Artificial Intelligence and Human-Computer Interaction
Y. Ye and H. Zhou (Eds.)

doi:10.3233/FAIA250157

Trust-Modulated Authority Allocation in Human-Guided Goal Recognition Tasks

Ruiyu XIA[a], Yunbo ZHAO[a,b,c], Junsen LU[a], Yang WANG[a], Pengfei LI[a] and Yu KANG[a,1]

[a] *Department of Automation, University of Science and Technology of China, Hefei, China*

[b] *Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China*

[c] *Institute of Advanced Technology, University of Science and Technology of China, Hefei, China*

ORCiD ID: Ruiyu XIA https://orcid.org/0009-0002-2084-6022
Yu KANG https://orcid.org/0000-0002-8706-3252

Abstract. In shared control teleoperation, the machine infers the humans' goal to provide effective assistance, which we call human-guided goal recognition. However, current methods mainly use algorithm confidence to assign control authority during the process, which makes it difficult to correct machine inference errors under high confidence. To address this problem, we propose a trust model that considers machine capability fluctuations and human-machine interaction experience. We also add trust as a dynamic assessment of machine capabilities to authority allocation to improve the success rate of the tasks. Finally, we verify the effectiveness of the proposed method through experiments.

Keywords. Trust model, human-guided goal recognition, human-machine authority allocation

1. Introduction

In shared control teleoperation, the machine typically infers human goals based on their control input and provide autonomous assistance towards the predicted goal [1–4]. For example, in Brain-Computer-interface (BCI) controlled manipulation, the system helps humans accomplish goal grasping by inferring their desired goals [5]. For the convenience of subsequent description, we refer to this process as human-guided goal recognition. This process not only reduces human workload, but also improves the performance of human-machine collaboration.

Establishing appropriate human-machine authority allocation is a prerequisite for effective machine assistance. The majority of studies use intention inference success rate or algorithm confidence as the basis for determining machine control authority during the process [3–5]. However, this form presents a challenge for humans in correcting machine behavior when it makes a high-confidence error prediction. Therefore, we need a more appropriate method to divide human-machine control authority.

[1] Corresponding Author: Yu KANG, E-mail: kangduyu@ustc.edu.cn.

Trust, which is commonly acknowledged as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [6], might be a potential solution. It represents the human assessment of the actual machine capabilities and the expected extent of machine assistance. Therefore, considering trust in authority allocation can better evaluate the machine capabilities, allowing humans to correct machine errors.

One possible reason why existing studies have not considered trust is that trust models suitable for this scenario have not yet been proposed. The machine capabilities in these tasks are generally given at the end of the task and are mostly evaluated in a binary way. Existing machine capability-based trust models fluctuate widely in this scenario, making accurate trust tracking difficult [7,8]. Hu demonstrated in experiments that previous trust and human-machine interaction experience affect current trust [9]. However, the proposed model is difficult to meet the requirement of continuously updating trust based on machine performance in real-time control.

Therefore, to the best of our knowledge, this paper is the first to propose a time-series trust model that considers machine capability fluctuations and human-machine interaction experience in this scenario and considers trust in the allocation of control authority to improve system performance. The trust-modulated dynamic authority allocation is introduced in Section 2; the trust definition and computation model for this context are described in Section 3; the experimental platform design and results are outlined in Section 4; and the conclusion is presented in Section 5.

