Effective Human-in-the-loop Control Handover via Confidence-Aware Autonomy

Breanne Crockett*
Univ. of Colorado Boulder
Boulder, Colorado, USA
breanne.crockett@colorado.edu

Kyler Ruvane*
Univ. of Colorado Boulder
Boulder, Colorado, USA
kyler.ruvane@colorado.edu

Matthew B. Luebbers
Univ. of Colorado Boulder
Boulder, Colorado, USA
matthew.luebbers@colorado.edu

Bradley Hayes
Univ. of Colorado Boulder
Boulder, Colorado, USA
bradley.hayes@colorado.edu

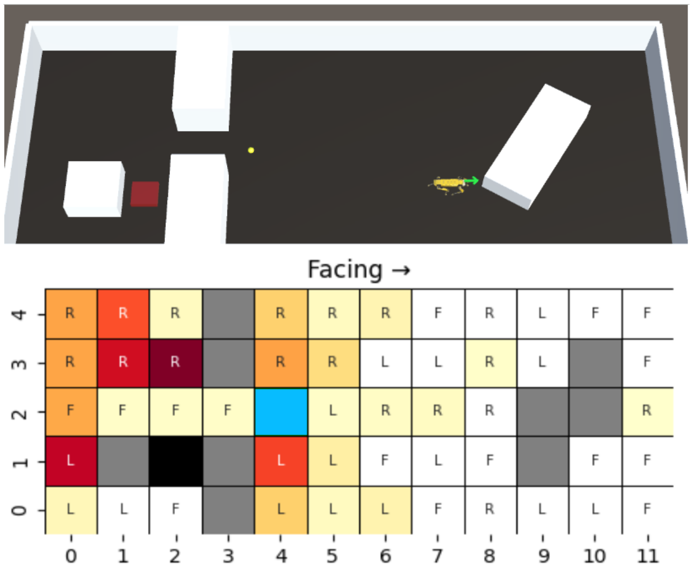
***Abstract*—Robotics is traditionally divided into two operational paradigms – autonomous control and teleoperation. Both approaches are affected by the inherent strengths and weaknesses of autonomous systems and human operators. Therefore, it is beneficial for many tasks to blend the two operation strategies, incorporating human-in-the-loop supervision with autonomous control. In this work, we explore the question of control handover: when should a robot act autonomously, when should a human supervisor take control, and who should decide this? We first analyze four candidate metrics for estimating confidence in a policy learned via reinforcement learning (count of examples, choice difficulty, Gaussian choice difficulty, and historical upper confidence bound), using those metrics to autonomously trigger handover requests to a human supervisor whenever a robot's confidence is low. Through a simulation evaluation, we found historical upper confidence bound to be the most correct metric, achieving the highest accuracy on the timing of handover requests. Using this finding, we conducted a human-subjects evaluation, showing that in a human-supervised robotic navigation task, robot-to-human handover triggered autonomously using our method outperformed human-initiated handover, both on robotic task performance and on subjective human measures of workload and usability.**

***Index Terms*—Human-in-the-loop, Control Handover, Agent Confidence, Reinforcement Learning**

I. Introduction

Robotic agents are capable of being operated either by an autonomous controller or by a human via teleoperation. Both paradigms have advantages and drawbacks. Teleoperation is often challenging and time consuming for human operators, especially for high degree-of-freedom robotic platforms [1]. It is also hard to scale, with an individual human generally unable to control more than one robot at a time. In contrast, autonomous control requires no direct human involvement, enhancing scalability but often adding substantial risk. In complex environments, designers often will not trust a robot to make correct decisions in dangerous states unforeseen by its training, preferring human control in such cases. Human-in-the-loop systems broadly aim to combine both autonomous control and teleoperation schema in an intelligent way to balance the respective strengths and weaknesses of each approach [2].

