hand}}} - \frac{\text{v}_{\text{EE}}(t) - \text{v}_{\text{min,EE}}}{\text{v}_{\text{max,EE}} - \text{v}_{\text{min,EE}}} \right|, \tag{5}$$

where $v_{\text{hand}} \in \mathbb{R}$ and $v_{\text{EE}} \in \mathbb{R}$ represent the magnitudes of the human hand velocity and the robot end-effector velocity, respectively. $v_{\text{min,hand}}$ and $v_{\text{max,hand}}$ denote the minimum and maximum observed velocity values for the human hand, while $v_{\text{min,EE}}$ and $v_{\text{max,EE}}$ represent the lower and upper bounds of the robot end-effector velocity. As a result, $\sigma \approx 0$ indicates a significant speed mismatch, while larger values signify closely matched velocities.

Reaction Time - This measure quantifies the temporal latency between the robot's actions and the human operator's responses. Research suggests that tasks involving complex decision-making may increase delays in movement initiation [21], reflecting heightened cognitive demands. As a key trust factor, longer reaction times may denote uncertainty or reduced confidence in the robot's performance.

Let $t_R \in \mathbb{R}$ represent the timing of the robot's motion initiation, and $t_H \in \mathbb{R}$ the human's action start time. The temporal misalignment $\Delta t \in [0,1]$ is defined as:

$$\Delta t = \frac{|t_H - t_R|}{\tau}, \tag{6}$$

where $\tau \in \mathbb{R}$ is the total execution time of the robot's trajectory. A value of $\Delta t \approx 0$ indicates near-perfect synchronization, while $\Delta t \approx 1$ suggests a significant delay in the human's response to the robot.

2) Robot-Related Trust Indicators: Robot-related trust indicators are calculated based on the robot's behavioral characteristics, which are known to influence human perception and trust in the interaction.

Predictability - Existing research has demonstrated that smooth robot maneuvers with low jerk values enhance predictability by aligning with human expectations, thereby increasing operator confidence and fluidity in interaction [22], [23]. This effect is likely due to their similarity to natural arm motions [24]. In contrast, high-jerk movements can appear erratic, reducing predictability and trust.

To quantify the impact of robot smoothness on human trust, an indicator is defined as the normalized integral of the squared jerk over the time interval $[t_0, t_f]$, where t_0 and t_f represent the starting and ending instants of time of the robot's trajectory. The jerk $j(t) \in \mathbb{R}$, which is the magnitude of the third derivative of position, is constrained by the feasible minimum and maximum jerk values of the robot's end effector, $j_{\min}$ and $j_{\max} \in \mathbb{R}$. The *Predictability* indicator $\rho(t) \in [0,1]$ is then computed as:

$$\rho(t) = 1 - \frac{\int_{t_0}^{t_f} j(t)^2 dt - j_{\min}^2 \tau_{\min}}{j_{\max}^2 \tau_{\max} - j_{\min}^2 \tau_{\min}}, \tag{7}$$

where $\tau_{\min}$ and $\tau_{\max}$ represent the minimum and maximum durations of the robot's trajectory. A $\rho \approx 0$ value reflects significant jerk fluctuations, while $\rho \approx 1$ signifies smooth and consistent motion.

Legibility - This factor evaluates the curvature of the robot's trajectory, which is used in literature as a measure of legibility—how clearly the robot's motion communicates its intent or goal to a human observer based on a partially observed trajectory [25].

The curvature coefficient $\kappa \in \mathbb{R}$ of the robot end-effector trajectory $\gamma(t)$ is computed as:

$$\kappa(t) = \frac{\sqrt{\|\dot{\gamma}(t)\|^2 \|\ddot{\gamma}(t)\|^2 - (\dot{\gamma}(t) \cdot \ddot{\gamma}(t))^2}}{\|\dot{\gamma}(t)\|^3}, \tag{8}$$

where $\dot{\gamma}(t)$ and $\ddot{\gamma}(t)$ are the first and second derivative, representing the robot's velocity and acceleration, respectively.