2. Trust-modulated dynamic authority allocation

In the human-guided goal recognition tasks, human-machine hybrid control can be expressed as Eq. (1).

$$u_s = u_h \cdot (1-\eta) + u_m \cdot \eta \tag{1}$$

Here, u_h represents the human control input, u_m represents the machine control algorithm input, and u_s is the final control output sent to the machine. The factor η determines the division of control authority between the human user and the machine. To ensure that humans can correct machine inference errors, we design factor η as a function related to the inference algorithm confidence ρ and human-machine trust T.

$$\eta = \begin{cases} 0 & if\ \rho < \zeta_1 \\ \dfrac{T(\rho-\zeta_1)}{(\zeta_2-\zeta_1)} & if\ \zeta_1 \le \rho \le \zeta_2 \\ T & if\ \rho > \zeta_2 \end{cases} \tag{2}$$

As shown in Eq. (2), confidence ρ measures the algorithm certainty. When ρ exceeds the lower threshold ζ_1, the machine control authority increases proportionally with ρ until it exceeds the upper threshold ζ_2, at which point it provides maximum assistance as defined by the dynamic trust level T.

Remark 1: When both T and ρ are high, humans may still find it difficult to correct machine errors. Switching control to human based on the detection of long-term conflicts between human and machine control could serve as a final safeguard.

3. Human-Machine Dynamics Trust model

Definition 1 (Human-Machine Dynamics Trust). In human-guided goal recognition tasks, Human-Machine Dynamic Trust is defined as the human subjective perception of the machine capability. As seen in Eq. (3), it is influenced by the trust value T from the previous moment, the machine capability fluctuation ΔP, and the human-machine interaction experience E.

$$\begin{aligned} T(k) &= \Gamma\big(T(k-1), \Delta P(k), E(k)\big), k \geq 0 \\ T(t) &= T(k), t \in [t_k, t_{k+1}), k \geq 0 \end{aligned} \tag{3}$$

Considering trust remains stable in the absence of new information or experience [10], we model the Human-Machine Dynamic Trust as a piecewise constant function varying over time, where t_k denotes the time humans update their trust after receiving the specific event for k-th time. And $t_0 \coloneqq 0$ represents the start time of the human-machine system. Previous studies have shown a strong correlation between robot performance and human trust [9,11,12]. Therefore, considering that both long-term reliability and short-term performance affect trust, we incorporated machine capability fluctuation ΔP and human-machine interaction experience E into the model.

We define the machine capability fluctuation $\Delta P(k)$ as the difference in machine capability within the interval $[k-1, k)$:

$$\Delta P(k) = P(k) - P(k-1) \tag{4}$$

In the human-guided goal recognition task, the machine intent inference capability $P(k)$ is determined by the Euclidean distance between inferred and actual goals denoted as δ_k. If δ_k exceeds the threshold ε, the intent inference fails:

$$P(k) = \begin{cases} 1 & if\ \delta_k < \varepsilon \\ 0 & if\ \delta_k > \varepsilon \end{cases} \tag{5}$$

Based on the human cognitive characteristic of being more sensitive to recent interaction experiences [11], we introduce a dynamic weighting strategy with a forgetting mechanism in the design of human-machine interaction experience $E(k)$ to reflect that only the most recent n fluctuations in machine capability will affect trust. The weight allocation follows an exponential decay pattern, ensuring that the most recent data receives the highest weight. $E(k)$ is calculated as follows:

$$E(k) = \begin{cases} \dfrac{\sum_{i=\max(0,k-n+1)}^{k} \lambda^{k-i} P(i)}{\sum_{i=\max(0,k-n+1)}^{k} \lambda^{k-i}}, k > n \\ \dfrac{\sum_{i=0}^{k} \lambda^{k-i} P(i)}{\sum_{i=0}^{k} \lambda^{k-i}} \quad\quad , k \leq n \end{cases} \tag{6}$$

$P(i)$ represents the machine capability at the i-th interaction; $\lambda^i = 0.95^i (0 < \lambda < 1)$ is the decay factor, determining the contribution of $P(i)$ to $E(k)$ at time i;the denominator ensures that the sum of the weights is 1, maintaining the normalization property of $E(k)$.

Based on previous research, we propose that when humans evaluate trust, different cognitive factors carry varying weights, and humans are more sensitive to declines in machine capability [9,12]. And E represents the long-term machine reliability, allowing us to assume its weight is stable. Therefore, Eq. (3) in Definition 1 can be refined in Definition 2, which is an original contribution to address trust modeling.

