IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Reconciling Conflicting Intents: Bidirectional Trust-Based Variable Autonomy for Mobile Robots

Yinglin Li, Rongxin Cui, *Member, IEEE*, Weisheng Yan, Shi Zhang, and Chenguang Yang, *Fellow, IEEE*

***Abstract*—In the realm of semi-autonomous mobile robots designed for remote operation with humans, current variable autonomy approaches struggle to reconcile conflicting intents while ensuring compliance, autonomy, and safety. To address this challenge, we propose a bidirectional trust-based variable autonomy (BTVA) control approach. By incorporating diverse trust factors and leveraging Kalman filtering techniques, we establish a core abstraction layer to construct the state-space model of bidirectional computational trust. This bidirectional trust is integrated into the variable autonomy control loop. Real-time modulation of the degree of automation is achieved through variable weight receding horizon optimization. Through a within-group experimental study with twenty participants in a semi-autonomous navigation task, we validate the effectiveness of our method in goal transfer and assisted teleoperation. Statistical analysis reveals that our method achieves a balance between rapid response and trajectory smoothness. Compared with binary control switching, this method reduces operator workload by 14.3% and enhances system usability by 9.9%.**

***Index Terms*—Bidirectional trust, conflicting intents, degree of automation, human-robot collaboration, variable autonomy.**

I. Introduction

REMOTE robotics has revolutionized human capabilities, unlocking new frontiers in accessing challenging environments such as the deep sea, outer space, and post-disaster zones [1]. Despite these advancements, robots still lack full autonomy in task execution due to limitations in situational awareness and decision-making [2]. To overcome these limitations, collaborative semi-autonomous robotic systems have been developed, integrating the strengths of robots and humans through variable autonomy [3] and shared control [4]. However, the autonomous natures of both entities often leads to conflicting intents. Fig. 1 illustrates a typical task scenario where the robot autonomously navigates to the predefined goal D (robot intent). During task execution, the operator at point B notices a point of interest E (human intent), prompting a transfer in the team's task goal from D to E. This necessitates the operator's intervention for robot control. The operator adjusts the path, directing the robot towards the noticed point of interest for closer observation. Upon fulfilling the human intent, the task goal transfers back from E to D. The operator then relinquishes control, enabling the robot to re-plan and resume operations in line with its original robot intent. The paramount focus is on goal transfer during human intervention or control relinquishment to ensure compliance or autonomy, complemented by assisted teleoperation to ensure safety. Designing a variable autonomy method to tackle these two issues remains a significant challenge.

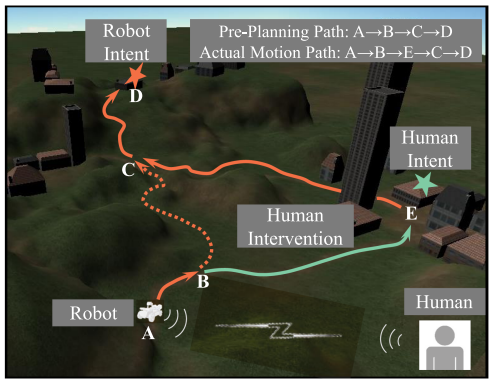


Fig. 1. Examples of intent conflicts arising from human intervention in navigation tasks. The robot pre-planned a path from the starting point A to the goal D as A → B→C → D. However, intervention by the operator at point B and its release at point E altered the robot's path to become A → B→E → C→D.

Variable autonomy systems facilitate control exchange between human operators and robots by transitioning between different levels of automation (LoA), reflecting the degree of independent decision-making and action by robots or artificial intelligent (AI) agents [5]. In [6], a research group presented an expert-guided mixed-initiative control switcher for remotely operating mobile robots, allowing LoA transitions initiated by human operators or in response to deteriorating robot performance. By addressing control conflicts from a negotiation perspective [7] and considering human state [8], they tackled the issue. Measuring decision variables is critical for achieving variable autonomy [9]. Existing work either focuses on reacting to sensor inputs (such as obstacles) [10], [11], predefined priorities [12], and haptic feedback [11] to ensure teleoperation safety, or on binary switching modes [13], [14] and timing of LoA [6],

Manuscript received 15 March 2024; accepted 23 April 2024. Date of publication 2 May 2024; date of current version 8 May 2024. This letter was recommended for publication by Associate Editor Francois Ferland and Editor Angelika Peer upon evaluation of the reviewers' comments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant U22A2066 and Grant U21B2047. *(Corresponding author: Rongxin Cui.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Northwestern Polytechnical University under Application No. 202302026, and performed in line with the Helsinki Declaration.

Yinglin Li, Rongxin Cui, Weisheng Yan, and Shi Zhang are with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: liylin6688@mail.nwpu.edu.cn; r.cui@nwpu.edu.cn; wsyan@nwpu.edu.cn; shizhang@mail.nwpu.edu.cn).

Chenguang Yang is with the Department of Computer Science, University of Liverpool, L69 3BX Liverpool, U.K. (e-mail: cyang@ieee.org).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3396100, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3396100

[7], [8]. However, there is rarely a unified framework to concurrently address goal transfer and assistive teleoperation. Unlike the discrete nature of LoA, degree of automation (DoA) refers to a continuous automation concept, enabling finer control. Leveraging a framework for quantifying performance degradation and assessing robot health, researchers can adjust the quantity of LoA or DoA in real-time [15]. Results demonstrated comparable performance between adjusting DoA and adjusting LoA [16]. Nevertheless, this framework primarily prioritize improving robot performance and overlook crucial aspects of human cognition and human-robot interaction (HRI), such as trust. The dynamic switching of DoA to reconcile conflicting intents remains a challenging and underexplored problem in the literature.

The work in [17] introduced a bidirectional trust model, enabling the weighted combination scheme in mobile robot systems that enhances overall performance through gradual adaptation to authority allocation. Researchers have incorporated human trust in robots (briefly called human trust) across various domains, such as shared control [18] and driver action correction [19]. Evidence from variable autonomy systems [20] suggests that human trust improve as operators become more familiar and skilled in task execution, inspiring further exploration of addressing our problem from trust perspective.