The radius of curvature $r \in \mathbb{R}$, which reflects the size of the approximating circle at a point on the trajectory, is then calculated as $\left(\frac{1}{k}\right)$. Using the vector cross product identity $|\mathbf{a}\times\mathbf{b}|^2 = |\mathbf{a}|^2|\mathbf{b}|^2 - (\mathbf{a}\cdot\mathbf{b})^2$, we obtain:

$$r(t) = \frac{\|\dot{\gamma}(t)\|^3}{\|\dot{\gamma}(t)\times\ddot{\gamma}(t)\|}. \tag{9}$$

The *Legibility* indicator $L \in [0,1]$ is obtained by scaling the radius relative to its maximum value $r_{\max}$:

$$L(t) = \frac{r(t)}{r_{\max}}. \tag{10}$$

A value of $L \approx 0$ corresponds to a robot trajectory with a smaller radius of curvature. In contrast, when $L \approx 1$, the trajectory is associated with a larger radius of curvature, potentially making its actions easier to interpret.

B. Dataset Collection

To evaluate the relevance of the behavioral indicators for modeling human trust in the robot and confidence in HRC, the study relies on reliable ground truth data. As obtaining absolute trust values in realistic scenarios is challenging, the adaptive trajectory framework presented in [11] is adopted. This framework employs the PBO algorithm [26] to iteratively refine the interaction parameters that shape the robot's trajectory, based on repeated explicit trust feedback. At each iteration, new parameters are selected by minimizing an acquisition function that balances surrogate model exploitation and exploration, as detailed in [11].

1) Preference Labels: Specifically, for a robot's trajectory from point **A** to point **B**, three key interaction parameters (refer to Fig. 1) that influence the user's trust are adjusted to align the robot's behavior with human preferences and expectations:

i) the *total execution time* τ of the trajectory, which determines the robot's velocity profile;
ii) the *separation distance* d maintained by the robot from the user's body;
iii) the *maximum height* h, controlling the vertical proximity of the robot's end effector to the user's head.

These parameters form the decision vector $\mathbf{x} = [\tau, d, h] \in \mathbb{R}^3$. At the first iteration, the PBO algorithm proposes two distinct parameter sets $\mathbf{x}_1, \mathbf{x}_2 \in X$, where $\mathbf{x}_1 \neq \mathbf{x}_2$. The human operator is then asked to indicate a preference using the *preference function* $\pi : \mathbb{R}^n \times \mathbb{R}^n \to \{-1, 1\}$, defined as:

$$\pi(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} -1 & \text{if } \mathbf{x}_1 \text{ is preferred over } \mathbf{x}_2 \\ 1 & \text{if } \mathbf{x}_2 \text{ is preferred over } \mathbf{x}_1 \end{cases} \tag{11}$$

Each parameter set $\mathbf{x}_i$ refines the robot's trajectory γ to reach point **B** as follows: i) the total trajectory duration is set to τ_i; ii) a circle with radius d_i around the human defines a protected zone, with waypoints inside this zone adjusted to lie on the circle's edge; iii) the z-component of the trajectory is modified so that the motion's maximum height corresponds to h_i. Details on trajectory adaptation can be found in [11].

In each iteration, a new trajectory γ, defined by a parameter set $\mathbf{x}_i$, is proposed to the human operator. The operator selects the trajectory that best promotes trust, either by retaining the previous best trajectory (assigning -1 to π) or by adopting the refined trajectory (assigning 1 to π). This iterative process identifies the optimal parameter vector, $\mathbf{x}^*$, that minimizes the preference function π within the bounded domain $[\mathbf{x}_{\min}, \mathbf{x}_{\max}]$, thereby maximizing trust in the robot. This work focuses on participant preferences when comparing two robot behaviors. While these do not directly measure absolute trust, they provide valuable insight into trust dynamics across behaviors. Combined with human behavioral indicators, the labels indicate whether parameter changes increased or decreased trust and confidence.