Definition 2 (Human-Machine Dynamics Trust Model).

$$\begin{cases} T(k) = \alpha^* T(k-1) + \beta^* \Delta P(k) + \gamma E(k), k > 0 \\ T(t) = T_m^h(k), t \in [t_k, t_{k+1}), k > 0 \end{cases} \tag{7}$$

where

$$\alpha^* + \beta^* + \gamma = 1;$$
$$(\alpha^*, \beta^*, \gamma) = \begin{cases} (\alpha^+, \beta^+, \gamma),\ \Delta P(k) > 0; \\ (\alpha^0, 0, \gamma), \quad \Delta P(k) = 0; \\ (\alpha^-, \beta^-, \gamma),\ \Delta P(k) < 0. \end{cases} \tag{8}$$

Noted that $T(0)$ represents the initial human trust, obtained by the scale, while $P(0) = P_0$ represents the machine objective capability, measured by the average accuracy of the intent inference algorithm in human-machine experiments. Since the trust value $T \in [0,1]$, all parameters (α^*, β^* and γ) are set within the range (0, 1) to ensure model stability, with their sum equal to 1.

Several studies have shown that trust model parameters may be influenced by factors such as nationality, cognitive background, age, and so forth [9,11,12]. Therefore, we construct a personalized trust model for everyone, using the least squares method to solve for the parameters in the model [13].

4. Experiment and result

4.1. Setup

We construct our simulation environment using the Lunar Lander simulator developed by OpenAI Gym[2] (Figure 1), which provides a highly controllable and reproducible setting suitable for AI-assisted drones in search-and-rescue missions. The experimental interaction flow is shown in Figure 2. To align with our task setup, we replaced the goal point with randomly generated coordinates g, while keeping other settings unchanged.

The machine autonomous control strategy is driven by the DDQN algorithm [14], which achieves a 95% success landing rate after training with known goals. For intent inference, we employ the Naive Bayes network proposed by Jain [1]. The parameters in Eqs. (2), (5), and (6) are set as: $\zeta_1 = 40\%, \zeta_2 = 80\%, n = 6, \varepsilon = 0.167$.

[2] Documentation for the Lunar Lander simulator can be found on the Gymnasium official website: https://gymnasium.farama.org/environments/box2d/lunar_lander/

Participants: two participants, aged 24, are recruited through selection. They use keyboard arrow keys for interaction and complete 20 practice trials to familiarize themselves with the interface before the formal experiment

We set up two experiments to validate the method: Experiment 1 uses the factor η^* with the machine objective capability P_0 as its auxiliary upper limit for 100 trials, to obtain the trust model parameters. While Experiment 2 uses the trust-modulated factor η for 100 trials to verify the validity of the trust model and the effectiveness of trust modulation in authority allocation. After each trial, a pop-up window collects changes in participant trust. The system automatically collects the accuracy of intent inference and the task success rate.

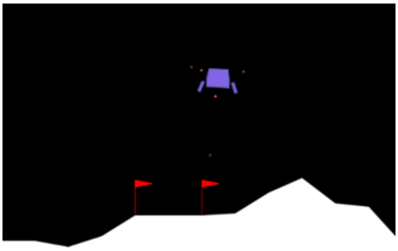


Figure 1. Simulated environment.

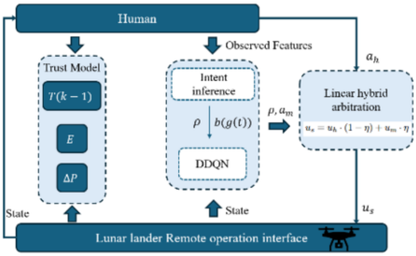


Figure 2. Simulation flow chart.