The definition of trust, as presented in [21], characterizes it as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability. Current research has made strides in exploring trust factors [22] and establishing trust models [23], [24]. Yet, these efforts mostly focus on unidirectional trust models [25], [26], overlooking bidirectional trust in scenarios with peer-to-peer relationships between autonomous agents and operators. In [27], a capability-based bidirectional multi-task trust model was introduced for predicting human trust or robot trust in humans (briefly called robot trust). Considering variations in factors across domains and environments, adjustments to the input factors of these trust models may be necessary. Therefore, careful consideration of runtime performance impacts related to humans, tasks, and the environment is required for direct application to online DoA adjustment.

To our knowledge, an explicit bidirectional trust model designed to adaptively adjust DoA based on the distinct characteristics of HRI remains an open area of research. This letter aims to bridge this gap by introducing a variable autonomy control system that enables goal transfer and assisted teleoperation, enhancing the capabilities of human-robot teams in reconciling conflicting intents. Unlike human-initiated [5] and mixed-initiative [28] approaches, we achieve this through robot-initiated behavior guided by the bidirectional trust.

The main contributions include (1) formulating a set of five task-specific trust factors for human and robot trust, (2) introducing a bidirectional computational trust model (BTM) that combines diverse trust factors using Kalman filtering techniques, (3) developing a variable weight receding horizon optimization framework to optimize DoA using bidirectional trust as the weighting parameter, and (4) evaluating the performance of the proposed method through a scientifically repeatable semi-autonomous navigation experiment considering conflicting intents on remote mobile robots.

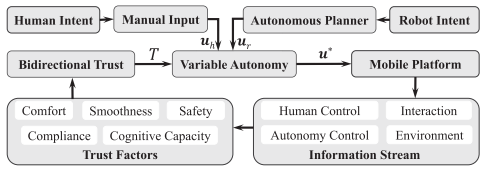


Fig. 2. Variable autonomy framework for bidirectional trust-based DoA modulation.

II. Trust Factors

We consider a semi-autonomous navigation task performed by a human-robot team, as depicted in Fig. 1. The robot is equipped with onboard computers, cameras, IMU, odometry, LiDAR, and other devices, enabling autonomous planning and navigation. The operator, as the teammate, has the ability to manually control the robot using a joystick without force feedback. Visual feedback is provided to the operator through the WiFi network, displaying the camera's first-person perspective.

Fig. 2 depicts the information flow and components of our proposed framework. The robot collects real-time information stream through sensors, extracting key trust factors. These factors form the bidirectional trust, denoted as $T = [T_h, T_r]^\top$. Here, $T_h \in [0, 1]$ represents the human trust, and $T_r \in [0, 1]$ represents the robot trust, with higher values indicating greater trust. When the operator generates manual control commands $\boldsymbol{u}_h \in \mathbb{R}^2$ based on human's intent (goal), and the robot's planner generates autonomous control commands $\boldsymbol{u}_r \in \mathbb{R}^2$ based on robot's intent (goal), the variable autonomy controller utilizes the current bidirectional trust level to determine the shared output $\boldsymbol{u}^* \in \mathbb{R}^2$, which is then applied to the robot.

Trust fluctuations may depend on operator, robot, and environmental characteristics, as well as their interactive closed-loop feedback, rather than solely on any individual factor [21]. These factors, termed trust factors, can be defined and quantified to establish the BTM. While an ideal BTM would consider all potential factors, research [22] shows that a focused set of trust factors can provide robust insights, particularly in specific robot types, tasks, and forms of HRI.

We present a practical and viable example with five trust factors relevant to goal transfer and assisted teleoperation. These trust factors are determined through preliminary experiments, literature analysis, and observations across various robot platforms. To ensure trust consistency, we have analyzed and quantified each factor, assigning numerical values on a scale $[0, 1]$. A value of 1 denotes optimal performance, while 0 indicates the lowest performance.

A. Factors of Robot Trust in Humans

1) Safety: With flawless autonomous obstacle avoidance, our focus is on continuous obstacle detection during operator control to prevent potential harm. Robot trust diminishes when manual operations result in collisions with obstacles. Safety is defined as the likelihood of collision with surrounding objects, ranging

from low to high probability. The robot's velocity observation, denoted as $\boldsymbol{z} = [v_z, \omega_z]^\top$, captures linear velocity v_z and angular velocity ω_z from the robot's sensors. Safety is assessed by monitoring the minimum distance Δd_k to obstacles and the linear velocity observation $v_{z,k}$. Distance is measured using the LiDAR, while the robot's velocity is estimated through Kalman filtering with IMU and odometry measurements. Using an exponential function, the safety factor $\mathcal{S}_k \in [0,1]$ at time k is described as

$$\mathcal{S}_k = e^{-\frac{\gamma_s v_{z,k}^2}{2\Delta d_k}} \tag{1}$$

where $\gamma_s \in \mathbb{R}^+$ denotes the safety coefficient. Rather than using distance [11], we use the deceleration $v_{z,k}^2/(2\Delta d_k)$ as an indicator of safety level.

2) Cognitive Capacity: Extended human interventions may burden operators cognitively, impacting operational performance and diminishing robot trust [22]. Incorporating the utilization in [25], we quantify operator cognitive capability as the function of the duration controlling the robot. Cognitive capacity factor $\mathcal{L}_k \in [0,1]$ at time k is expressed by

$$\mathcal{L}_k = (1-\gamma_w)\,\mathcal{L}_{k-1} + (1-\pi_k)\,\gamma_w \tag{2}$$

where the time constant $\gamma_w \in [0,1]$ determines the sensitivity of the operator's cognitive capacity at the current time to the previous time. The variable $\pi_k \in \{0,1\}$ denotes the robot's control mode at time k, with $\pi_k = 1$ indicating human intervention and $\pi_k = 0$ indicating no human intervention. The longer the human intervention, the lower the cognitive capacity $\mathcal{L}_k$. After humans release control, $\mathcal{L}_k$ increases due to regained mental energy.