Ideally, a robot should operate autonomously when it will perform well on its own, and defer to a human operator by handing over control when it will perform poorly. Actually achieving this behavior, however, remains an open research question, and forms the focus of this work. We are interested in answering both whether human-initiated or robot-initiated handover of control leads to better combined human-robot performance, as well as what triggers for robot-initiated handover perform best. If a human decides to take over control of a robot, it is usually because that human has low confidence in the robot's ability to take optimal actions. We extend this idea to the robot-initiated case, examining a number of candidate agent self-confidence measures from reinforcement learning (RL) literature for their suitability as triggers for human-in-the-loop control handover, giving control to a human operator when the robot lacks confidence and operating autonomously otherwise. This method provides the ancillary benefit of obtaining human expert training data precisely in the spots where



	0	1	2	3	4	5	6	7	8	9	10	11
4	R	R	R		R	R	R	F	R	L	F	F
3	R	R	R		R	R	L	L	R	L		F
2	F	F	F	F		L	R	R	R			R
1	L				L	L	F	L	F		F	F
0	L	L	F		L	L	L	F	R	L	L	F

Fig. 1. Top: An autonomous robot attempts to travel to the yellow tennis ball (blue square in bottom image), avoiding the red trap square (black square in bottom image). Bottom: A visual representation of the robot's policy at every position when it is rotated towards the east of the environment - the dog can go forward ”F” one square, turn right ”R” 45 degrees, or turn left ”L” 45 degrees. Also shown is the learned confidence in each state - states colored white have very low confidence while states colored dark red have very high confidence.

[*] These authors contributed equally to this work.

the robot is least confident, allowing the robot to respond by adjusting its value function moving forward through inverse reinforcement learning [3], requiring fewer handovers in the future.

We perform two evaluations, forming the primary contributions of this work: 1) we implement and then analyze the correctness of four candidate handover triggers derived from agent confidence measures using a simulated human-robot navigation task, and 2) we perform a human-subjects study using the same task to test the effectiveness of our automated, robot-initiated handover method against human-initiated handover on a number of objective and subjective metrics.

II. Related Work

Prior research has been conducted into developing agent self-confidence measures for RL algorithms, often with the aim of improving agent training. Confidence in this context is often defined as the sureness about an estimated value for taking a given action in a given state. Many works achieve this by calculating confidence bounds on the value at each state-action pair, such as Mannor et al. [4], who present a method for determining the confidence bounds of value function estimates in discrete Markov processes using statistical variance. White and White [5] describe a similar method, computing the confidence bounds of value function estimates using a history of those estimates and bootstrap sampling. Confidence can also be modeled as by the difference between the estimated values of the highest valued and second highest valued actions from a state, otherwise known as the choice difficulty [6]. All of these measures are designed for use during agent training, providing means to track exploration and exploitation [7] and as an indicator of when to stop training. Our work, on the other hand, explores using these measures to mediate handover in human-in-the-loop robotic systems.

Additional techniques constantly maintain a distribution of possible value function estimates for each state-action pair in order to train risk-aware agents. Though computationally expensive, these distributions carry more information than a simple expected value, enabling alternate and complex decision criteria. For example, an agent could elect to choose an action with the lowest probability of a negative reward in certain situations. Bellemare et al. [8] argue the value of such distribution-estimating techniques for improving risk-aware behavior in RL agents. Morimura et al. [9] introduce an approach for producing non-parametric approximations of these value-function distributions, also demonstrating improvement in high-risk RL environments. Cao et al. [10] employ a distributional technique for confidence-aware RL in a self-driving car domain, selecting an RL-based or a rule-based policy depending on their respective confidence levels. This represents a similar motivation to our work – instead of switching control between two policies, we are interested in switching between human and robotic control, using confidence measures to do so.

III. Confidence Measure Simulation Evaluation

Through simulation, we evaluated four RL-based agent confidence metrics on their feasibility as an autonomous trigger for robot-to-human control handover.