4.2. Result

By analyzing the experimental data from Experiment 1, the trust model parameter values in Eqs. (7) and (8) are summarized in Table 1. From the parameter values, H2 is more sensitive to ΔP and E, while H1 tends to be more cautious. This is demonstrated in Experiment 2. A detailed analysis is provided below.

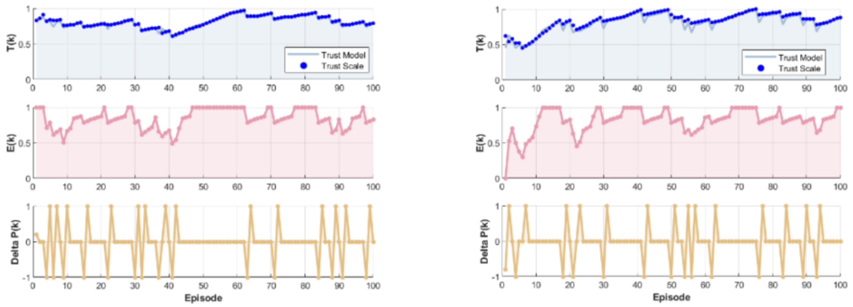
Table 1. Trust parameter changes under different conditions.

Human Subjects	$\boldsymbol{\alpha^+}$	$\boldsymbol{\beta^+}$	$\boldsymbol{\alpha^0}$	$\boldsymbol{\alpha^-}$	$\boldsymbol{\beta^-}$	$\boldsymbol{\gamma}$
H1	0.8637	0.0247	0.8884	0.8480	0.0404	0.1116
H2	0.7346	0.0646	0.7992	0.7280	0.0712	0.2008

4.2.1. Trust model validity

The trust curve of the participants in Experiment 2 are shown in Figure 3. H2 exhibits greater trust variations when encountering continuous capability fluctuations, while H1 is more stable. Moreover, H2 trust values increase faster than H1, further supporting the above view. Additionally, we find that when trust levels are high, a sudden drop in machine performance leads to a greater loss of trust. This may be because humans pay less attention to the control process at high trust levels, making it difficult to correct mistakes promptly, which leads to blaming the machine.

The small difference between the trust model value and the trust scale value verifies the validity of the model. In addition, Figure 3 shows that the average trust value of the participants fluctuates around 0.8, which indicates that after long-term interaction, human-machine trust value is roughly equal to the actual capability.



(a) Participant H1

(b) Participant H2

Figure 3. Trust dynamically changed based on machine capability fluctuations and interaction experiences.

4.2.2. Effectiveness of the trust-modulated dynamic authority allocation.

To better demonstrate the effect of adding trust to authority allocation, we divide trust into 10 intervals by Eq. (9), and the success rate f_n of each trust intervals T_n can be defined as the ratio of successful trials S_n to the total number of trials N_n within that interval.

$$T_n = [0.1 \times (n-1), 0.1 \times n) \quad \text{where } n \in \{1,2,\dots,10\}. \tag{9}$$

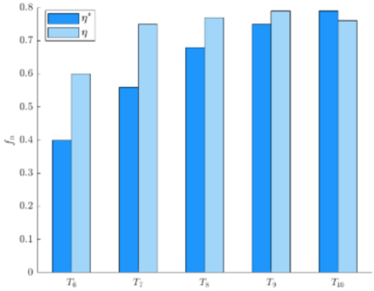


Figure 4. Success rate under different authority allocation designs.

As shown in Figure 4, after adding trust to the authority allocation, the success rate has improved in most cases, and the lower the trust, the greater the improvement in success rate. Since lower trust means that the machine inference capability is worse at that moment, the trust-modulated dynamic authority allocation gives human more authority to correct incorrect inferences. However, we also find that this model makes it difficult for humans to correct when trust is high. One possible solution is to enforce human control when a prolonged human-machine control conflict is detected.