2) Data Pre-Processing and Trust Indicators Selection: To prepare the input data for the machine learning model, the average value of each trust indicator is computed for each task iteration, per participant. Given that human preferences are represented by binary labels (-1 for the previously preferred trajectory, 1 for the current one), we compute differences in averaged trust indicator values between the two trajectories. These differences capture the variation in average feature values between the current and previously preferred parameter sets.

A *Pearson correlation* analysis is performed to identify potential redundancies among the trust indicators. The Pearson correlation coefficient c between two variables A and B is calculated as:

$$c = \frac{\sum_{h=1}^{o}(A_h - \bar{A})(B_h - \bar{B})}{\sqrt{\sum_{h=1}^{o}(A_h - \bar{A})^2}\sqrt{\sum_{h=1}^{o}(B_h - \bar{B})^2}} \tag{12}$$

where A and B are the two trust indicators being compared, A_h and B_h represent the h-th observation of A and B, $\bar{A}$ and $\bar{B}$ are their respective mean values, and o is the total number of observations. The *tree-based feature importance* method evaluates the contribution of trust indicators to model predictions. *XGBoost* is selected for its robustness against overfitting and its boosting algorithm, which captures complex feature patterns. Feature importance is measured using the *gain* metric, reflecting each feature's average performance improvement during training. A *5%* threshold of the highest importance score filters for the most influential features in further analysis.

The final preprocessing step applies data augmentation to generate synthetic samples, enhancing dataset robustness. Due to the limited size of the original dataset, this was essential to mitigate overfitting and improve generalization. In low-data regimes, models often mistake noise for signal, resulting in unreliable outcomes. To address this, we employed *probabilistic sampling*, drawing synthetic data independently for each feature from *univariate Gaussian distributions* fitted to the global empirical mean and variance. This preserved the dataset's statistical structure while increasing sample density, yielding a more stable and generalizable learning process.

C. Machine Learning Classification of Human Preferences

After identifying the relevant behavioral indicators for trust modeling, the next step is to evaluate whether these indicators can classify participants' expressed trust preferences for different robot behaviors. To classify human preferences for robot trajectories, we employed *Random Forest*, *K-Nearest Neighbors*

TABLE I: Hyperparameter Tuning Summary for Each Model.

Model	Hyperparameter	Description	Option Values	Optimized Value
Random Forest	max_depth	Max depth of the tree	[None, 10, 20]	None
	min_samples_leaf	Min samples at a leaf node	[1, 2, 4]	1
	min_samples_split	Min samples to split a node	[2, 5, 10]	2
	n_estimators	Number of trees	[50, 100, 150]	150
KNN Classifier	n_neighbors	Number of neighbors	[3, 5, 7, 10]	3
	weights	Weight function	['uniform', 'distance']	distance
	algorithm	Nearest neighbors algorithm	['auto', 'ball_tree', 'kd_tree', 'brute']	auto
SVM	C	Regularization	[0.1, 1, 10, 100]	100
	kernel	Kernel type	['linear', 'rbf', 'poly']	rbf
	gamma	Kernel coefficient	['scale', 'auto']	scale
Voting Classifier	voting method	Prediction combination method	['hard', 'soft']	soft
	base models	Included models	[Random Forest, KNN, SVM]	Optimized configurations

(KNN), and *Support Vector Machine* (SVM). *Random Forest* reduces overfitting and handles noisy data through ensemble learning. *KNN* captures local patterns by classifying based on neighbor proximity. *SVM* excels in high-dimensional spaces, providing clear decision boundaries and strong generalization to distinguish subtle trust differences. Additionally, a *Voting Classifier* was implemented to combine predictions from these models, leveraging their strengths to enhance reliability, mitigate overfitting, and improve classification performance.

The dataset was split into training and testing sets, with *80%* allocated to *training* and *20%* to *testing*. *Stratification* was applied to maintain consistent class distribution across both sets. With reference to hyperparameter optimization, the *GridSearchCV* technique was employed to fine-tune the model's parameters. With reference to [27], *Nested cross-validation* was used to assess model performance and minimize the risk of overfitting. Both the outer and inner cross-validation processes utilized *Stratified K-Fold Cross-Validation* with *5 folds*, preserving the class distribution across each fold. This approach is particularly advantageous for small datasets as it provides a more reliable estimate of model performance. In the inner cross-validation, *GridSearchCV* explored a specified range of hyperparameters, optimizing the classifier based on accuracy. Meanwhile, the outer cross-validation assessed the model's generalization to unseen data, yielding nested accuracy scores that reflect its reliability. This comprehensive approach leverages all available data while preserving the target class distribution. The tested hyperparameter ranges and the final selections for each model are summarized in Table I.