3) Smoothness: Experienced operators typically exhibit smoother control over robots, signifying superior operational quality and mitigating abrupt, potentially destabilizing movements [26]. Robots tend to exhibit higher trust in human colleagues displaying enhanced operational quality. To assess the quality of manual control, we use smoothness as a metric, focusing on jerk, which is the first-order time derivative of acceleration $\dot{\boldsymbol{a}}_k$. We consider both linear and angular jerk, as excessive jerk can lead to motion issues such as impacts, jitters, and drift. Acceleration $\boldsymbol{a}_k$ is measured using the IMU, and jerk is calculated by differencing consecutive measurements. Employing an exponential function, the smoothness factor $\mathcal{E}_k \in [0,1]$ at time k is described as

$$\mathcal{E}_k = e^{-\|\gamma_m \dot{\boldsymbol{a}}_k\|_2} \tag{3}$$

where $\gamma_m \in \mathbb{R}_{>0}^{2\times 2}$ represents the smoothness coefficient, and $\|\cdot\|_2$ denotes the Euclidean norm. $\mathcal{E}_k$ increases as $\dot{\boldsymbol{a}}_k$ decreases.

B. Factors of Human Trust in Robots

1) Comfort: When human observations of the robot's performance deviate from their preferences, human trust in the robot consequently diminishes [29]. To ensure optimal visual observation, the robot's movement speed should align with our preferred pace. Excessive speed can result in image blurring, while overly slow movements can test the operator's patience. To quantify the operator's satisfaction with environmental perception, we propose a comfort factor. Using an exponential function, the comfort $\mathcal{I}_k \in [0,1]$ at time k is described as

$$\mathcal{I}_k = e^{-\|\gamma_b(|\boldsymbol{z}_k| - \hat{\boldsymbol{z}})\|_2} \tag{4}$$

where the comfort coefficient $\gamma_b \in \mathbb{R}_{>0}^{2\times 2}$ captures the operator's sensitivity to deviations from the preferred robot velocity denoted as $\hat{\boldsymbol{z}}$. The comfort $\mathcal{I}_k$ increases as the offset $(|\boldsymbol{z}_k| - \hat{\boldsymbol{z}})$ decreases, reflecting the operator's satisfaction.

2) Compliance: The disparity between the robot's actual speed and the operator's expectations often leads to operator frustration, prompting attempts to seize control and thereby diminishing trust in the robot [9]. The compliance factor quantifies the divergence between the robot's motion and the operator's expectations during human intervention. Given the fixed forward-facing onboard camera, we constrain the robot to forward movement or rotation to ensure operational safety. Compliance is primarily observed in the rotational component. We define compliance as the cosine similarity between the robot's angular velocity observation and the operator's angular velocity control input. The operator's control input is denoted as $\boldsymbol{u}_h = [v_h, \omega_h]^\top$, where v_h and ω_h represent the linear and angular velocity components derived from the joystick. Using the sigmoid function, the compliance $\mathcal{C}_k \in [0,1]$ at time k is described as

$$\mathcal{C}_k = \frac{1}{1 + e^{-\gamma_c \cos(\omega_{h,k}, \omega_{z,k})}} \tag{5}$$

where $\gamma_c \in \mathbb{R}_{>0}$ is the compliance coefficient and depends on the human's tolerance for non-compliance. $\mathcal{C}_k$ monotonically increases with lower angular difference $|\omega_{h,k} - \omega_{z,k}|$.

III. Variable Autonomy Control

A. Bidirectional Computational Trust Evolution

Inspired by the qualitative trust framework based on Kalman filter in [23] and the temporal trust model in [25], we have opted to employ a discrete state estimation approach based on the classical Kalman filter for trust estimation. To represent bidirectional trust $T = [T_h, T_r]^\top$ as a state variable, we need to perform mathematical derivations for the state-space model that describes the dynamics of trust. The trust state prediction process can be expressed as

$$T_k = A_k T_{k-1} + B_k + \sigma_k \tag{6}$$

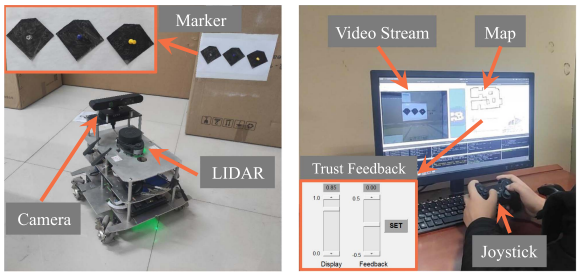
where T_k represents the predicted value of the trust state at the current time, and T_{k-1} is the optimal estimate of the trust state from the previous time step. $\sigma_k \sim \mathcal{N}(0, Q)$ is the system noise following Gaussian distribution. Transfer matrix A_k and estimation vector B_k are designed as

$$A_k = \mathrm{diag}(1-\gamma_h, 1-\gamma_r) \tag{7}$$

$$B_k = [\gamma_h\,(\mathcal{I}_k + \pi_k \mathcal{C}_k)\,, \gamma_h \mathcal{L}_k]^\top \tag{8}$$

where $\gamma_h \in [0,1]$ and $\gamma_r \in [0,1]$ are propagation coefficients. Human trust is predicted through comfort $\mathcal{I}$ and compliance $\mathcal{C}$. Robot trust is predicted through cognitive capacity $\mathcal{L}$. The observation process of the trust state can be expressed as

$$Y_k = H_k + \epsilon_k \tag{9}$$



(a) (b)

Fig. 3. Experimental setup. (a) The mobile robot and markers are positioned in the workspace. (b) The operator controls the robot using a joystick while monitoring it via GUI.

where Y_k represents the trust observation. $\epsilon_k \sim \mathcal{N}(0, R)$ is the observation noise following Gaussian distribution. Observation vector H_k is designed as

$$H_k = \left[T_{hc,k}, \left(\eta\mathcal{E}_k + \eta - 1\right)\mathcal{S}_k\pi_k\right]^\top \tag{10}$$

where safety $\mathcal{S}$ and smoothness $\mathcal{E}$ serve as observations for robot trust calibration. We identify $\mathcal{S}$ as the critical factor and $\mathcal{E}$ as the non-critical factor [26]. $\eta \in [0, 1]$ is the maximum contribution of $\mathcal{S}$. To calibrate human trust, an interactive interface (Fig. 3(b)) is designed to acquire the human trust observation T_{hc}. T_{hc} is not continuously available but is provided by the operator only when deemed necessary for calibration through the interface.