A. Simulation Environment

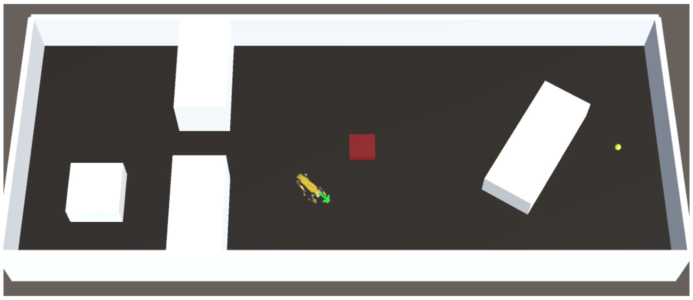


Fig. 2. Simulated navigation domain used for evaluation: a quadruped robot tries to find a tennis ball while avoiding the red trap state.

We chose a simulated discrete navigation task as our evaluation domain. The autonomous agent (a quadruped robot) must traverse a maze to locate and reach its end goal, a tennis ball, while avoiding a trap square which will render the robot unable to move. The robot's state is characterized by its location on the 5x12 maze grid, and its heading. Since the robot can face in one of eight directions, the environment contains a total of 480 states (5 rows x 12 columns x 8 orientations per cell). The robot can take one of three deterministic actions: rotating to the left or to the right by 45 degrees, or moving forward one square (diagonal moves are allowed). Attempting to move forward into a wall results in the robot staying at its current position.

The robot recovers a large positive or negative reward upon reaching a terminal state of an episode (+100 for reaching the tennis ball and -100 for landing in the trap state). Upon reaching any other state, the robot recovers a reward of -1. This reward structure incentivizes the robot to locate and reach the goal in as few actions as possible.

B. Robot Policy

To evaluate the various candidate handover triggers, we trained policies for the robot across 200 environments with randomized goal and trap placement through tabular q-learning [11]. For each environment, to emulate the experience of human supervision in complex domains where perfect policies cannot reasonably be obtained, we deliberately stopped the agent's training prior to a fully accurate policy being reached. If the robot were always capable of taking the correct action given its state, it would gain no utility from human oversight. For the purposes of our analysis, we define an action within a policy to be "accurate" if following the policy rollout from that action leads to the goal state within 30 actions. Each policy was trained for 9,000 episodes, leading to an average accuracy of 91.8%.