D. Model Explainability to Analyze the Influence of Behavioral Trust Indicators

To gain insights into the influence of behavioral trust indicators, we employed a model explainability technique to quantify the contribution and impact of each feature on the model's predictions. Specifically, we utilized SHAP values, which enhance transparency in the model's decision-making process by identifying the most influential trust indicators and determining their positive or negative contributions to the predictions. As described in Section II-B2, the input to the model consists of differences in the averaged trust indicator values between two proposed trajectories. A positive value for an input feature indicates an increase in the trust indicator with the new proposed trajectory, while a negative value indicates a decrease. Similarly, a positive SHAP value associated with a model prediction signifies a contribution to the positive class (indicating a preference for the new trajectory), whereas a negative SHAP value reflects a contribution to the negative class (indicating a preference for the previously considered best trajectory). This approach allows us to evaluate the relationship between variations in trust indicators and expressed trust preferences, determining whether an increase (or decrease) in the indicator value with the proposed trajectory fosters trust (if the new trajectory is preferred) or diminishes trust (if the previous trajectory is preferred), and, consequently, whether higher or lower indicator values are more desirable.

SHAP values define weights for the behavioral trust indicators in Section II-A1, enabling a *continuous trust score* to support online adaptation of robot behavior to foster trust. Since trust is subjective, operators may prefer different trust-enhancing behaviors, reflected in their behavioral indicators. Personalizing the *continuous trust score* with SHAP values allows quantifying the contribution of each indicator for individual users.

III. EXPERIMENTS

The following section outlines the task and experimental procedure used to evaluate the proposed framework, and describes the sensor data acquisition setup for calculating the trust indicators defined in Section II-A.

A. Task Description and PBO Parameters

The proposed scenario featured a cyclic HRC task set in a chemical industrial environment, as illustrated in Fig. 2. In this setup, the robot manipulator assisted the human operator in the process of mixing chemicals. The task began with the participant placing a beaker labeled B_1, containing a chemical, at location (**C**). Simultaneously, the robot retrieved another beaker, B_2, located at (**A**) and pre-filled with a chemical. To enhance realism, participants were told that the chemicals were hazardous, though coarse salt was used as a safe substitute.

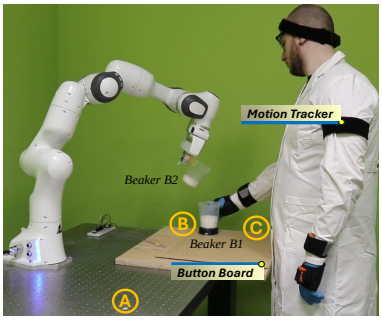


Fig. 2: The experimental scenario with the robot assisting the human operator in mixing chemicals.

Upon hearing a cue, the participant moved beaker B_1 from (**C**) to (**B**) and held it steady, while the robot transported beaker B_2 to (**B**) and poured its contents into beaker B_1. Once pouring was complete, the robot returned beaker B_2 to (**A**), and the participant returned beaker B_1 to (**C**).

Regarding the interaction parameters described in Section II-B, the *total execution time* τ, representing the time taken for the robot to move beaker $B2$ to the pouring position **B**, was optimized by the PBO within the range $[5,10]$s. The *separation distance* d varied within $[0.52,0.65]$m, while the *maximum height* h corresponded to the pouring height, ranging from $[0.26,0.40]$m. The maximum separation distance $d_{\max}$ was estimated empirically from preliminary trials based on the maximum observed head displacement during task execution. A similar procedure was used to determine $v_{\min/\max,\text{hand}}$ and $v_{\min/\max,\text{EE}}$, based on the observed velocity ranges of the human hand and the robot end-effector, respectively.