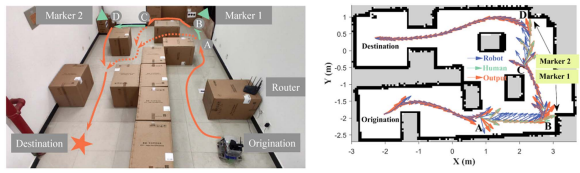
The Kalman filter allows for optimal state estimation by using the predicted value T_k of the system state from the previous time step and the observed value H_k of the current time step.

B. DoA Optimization

Utilizing bidirectional trust T, we devise a variable weight receding horizon optimization framework for variable autonomy. To address conflicting intents, we formulate it as an optimization problem involving a set of DoA. The reference signals encompass manual control input $\boldsymbol{u}_{h,k}$ and autonomous control input $\boldsymbol{u}_{r,k}$, while the potential actions $\boldsymbol{u}_k = [v_k, \omega_k]^\top$ are treated as candidate actions in the optimization. $\boldsymbol{u}_k$ represents a set of candidate control commands, constituting an uncountable set comprised of all control commands constrained within the maximum allowed velocity $\boldsymbol{u}_{max}$ and the minimum allowed velocity $\boldsymbol{u}_{min}$. At time k, the control error is defined as the difference between the candidate action and the reference signal, denoted as $\boldsymbol{e}_k = [\boldsymbol{u}_{h,k} - \boldsymbol{u}_k; \boldsymbol{u}_{r,k} - \boldsymbol{u}_k]^\top$. The cost function is designed as

$$\begin{aligned} J\left(\boldsymbol{u}_{r,k}, \boldsymbol{u}_{h,k}, T_k\right) = &\sum_{i=1}^{N_p} \boldsymbol{e}_{k+i|k}^\top \boldsymbol{\Lambda}_{k+i|k} \boldsymbol{P} \boldsymbol{e}_{k+i|k} \\ &+ \sum_{i=1}^{N_c} \boldsymbol{u}_{k+i|k}^\top \boldsymbol{W} \boldsymbol{u}_{k+i|k} \end{aligned} \tag{11}$$

where N_p is the prediction horizon, and N_c denotes the control horizon. $\boldsymbol{\Lambda} = \mathrm{diag}(T_{h,k}, T_{h,k}, T_{r,k}, T_{r,k})$ is variable weight matrix that penalizes the control error, and $\boldsymbol{P} \in \mathbb{R}^{4\times4}$ regulates trust's contribution to diverse control dimensions. $\boldsymbol{W} \in \mathbb{R}^2$ is the control matrix used to penalize control effort.



(a) (b)

Fig. 4. (a) Human intervention (green line) deviates the robot from the pre-planned path (dashed orange line) to create the new trajectory (solid line). (b) Robot trajectory and angular velocity for participants #5 using BTVA.

Assuming $\boldsymbol{u}_{h,k+i} = \boldsymbol{u}_{h,k}$, and $\boldsymbol{u}_{r,k}$ is generated by the autonomous planner. Accounting for constraints, the BTVA is characterized as

$$\begin{aligned} \boldsymbol{u}_k^* &= \arg\min_{\boldsymbol{u}_k} J\left(\boldsymbol{u}_{r,k}, \boldsymbol{u}_{h,k}, T_k\right) \\ s.t. \quad &\begin{cases} \boldsymbol{u}_{min} \le \boldsymbol{u}_k \le \boldsymbol{u}_{max} \\ v_k \le \sqrt{2a_{max}\Delta d_k} \end{cases} \end{aligned} \tag{12}$$

where the maximum deceleration a_{max} ensures that $\boldsymbol{u}_k$ remains within a feasible range considering obstacle constraints.

In essence, (12) minimizes errors and control efforts resulting from bidirectional trust deficits between the robot and the human. Bidirectional trust serves a dual role in the DoA modulation. The relative magnitudes of human and robot trust indicate the contributions of autonomous and manual control inputs to the optimal output, facilitating control switchovers during goal transfer. Lower human trust values lead to compliance with manual control inputs, and vice versa. Additionally, insufficient human or robot trust, determined by $\boldsymbol{P}$ and $\boldsymbol{W}$, emphasizes the second term in (11), favoring smaller control outputs to hasten trust recovery and ensure robot safety.

IV. Experimental Evaluation

We conducted a within-group experimental study to compare our BTVA with two alternatives in a semi-autonomous navigation task. This study was approval by Northwestern Polytechnical University's Institutional Review Board.

A. Experimental Setup

We created a $6.6\,\mathrm{m}\times3.9\,\mathrm{m}$ maze-like test arena, as depicted in Fig. 4(a), using a mobile robot as the platform. The robot's kinematics and dynamics models were aligned with [11]. The sensors equipped on the robot, as illustrated in Fig. 3(a), were consistent with the description provided in Section II. The remote operator control station shown in Fig. 3(b) included a laptop, joystick, mouse, and a display control interface screen. The operator accessed situational information solely through the screen, which provided real-time video from the robot's front camera and overlaid the estimated robot position on a 2D SLAM map. The SLAM map was generated offline by controlling the robot to cover the entire environment before the experiment, employing LiDAR and SLAM algorithms. The display interface also provides real-time visualization of human trust predictions calculated via (6) and includes a slider ranging from -0.5 to 0.5.

This slider enables operators to input their subjective trust deviation from the current prediction. After capturing human input, the system automatically calculates human trust observations T_{hc}.

We integrated the ROS-based NAV[1] project as our variable autonomy controller on the robot platform, combining the Dijkstra algorithm for global planning and the dynamic window approach for local planning in the autonomous planner.