5. Conclusion

This paper presents a human-machine trust model integrated with authority allocation as a dynamic assessment of machine capabilities, with its effectiveness demonstrated through experiments. Future work will focus on conducting experiments in real-life scenarios or using virtual reality (VR) technology to create more realistic teleoperation environments. Additionally, the model can be extended by broadening the definition of machine capabilities to include critical factors beyond reasoning abilities, such as control capabilities, safety, and acceptance of recommendations. This will enhance its applicability across various fields. For example, in AI-assisted art creation, using trust as a criterion for evaluating the quality of AI-generated paintings will undoubtedly guide algorithms toward better aligning with human needs.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Nos. 62173317 and 62473352) and the Zhuimeng Fund of Jianghuai Advanced Technology Innovation Center (No. Q/JH-063-2023-T02).

References

[1] Jain S, Argall B. Recursive Bayesian Human Intent Recognition in Shared-Control Robotics. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE; 2018. p. 3905–3912, doi: 10.1109/IROS.2018.8593766.

[2] Haninger K, Hegeler C, Peternel L. Model Predictive Control with Gaussian Processes for Flexible Multi-Modal Physical Human Robot Interaction. 2022 International Conference on Robotics and Automation (ICRA). Philadelphia, PA, USA: IEEE; 2022. p. 6948–6955, doi: 10.1109/ICRA46639.2022.9811590.

[3] Reddy S, Dragan A, Levine S. Shared Autonomy via Deep Reinforcement Learning. Robotics: Science and Systems XIV. Robotics: Science and Systems Foundation; 2018, doi: 10.15607/RSS.2018.XIV.005.

[4] Song P, Li P, Aertbeliën E, et al. Robot Trajectron: Trajectory Prediction-based Shared Control for Robot Manipulation. 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE; 2024. p. 5585–5591, doi: 10.1109/ICRA57147.2024.10611507.

[5] Muelling K, Venkatraman A, Valois J-S, et al. Autonomy infused teleoperation with application to brain computer interface controlled manipulation. Auton Robot. 2017;41(6):1401–1422, doi: 10.1007/s10514-017-9622-4.

[6] Hoff KA, Bashir M. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Hum Factors. 2015;57(3):407–434, doi: 10.1177/0018720814547570.

[7] Li Y, Cui R, Yan W, et al. Reconciling Conflicting Intents: Bidirectional Trust-Based Variable Autonomy for Mobile Robots. IEEE Robot Autom Lett. 2024;9(6):5615–5622, doi: 10.1109/LRA.2024.3396100.

[8] Wang Q, Liu D, Carmichael MG, et al. Robot Trust and Self-Confidence Based Role Arbitration Metho d for Physical Human-Robot Collaboration. 2023 IEEE International Conference on Robotics and Auto mation (ICRA). London, United Kingdom: IEEE; 2023. p. 9896–9902, doi: 10.1109/ICRA48891.2023.10160711.

[9] Hu W-L, Akash K, Reid T, et al. Computational Modeling of the Dynamics of Human Trust During Human–Machine Interactions. IEEE Trans Human-Mach Syst. 2019;49(6):485–497, doi: 10.1109/THMS.2018.2874188.

[10] Ekman F, Johansson M, Sochor J. Creating Appropriate Trust in Automated Vehicle Systems: A Framework for HMI Design. IEEE Trans Human-Mach Syst. 2018;48(1):95–101, doi: 10.1109/THMS.2017.2776209.

[11] Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Human factors. 2004;46(1):50–80.

[12] Schaefer KE, Chen JYC, Szalma JL, et al. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. Hum Factors. 2016;58(3):377–400, doi: 10.1177/0018720816634228.

[13] Björck Å. Least squares methods. Handbook of Numerical Analysis. Elsevier; 1990. p. 465–652, doi: 1 0.1016/S1570-8659(05)80036-5.

[14] Van Hasselt H, Guez A, Silver D. Deep Reinforcement Learning with Double Q-Learning. AAAI. 2016;30(1), doi: 10.1609/aaai.v30i1.10295.