B. Experimental Setup

The robot manipulator used in this study was the *Franka Emika Panda*, equipped with a two-fingered parallel gripper and controlled at 1 kHz via the *Robot Operating System* (ROS).

The *Xsens MVN Awinda* motion tracking system was utilized to record the initial head position $\mathbf{p}_{\text{init}}$ following calibration and to monitor head and body movements throughout the task. A button located at (**C**) detected the moment when beaker B_1 was raised (indicated by the button being unpressed), marking the initiation time of human motion (t_H).

C. Experimental Protocol

The study included 14 healthy participants (8 males and 6 females, aged 29.21 ± 4.90 years), recruited from the student body and research staff at the Istituto Italiano di Tecnologia[1].

Participants completed the task 16 times, providing feedback after each cycle for a total of 15 evaluations. Initially, 2 cycles with different parameter sets were conducted, and participants indicated their preference (*feedback*) between them. Subsequently, at the end of each cycle, participants compared the current parameter set proposed by the PBO to their preferred one so far. If needed, they could repeat the robot trajectory with the previous preferred set.

[1]Experiments were carried out in accordance with the Helsinki Declaration, with the protocol receiving approval from the ASL3 Genovese ethics committee (Protocol IIT_HRII_ERGOLEAN 156/2020).

IV. EXPERIMENTAL RESULTS

This section presents key findings from evaluating the machine learning model that categorizes human preferences for trust-enhancing robot behaviors in industrial HRC. The evaluation assesses predictive accuracy and reliability using human- and robot-related trust indicators, and examines how these behavioral indicators influence predictions and shape trust levels.

A. Performance Analysis of Trust Classification Models

Before evaluating the model, the dataset underwent comprehensive pre-processing (detailed in Section II-B2) to create a compact and meaningful dataset. The *Pearson correlation* analysis revealed a correlation coefficient of $r = 0.65$ between the two trust indicators, *Human Attention to Task* and *Human Attention to End-Effector*, indicating redundancy in the model input data. For effective feature selection, a *tree-based feature importance* analysis using *XGBoost* was conducted to assess the contribution of each indicator. While all features were initially considered valid, *Human Attention to End-Effector* was excluded due to its redundancy, as highlighted by the correlation analysis, and its lower importance in the feature importance evaluation. To enhance robustness, *data augmentation* was performed on the initial dataset of 210 samples. A total of 630 synthetic samples were generated (three times the original size), striking a balance between increasing sample size and preserving inherent data patterns. *Nearest Neighbor Labeling* assigned labels to synthetic samples based on the majority vote of the 5 nearest original data points. This process increased the dataset size to 840 samples. To ensure class balance and avoid bias, *Synthetic Minority Over-sampling Technique* was applied to generate additional samples for the minority class, achieving a balanced class distribution in the augmented dataset.

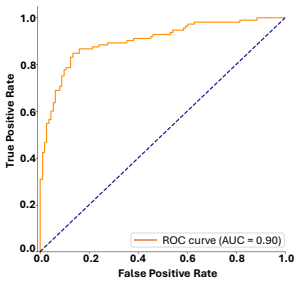
Table II summarizes the performance of all tested models. The *Voting Classifier* was the most robust, achieving an accuracy of 84.07%, surpassing individual models in both classification accuracy and stability. Its performance was further validated with the *Receiver Operating Characteristic* (ROC) curve (Fig.3a), which illustrates the true positive and false positive rates, and an *Area Under the Curve* (AUC) of 0.90, confirming strong class distinction. The confusion matrix (Fig.3b) shows true/false positives and negatives. Additionally, the model achieved a *precision* of 0.85, *recall* of 0.84, and an *F1 score* of 0.84, indicating a balanced performance with reliable positive identification and a good trade-off between false positives and false negatives.