To solve (12), we employ the L-BFGS-B algorithm from the `scipy.optimize.minimize` function in the Python scipy library. This solver iteratively optimizes a local second-order approximation model of the objective function. At each discrete time step, only the first control signal from the optimal control sequence is implemented on the robot. The sampling frequency is set at 10 Hz, and the robot's velocity is constrained within $\boldsymbol{u}_{\max} = [0.3\,\mathrm{m/s}, 1\,\mathrm{rad/s}]^\top$ for the maximum velocity and $\boldsymbol{u}_{\min} = [0\,\mathrm{m/s}, -1\,\mathrm{rad/s}]^\top$ for the minimum velocity. Additionally, the maximum deceleration is set to $a_{\max} = 0.1\,\mathrm{m/s^2}$. The step sizes for linear and angular velocities are $0.03\,\mathrm{m/s}$ and $0.1\,\mathrm{rad/s}$, respectively. Both N_p and N_c are set to 10. Moreover, we have $\boldsymbol{W} = \mathrm{diag}(0.3, 0.1)$ and $\boldsymbol{P} = \mathrm{diag}(100, 30, 100, 30)$.

Regarding the parameters chosen for the BTM, in Section II, we set $\gamma_s = 2$, $\mathcal{L}_0 = 1$, $\gamma_w = 0.999$, $\gamma_m = \mathrm{diag}(1.1, 0.33)$, $\hat{\boldsymbol{z}} = [0.2\,\mathrm{m/s}, 0.3\,\mathrm{rad/s}]^\top$, $\gamma_b = \mathrm{diag}(1.6, 0.5)$, $\gamma_c = 6$. In Section III-A, we have $\gamma_h = 0.1$, $\gamma_r = 0.1$, $\eta = 0.7$, $Q = \mathrm{diag}(0.11, 0.02)$, $R = \mathrm{diag}(0.16, 0.04)$, $T_0 = [1, 1]^\top$. In our implementation, parameters are empirically determined, and their optimization is beyond the scope of this letter.

B. Tasks

To ensure reproducibility and validity, we used rotatable notched square markers, as shown in Fig. 3(a). These markers were placed in fixed locations within the environment depicted in Fig. 4(a) to mimic randomly occurring targets of interest.

The predetermined robot path is depicted in Fig. 4(a). The robot was tasked with autonomous planning and navigation from origination to destination. Our task aimed for prompt and precise observation of two markers while the robot navigated. Participants oversaw the robot's navigation and controlled it via a joystick upon observing the markers through video feedback, near points A and C. It was mandatory for participants to accurately identify the three notched square directions of the markers in sequence from left to right before releasing control and allowing the robot to continue its initial navigation task.

In our setup, without human intervention (green solid line), the autonomous planned route of the robot (orange dashed line) did not pass near the markers. Specifically, near points A and C, the robot would make a left turn instead of proceeding straight. This configuration aimed to create a mismatch between the robot's goal and human's goal.

C. Control Conditions and Measures

To validate the performance of the proposed algorithm, we further designed two control conditions for comparison.

HISC: In the absence of an existing unified framework, we devised a two-stage method called HISC. Human intervention and release of control marked the goal transfer. Shared control was implemented when $\pi = 1$, combining the operator's command (scaled by weight $\mu \in [0, 1]$) and the autonomous control command (scaled by $1 - \mu$) with $\mu = \mathcal{S}$.

BTSC: For comparison, we developed BTSC, a bidirectional trust-based shared control method. Unlike HISC, BTSC enabled both goal transfer and teleoperation assistance. The weight $\mu = T_r T_h$.

Empirical evidence supports the superiority of LoA switching [5] and assisted teleoperation [17] over manual control. We conclude that HISC, combining human-initiated goal transfer and shared control, outperforms pure manual operation. Thus, manual control is not included as a control scheme.

Five measures were used to compare the control schemes, including three objective measures: operational safety, operational smoothness, and time cost, and two subjective measures: system usability scale (SUS) [30] and NASA task load index (TLX) [31]. Operational safety and operational smoothness are computed by averaging the safety and smoothness factors during human intervention, respectively. Time cost corresponds to the entire task duration from initiation to completion, with TLX and SUS scores assessed following established protocols.

D. Testing Procedure

Twenty participants (17 males and 3 females), aged 23 to 30 years (mean $M = 24.7$, std. $SD = 2.6$), from the authors' academic institution, took part in this experimental study. Participants reported moderate familiarity with the robot ($M = 4.6$, $SD = 1.1$, measured on a 7-point scale).

Upon signing the informed consent form, participants received a 10-minute tutorial explaining the task and controls, followed by a 5-minute practice to familiarize themselves with the remote control platform. Each participant then completed one trial under each control scheme, determined using a Latin square design to minimize order effects. To prevent premature control release, each marker was randomly rotated in multiples of 90° in each trial. We recorded task time and the robot's states. After each trial, participants completed online NASA-TLX and SUS questionnaires to provide subjective evaluations.

V. Results

All participants completed the study without any dropouts, and both objective and subjective measures were recorded and analyzed[2].

A. Verification of Bidirectional Computational Trust

Although no specific experiment was conducted to directly validate the BTM and its trust factors, their implications can be inferred from the recorded state data during the task. Fig. 4(b) illustrates the motion trajectory and angular velocity profile of participant #5 in BTVA. Figs. 5 and 6 depict the evolution of

[1]NAV is available at https://github.com/ros-planning/navigation.

[2]For video results, see https://youtu.be/9jCQuaVTj7M.

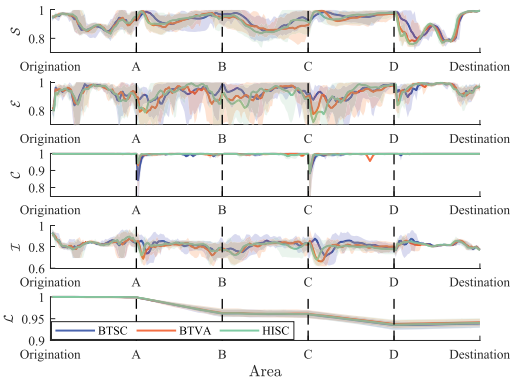


Fig. 5. Evolution of five trust factors among twenty participants across three control conditions. Shaded bands represent 95% confidence intervals.

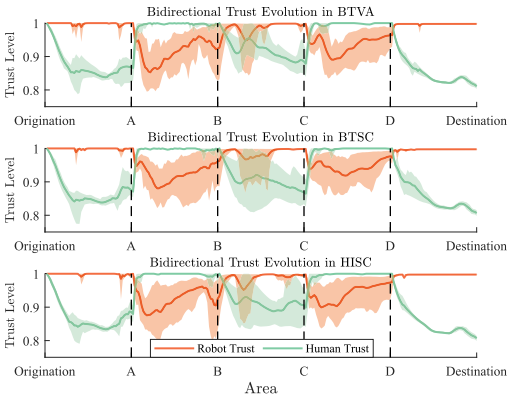


Fig. 6. Evolution of bidirectional trust in three control conditions with twenty participants. Shaded bands represent 95% confidence intervals.

trust factors and bidirectional trust for all participants in three control conditions, with the five areas on the x-axis corresponding to the labels marked in Fig. 4.

Between origination and point A, the robot autonomously navigated with the robot trust level of 1, while human trust decreased below 0.9 due to comfort. At point A, the operator intervened upon spotting the first marker. The transition from robot to human goal was safe but not smooth. Adjusting the robot's orientation reduced comfort, with a compliance drop due to angular velocity. Assuming full control increased speed, further decreasing comfort from linear velocity. Human intervention time rose, leading to reduced cognitive capacity. The robot estimated trust based on control duration, safety, and smoothness, while human trust grew, reflecting the operator's self-confidence. At point B, the operator identified notched square directions and lowered speed for safety during control release, resulting in reduced comfort and smoothness. Robot trust increased with higher cognitive capacity, while human trust decreased due to comfort. The robot autonomously reached point C, guided by the operator to point D upon spotting the second marker, and then proceeded autonomously to complete the task at the destination. Subsequent trust evolution and reasoning aligned with our analysis.

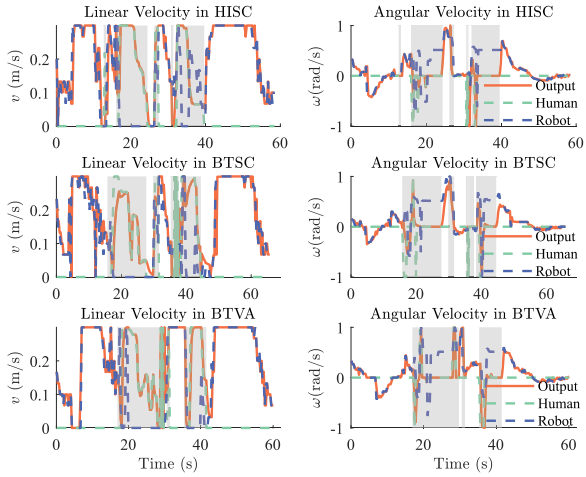


Fig. 7. DoA adjusting evolution for participant #5 using three control conditions. Shaded area represent human intervention.

In summary, these findings highlight the diverse roles of five trust factors in bidirectional trust. Despite potential local numerical variations under different control conditions, a consistent overarching trend emerges. This indicates that our trust factors and model evolve as expected, providing real-time insights into HRI when dealing with conflicting goals, in line with human intuition.

B. Variable Autonomy Verification

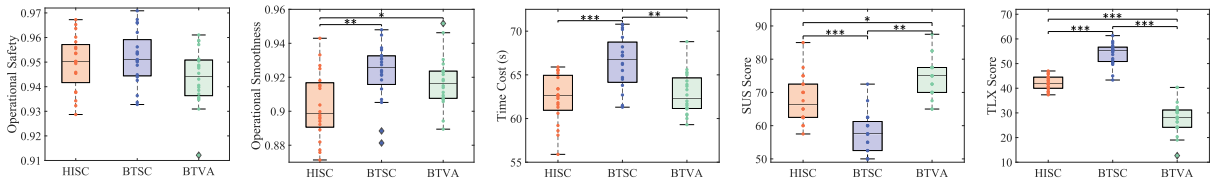
We conducted a one-way repeated measures ANOVA with Greenhouse-Geisser correction for violations of sphericity assumption, and used Bonferroni correction for pairwise comparisons. Statistical significance was set at $p < 0.05$. Results can be seen summarized in Fig. 8.

1) Operational Safety: No significant differences were observed between BTVA ($M = 0.943$, $SD = 0.004$) and BTSC ($M = 0.951$, $SD = 0.002$), and HISC ($M = 0.949$, $SD = 0.002$), $F(2, 38) = 2.364$, $p = 0.105$, $\eta_p^2 = 0.111$, and $power = 0.47$.

2) Operational Smoothness: Significant differences were observed between control schemes, $F(2, 38) = 7.985$, $p = 0.001$, $\eta_p^2 = 0.296$ and $power = 0.94$. Pairwise comparisons revealed that HISC ($M = 0.902, SD = 0.004$) was significantly lower than BTVA ($M = 0.915$, $SD = 0.003$), $p = 0.031$, and BTSC ($M = 0.923, SD = 0.004$), $p = 0.003$, but no significant difference was observed between BTVA and BTSC.

3) Time Cost: There were significant differences between control schemes, $F(2, 38) = 14.753$, $p < 0.001$, $\eta_p^2 = 0.435$ and $power = 0.995$. Pairwise comparisons indicated that BTSC ($M = 67.235$, $SD = 0.8$) took longer time compared with BTVA ($M = 63.335$, $SD = 0.523$), $p = 0.002$, and HISC ($M = 63.125$, $SD = 0.614$), $p < 0.001$. No significant differences were observed between BTVA and HISC.