B. Model Explainability with SHAP Analysis

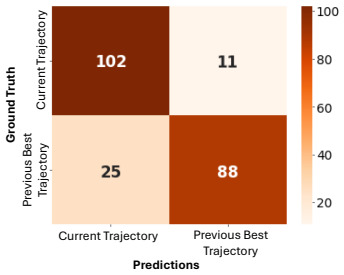
Fig. 4 illustrates the explainability of the model using SHAP values, which reveal the contributions of individual features to the *Voting Classifier*'s decision-making process. Specifically, the figure highlights the influence of behavioral indicators on human trust preferences in robot trajectories. Each dot represents a prediction, with its color indicating the feature value. Positive SHAP values indicate predicted preference for the new trajectory, while negative values reflect preference for

TABLE II: Machine learning models and performance indicators.

Model	Accuracy	AUC	Precision	Recall	F1-score
Random Forest	80.09%	0.87	0.81	0.80	0.80
KNN Classifier	80.97%	0.87	0.83	0.81	0.81
SVM	80.02%	0.84	0.80	0.80	0.80
Voting Classifier	84.07%	0.90	0.85	0.84	0.84



(a) ROC Curve with an AUC value of 0.90.



(b) Confusion Matrix.

Fig. 3: Evaluation of the Voting Classifier.

the previous best trajectory. Features are ranked by impact, with *Reaction Time* and *Human Attention to Task* being the most influential, with average absolute SHAP values of 0.197 and 0.183, respectively.

Notably, clusters of consistently colored points appear for each feature, suggesting that similar feature values consistently guided predictions in the same direction, which highlights the model's robustness in leveraging these indicators. Specifically, notable variations in *Reaction Time*—including decreases (represented by light blue dots) and substantial increases (depicted by light red dots)—are associated with negative SHAP values, indicating a lower level of trust in the newly proposed trajectory. In contrast, significant reductions in *Attention to the Task* (represented by light blue dots) typically correspond with a trust-driven preference for the new trajectory (associated with positive SHAP values), potentially signaling confidence in the robot's motion. For other indicators, the trends are more consistent and monotonic. Higher values of *Human-Robot Speed Synchronization* correlate with greater trust, suggesting that human operators prefer new trajectories that facilitate synchronization. For *Spatial Displacement*, increased movement by the human is associated with higher trust, reflecting a preference for the new trajectory, whereas reduced movement or freezing corresponds to diminished trust. Contrary to expectations, for *Legibility* and *Predictability*, significant increases in feature values (such as larger curvature radii or less jerky motions) do not consistently lead to an enhancement in trust, possibly because human operators may prioritize rapid, direct movements over smoother trajectories, which could be perceived as less efficient or slower. Note that similar, though expectedly less distinct, feature trends were identified in the SHAP analysis conducted without data augmentation.

Given the subjectivity inherent in trust and body language indicators, the analysis was also conducted on a per-participant basis. The importance of trust indicators for two different participants is shown in Fig. 5. For Participant 6, the most

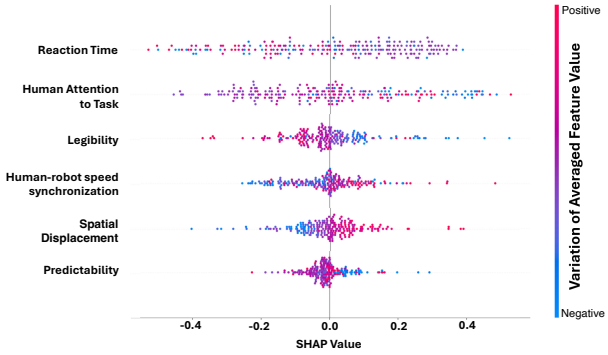
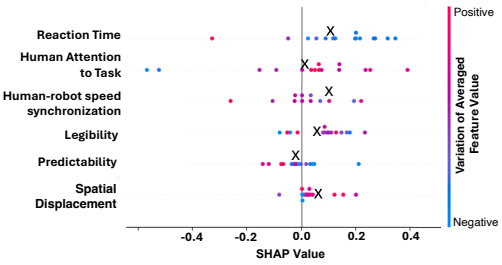
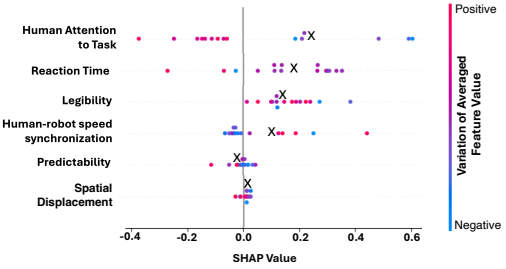


Fig. 4: SHAP plot showing feature contributions to model predictions.