4) SUS: Significant differences in usability were observed among control schemes, $F(2, 38) = 33.94$, $p < 0.001$, $\eta_p^2 = 0.645$ and $power = 1$, with a gradual decrease from BTVA ($M = 73.625$, $SD = 1.312$) to HISC ($M = 63.75$, $SD =$



(a) Operational Safety (b) Operational Smoothness (c) Time Cost (d) SUS (e) NASA-TLX

Fig. 8. Boxplot comparisons of subjective and objective measurements with a sample size of twenty participants using three control conditions. Key: $^{*}p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$.

1.57), $p = 0.057$, and BTSC ($M = 56.875$, $SD = 1.383$), $p < 0.001$.

5) TLX: Significant differences were observed between control schemes, $F(2, 38) = 128.091$, $p < 0.001$, $\eta_p^2 = 0.883$ and $power = 1$. Pairwise comparisons revealed a gradual increase in workload from BTVA ($M = 28.567$, $SD = 1.483$) to HISC ($M = 42.867$, $SD = 0.698$), $p < 0.001$, and BTSC ($M = 54.267$, $SD = 1.209$), $p < 0.001$.

VI. Discussion

A. Overview of Bidirectional Computational Trust

Analysis of our experimental data shows that the BTM intuitively aligns with trust reactions between individuals and exhibits a common trend in trust evolution, confirming the rationale. The successful use of the BTM as a two-dimensional modulation factor for adaptive DoA adjustment further validates our model.

In developing our remote operation system using the BTM, we offer two insights. Firstly, by considering trust factors in HRI and modeling them with the Kalman filter, our model distinguishes itself from black-box neural network trust models [24]. Interpretability is crucial, particularly in robotics and industrial AI applications. Secondly, the state space's mathematical structure allows for the incorporation of new trust factors, requiring only considerations of their intrinsic attributes, source, and consequential impact. This facilitates the integration of trust factors into existing models.

We also discuss ways to overcome certain limitations. Firstly, the BTM falls into the temporal model category, akin to those in [17], [19]. These models rely on parameters, which might not perfectly match the operator's real trust, despite our calibration using human trust feedback. Future work should involve dedicated psychological or psychophysical experiments for real-time human trust data collection and parameter estimation. Secondly, our selection of trust factors is based on literature reviews and expert judgment without quantitative justifications. Subsequent experiments could leverage machine learning algorithms and statistical analysis to automatically derive trust factors and identify optimal combinations instead of manual settings.

B. Extensions of Variable Autonomy

The presented BTVA complements the findings in [26] and [16]. The work in [26] aligns with our method of developing computational trust models using real-time sensor measurements, which are objective and devoid of human biases. The problem we study can be considered as an extension of the problem investigated in [16]. In addition to addressing DoA adjustment caused by human performance degradation, we also tackle DoA adaptive adjustment due to goal transfer.

Our experimental findings demonstrate that during robot autonomous navigation, the absence of direct HRI results in robots being uninformed about a significant portion of human-related information. Consequently, human trust typically falls below that placed in robot trust. Guided by our DoA modulation principle, robots maintain their autonomy. However, human intervention, driven by humans' inherent confidence in their actions, consistently elevates human trust, while robots continuously assess human performance in real-time, adjusting their trust accordingly to comply with human control. This interplay between human and robot trust fluctuates as goals transfer, facilitating a smooth transition from autonomy, aimed at fulfilling robot intent, to compliance, aimed at fulfilling human intent, as illustrated in Fig. 7. This underscores the efficacy of our proposed bidirectional trust framework, integrating trust factors as inputs, in providing valuable guidance for robots to achieve task goals at any given moment. While our example highlights semi-autonomous navigation, the proposed framework can be adapted to various tasks involving conflicting intents, such as environmental exploration, with only modifications to the BTM.

Based on quantitative statistical analysis of our experiments, all three control conditions demonstrated satisfactory safety performance, with no collision incidents observed in the recorded videos. This meets the safety requirements for practical application in physical robots. Compared with BTSC, BTVA reduced time cost by 5.8%, increased SUS by 16.75%, and reduced TLX by 25.7%. Compared with HISC, BTVA increased operational smoothness by 1.5%, improved SUS by 9.9%, and reduced TLX by 14.3%. Additionally, it was observed that compared with BTSC, HISC improved SUS by 6.9% and reduced TLX by 11.4%, which aligns with human intuition. Humans prioritize making the robot quickly obey their control rather than focusing solely on smooth trajectories.

The results of the within-group experimental study confirm the effectiveness of our variable autonomy approach, marking the first step for the BTM to modulate the DoA during goal transfer and assisted teleoperation. In the current work, while the BTM effectively reconciles conflicting intents by modulating DoA, conflicts persist, as shown in Fig. 4(b). Swift recognition and adaptation to human intents by the robot after human takeover may hold the key to conflict resolution. Future work can aim to achieve intent alignment by inferring and explicitly

measuring human intents [8] using environmental cues and operator control data.

VII. Conclusion

In this letter, we addressed conflicting intents in human-robot teams by developing a variable autonomy control method based on bidirectional trust. To describe trust, we designed five task-specific trust factors. These factors were combined using Kalman filtering to create the BTM. By using the BTM as the weighting parameter, we developed a variable weight receding horizon optimization method for DoA adjusting. Experimental studies with 20 participants demonstrated real-time detection and quantification of bidirectional trust during task execution. Our method facilitated smooth transitions in control authority, ensuring safety in teleoperation. Statistical analysis confirmed our approach exhibits an overall higher system usability and lower operator workload.

Acknowledgment

The authors would like to thank the volunteers for their participation in the experiments.

References

[1] M. Moniruzzaman, A. Rassau, D. Chai, and S. M. S. Islam, "Teleoperation methods and enhancement techniques for mobile robots: A comprehensive survey," *Robot. Auton. Syst.*, vol. 150, 2022, Art. no. 103973.

[2] C. Yang, Y. Zhu, and Y. Chen, "A review of human–machine cooperation in the robotics domain," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 1, pp. 12–25, Feb. 2022.

[3] S. A. Mostafa, M. S. Ahmad, and A. Mustapha, "Adjustable autonomy: A systematic literature review," *Artif. Intell. Rev.*, vol. 51, pp. 149–186, 2019.