(a) Impact of trust indicators for participant 6.



(b) Impact of trust indicators for participant 14.

Fig. 5: Importance of trust-related behavioral indicators for two participants, with X representing the mean SHAP values.

influential trust indicators were *Reaction Time* (average absolute SHAP = 0.186), where a decrease in value indicated higher trust in the robot, and *Human Attention to Task* (average absolute SHAP = 0.185), where an increase in value signaled a preference for the new trajectory. Interestingly, this deviated from the general trend observed across all participants in Fig. 4. Conversely, for Participant 14, the most influential feature was the variation in averaged *Human Attention to Task* (average absolute SHAP = 0.247), where lower attention was correlated with higher trust. The second most influential feature was *Reaction Time* (average absolute SHAP = 0.202), which similarly indicated higher trust at lower values.

V. DISCUSSION AND CONCLUSIONS

This paper presented a data-driven method for estimating trust based on human- and robot-related behavioral indicators. Using PBO, we gathered explicit user feedback on preferences between pairs of interaction parameters that determined the robot's trajectory in a collaborative chemical task. These preferences served as labels for a machine learning model

trained to predict trust preferences based solely on the proposed indicators.

Experimental results showed that machine learning models using human behavioral indicators can effectively predict human trust preferences in HRC (**RQ1**). Among the models tested, the *Voting Classifier* outperformed individual models such as Random Forest, KNN, and SVM, achieving an accuracy of *84.07%* and an *AUC* of *0.90* in selecting the preferred robot trajectory. This model could enable the automatic optimization of robot interaction parameters—such as human-robot separation distance, robot velocity, and vertical distance from the head—based on body language analysis, eliminating the need for direct user feedback.

The model explainability analysis revealed general trends in the proposed behavioral indicators and their relationship to trust-related preferences in robot trajectories (**RQ2**). This analysis identified consistent variations (increases or decreases) in behavioral indicators that signify an increase in trust across participants, revealing high or low values of specific indicators are preferable for promoting trust. Additionally, the subject-specific analysis of SHAP values highlighted the varying importance of trust-related features across participants, emphasizing the need for personalized trust models. This work lays the foundation for a continuous, personalized trust model that computes normalized trust indicators online and combines them into a weighted sum, using absolute average SHAP values and trend-based signs. The resulting online trust estimate informs the robot, enabling it to evaluate whether its adaptations and optimizations positively influence human trust.

Despite its promising contributions, the study has limitations. The framework was tested in a controlled pouring task within a chemical industry scenario, with researchers overseeing the process. A natural next step is to validate this approach across a broader range of tasks, particularly since the methodology was designed to be applicable to various HRC scenarios involving close-proximity interaction and robot trajectory planning. Additionally, we plan to deploy and validate the online model for trust assessment and trust-driven robot behavior adaptation, with the goal of fostering mutual understanding and trustworthy collaboration. Further research will also examine the integration of advanced learning models, such as deep learning, to enhance system performance and adaptability.

REFERENCES

[1] R. Gervasi, L. Mastrogiacomo, and F. Franceschini, "An experimental focus on learning effect and interaction quality in human–robot collaboration," *Production Engineering*, vol. 17, no. 3, pp. 355–380, 2023.

[2] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[3] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[4] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the "trust perception scale-hri"," in *Robust intelligence and trust in autonomous systems*. Springer, 2016, pp. 191–218.

[5] G. Campagna and M. Rehm, "A systematic review of trust assessments in human–robot interaction," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 2, pp. 1–35, 2025.

[6] A. Xu and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 221–228.