[4] D. A. Abbink et al., "A topology of shared control systems-finding common ground in diversity," *IEEE Trans. Hum.-Mach. Syst.*, vol. 48, no. 5, pp. 509–525, Oct. 2018.

[5] M. Chiou, R. Stolkin, G. Bieksaite, N. Hawes, K. L. Shapiro, and T. S. Harrison, "Experimental analysis of a variable autonomy framework for controlling a remotely operating mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 3581–3588.

[6] M. Chiou, N. Hawes, and R. Stolkin, "Mixed-initiative variable autonomy for remotely operated mobile robots," *ACM Trans. Hum.-Robot Interact.*, vol. 10, no. 4, pp. 1–34, 2021.

[7] S. Rothfuß, M. Chiou, J. Inga, S. Hohmann, and R. Stolkin, "A negotiation-theoretic framework for control authority transfer in mixed-initiative robotic systems," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2022, pp. 921–928.

[8] D. Panagopoulos et al., "A hierarchical variable autonomy mixed-initiative framework for human-robot teaming in mobile robotics," in *Proc. IEEE 3rd Int. Conf. Hum.- Mach. Syst.*, 2022, pp. 1–6.

[9] C. X. Miller, T. Gebrekristos, M. Young, E. Montague, and B. Argall, "An analysis of human-robot information streams to inform dynamic autonomy allocation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 1872–1878.

[10] P. Pappas, M. Chiou, G.-T. Epsimos, G. Nikolaou, and R. Stolkin, "VFH+ based shared control for remotely operated mobile robots," in *Proc. IEEE Int. Symp. Saf., Secur., Rescue Robot.*, 2020, pp. 366–373.

[11] J. Luo, Z. Lin, Y. Li, and C. Yang, "A teleoperation framework for mobile robots based on shared control," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 377–384, Apr. 2020.

[12] A. W. d. Jonge, J. G. Wildenbeest, H. Boessenkool, and D. A. Abbink, "The effect of trial-by-trial adaptation on conflicts in haptic shared control for free-air teleoperation tasks," *IEEE Trans. Haptics*, vol. 9, no. 1, pp. 111–120, Jan.–Mar. 2016.

[13] J. Ludwig, A. Haas, M. Flad, and S. Hohmann, "A comparison of concepts for control transitions from automation to human," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2018, pp. 3201–3206.

[14] A. Kucukyilmaz, T. M. Sezgin, and C. Basdogan, "Intention recognition for dynamic role exchange in haptic collaboration," *IEEE Trans. Haptics*, vol. 6, no. 1, pp. 58–68, First Quarter 2013.

[15] A. Ramesh, R. Stolkin, and M. Chiou, "Robot vitals and robot health: Towards systematically quantifying runtime performance degradation in robots under adverse conditions," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10729–10736, Oct. 2022.

[16] C. A. Braun, A. Ramesh, S. Rothfuß, M. Chiou, R. Stolkin, and S. Hohmann, "Model predictive control of the degree of automation optimizing robot health," in *Proc. IEEE 17th Int. Symp. Appl. Comput. Intell. Inform.*, 2023, pp. 381–386.

[17] H. Saeidi, J. R. Wagner, and Y. Wang, "A mixed-initiative haptic teleoperation strategy for mobile robotic systems based on bidirectional computational trust analysis," *IEEE Trans. Robot.*, vol. 33, no. 6, pp. 1500–1507, Dec. 2017.

[18] A. Broad, J. Schultz, M. Derry, T. Murphey, and B. Argall, "Trust adaptation leads to lower control effort in shared control of crane automation," *IEEE Robot. Automat. Lett.*, vol. 2, no. 1, pp. 239–246, Jan. 2017.

[19] H. Azevedo-Sa, S. K. Jayaraman, X. J. Yang, L. P. Robert, and D. M. Tilbury, "Context-adaptive management of drivers trust in automated vehicles," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 6908–6915, Oct. 2020.

[20] M. Chiou, F. McCabe, M. Grigoriou, and R. Stolkin, "Trust, shared understanding and locus of control in mixed-initiative robotic systems," in *Proc. IEEE 30th Int. Conf. Robot Hum. Interactive Commun.*, 2021, pp. 684–691.

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

[22] K. E. Schaefer, J. Y. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems," *Hum. Factors*, vol. 58, no. 3, pp. 377–400, 2016.

[23] T. B. Sheridan, "Extending three existing models to analysis of trust in automation: Signal detection, statistical parameter estimation, and model-based control," *Hum. Factors*, vol. 61, no. 7, pp. 1162–1170, 2019.

[24] C. Nam, P. Walker, H. Li, M. Lewis, and K. Sycara, "Models of trust in human control of swarms with varied levels of autonomy," *IEEE Trans. Human-Mach. Syst.*, vol. 50, no. 3, pp. 194–204, Jun. 2020.

[25] H. Saeidi and Y. Wang, "Incorporating trust and self-confidence analysis in the guidance and control of (semi) autonomous mobile robotic systems," *IEEE Robot. Automat. Lett.*, vol. 4, no. 2, pp. 239–246, Apr. 2019.

[26] Q. Wang, D. Liu, M. G. Carmichael, S. Aldini, and C.-T. Lin, "Computational model of robot trust in human co-worker for physical human-robot collaboration," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 3146–3153, Apr. 2022.

[27] H. Azevedo-Sa, X. J. Yang, L. P. Robert, and D. M. Tilbury, "A unified bi-directional model for natural and artificial trust in human–robot collaboration," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 5913–5920, Jul. 2021.

[28] S. Jiang and R. C. Arkin, "Mixed-initiative human-robot interaction: Definition, taxonomy, and survey," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2015, pp. 954–961.

[29] F. M. Verberne, J. Ham, and C. J. Midden, "Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars," *Hum. Factors*, vol. 54, no. 5, pp. 799–810, 2012.

[30] J. Brooke, "SUS-A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.

[31] D. Sharek, "A useable, online NASA-TLX tool," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2011, pp. 1375–1379.