[7] S. Shayesteh, A. Ojha, and H. Jebelli, "Workers' trust in collaborative construction robots: Eeg-based trust recognition in an immersive environment," *Automation and robotics in the architecture, engineering, and construction industry*, pp. 201–215, 2022.

[8] G. Campagna, M. Dadgostar, D. Chrysostomou, and M. Rehm, "A data-driven approach utilizing body motion data for trust evaluation in industrial human-robot collaboration," in *33rd IEEE International Conference on Robot and Human Interactive Communication, IEEE RO-MAN 2024*. IEEE, 2024.

[9] G. Campagna, D. Chrysostomou, and M. Rehm, "Analysis of facial features for trust evaluation in industrial human-robot collaboration," in *2024 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 2024, pp. 1–6.

[10] M. Lagomarsino, M. Lorenzini, M. D. Constable, E. De Momi, C. Becchio, and A. Ajoudani, "Maximising coefficiency of human-robot handovers through reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4378–4385, 2023.

[11] G. Campagna, M. Lagomarsino, M. Lorenzini, D. Chrysostomou, M. Rehm, and A. Ajoudani, "Promoting trust in industrial human-robot collaboration through preference-based optimization," *IEEE Robotics and Automation Letters*, 2024.

[12] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[13] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, J. Halszka, and J. van de Weijer, *Eye Tracking : A Comprehensive Guide to Methods and Measures*. Oxford University Press, 2011, p. 187.

[14] M. Lagomarsino, M. Lorenzini, P. Balatti, E. De Momi, and A. Ajoudani, "Pick the right co-worker: Online assessment of cognitive ergonomics in human–robot collaborative assembly," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 4, pp. 1928–1937, 2022.

[15] U. Weidenbacher, G. Layher, P.-M. Strauss, and H. Neumann, "A comprehensive head pose and gaze database," in *2007 3rd IET International Conference on Intelligent Environments*. IET, 2007, pp. 455–458.

[16] M. Lagomarsino, M. Lorenzini, and A. Ajoudani, "PRO-MIND: Proximity and Reactivity Optimisation of robot Motion to tune safety limits, human stress, and productivity in INDustrial setting," *IEEE Transactions on Robotics*, vol. 41, pp. 2067–2085, 2025.

[17] G. Campagna and M. Rehm, "Analysis of proximity and risk for trust evaluation in human-robot collaboration," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 2191–2196.

[18] J. Sweller, "Cognitive load theory," in *Psychology of learning and motivation*. Elsevier, 2011, vol. 55, pp. 37–76.

[19] T. Kanda, H. Ishiguro, M. Imai, and T. Ono, "Body movement analysis of human-robot interaction," in *Proceedings of Int. Joint Conference on Artificial Intelligence*, 2003.

[20] W. Bartkowski, A. Nowak, F. I. Czajkowski, A. Schmidt, and F. Müller, "In sync: Exploring synchronization to increase trust between humans and non-humanoid robots," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–14.

[21] M. A. Khan, I. M. Franks, D. Elliott, G. P. Lawrence, R. Chua, P.-M. Bernier, S. Hansen, and D. J. Weeks, "Inferring online and offline processing of visual feedback in target-directed movements from kinematic data," *Neuroscience & Biobehavioral Reviews*, vol. 30, no. 8, pp. 1106–1121, 2006.

[22] B. Kühnlenz and K. Kühnlenz, "Reduction of heart rate by robot trajectory profiles in cooperative hri," in *Proceedings of International Symposium on Robotics (ISR)*. IEEE, 2016, pp. 400–406.

[23] M. Lagomarsino, M. Lorenzini, E. De Momi, and A. Ajoudani, "Robot trajectory adaptation to optimise the trade-off between human cognitive ergonomics and workplace productivity in collaborative tasks," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 663–669.

[24] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *The Journal of Neuroscience*, vol. 5, no. 7, pp. 1688–1703, 1985.

[25] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.

[26] A. Bemporad and D. Piga, "Global optimization based on active preference learning with radial basis functions," *Machine Learning*, vol. 110, pp. 417–448, 2021.

[27] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PloS one*, vol. 14, no. 11, p. e0224365, 2019.