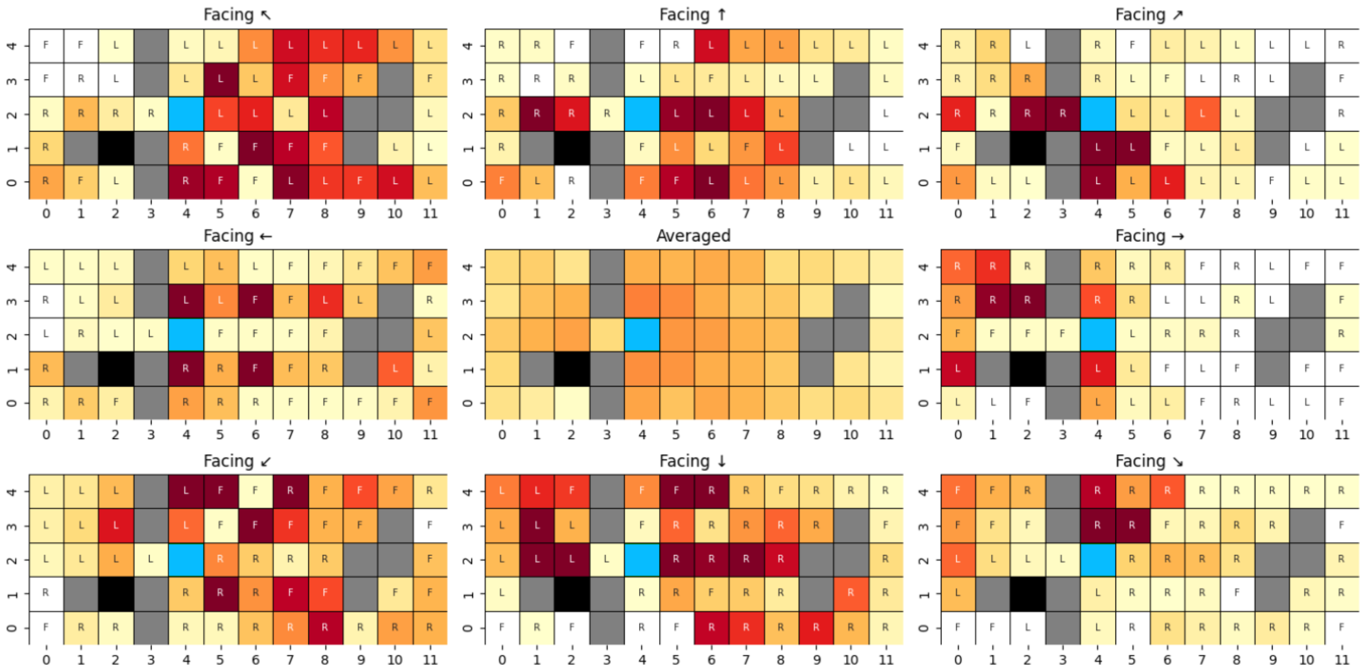


Fig. 3. The learned confidence values for every state of one simulated environment using Historical Upper Confidence Bound. Each image represents the policy for states at the given robot orientation, showing the chosen action with a letter ('F'orward, 'L'eft, or 'R'ight). Low confidence states are white, while high confidence states are dark red. Confidence-based control handover can be used to shield against poor or partially learned policies.

C. Evaluating Confidence Measures

We analyzed the correctness of four candidate action confidence measures for deciding when to handover control to a human supervisor. To obtain a direct comparison of measure quality, we set the threshold equivalently between conditions so that the 15% of states with the least confidence according to their respective confidence measures would trigger a handover. In practice, since certain states are traversed more commonly than others, the robot took approximately 96.8% of total actions autonomously, electing to handover to a human supervisor 3.2% of the time. Certain methods require extensions to the standard q-learning training loop and are described below. The four methods evaluated were (1) *Count of Examples*, (2) *Choice Difficulty*, (3) *Gaussian Choice Difficulty*, and (4) *Historical Upper Confidence Bound*.

(1) Count of Examples

This metric involves simply keeping track of the number of times each state was visited during training of the reinforcement learner agent. Intuitively, agents should be more confident in estimating the value of actions from a given state if it has visited that state and tried those actions many times. To support this metric, an additional table `q_count` is maintained throughout the training process. When a state is visited during the course of an episode, the count for that state is incremented by one. A visited state can be counted at most once during an episode.

(2) Choice Difficulty

This metric employs a choice difficulty heuristic to estimate confidence, similar to the technique shown in [6]. For each state, the difference between the highest-valued action and the second highest-valued action is computed, with a larger difference indicating higher confidence. The intuition is that human-provided insight is most useful whenever the autonomous agent does not possess an obvious best action. This method does not require any additional details to be tracked during training; the q-table is sufficient.

(3) Gaussian Choice Difficulty

This metric is an extension of *Choice Difficulty*, modelling each q-value as Gaussian distributions parametrized by a mean and standard deviation, and using the distribution to calculate a likely 90th percentile value for each q-value. These upper-bounded values are then used in the same manner as the *Choice Difficulty* approach, with higher difference between the first and second highest-valued actions leading to higher confidence. To support this, an additional table `q_dist` is used to track the mean and variance of each encountered state, according to Welford's online algorithm for online mean and variance estimation [12].

(4) Historical Upper Confidence Bound

This metric calculates an upper confidence bound for the agent's q-function using the approach described by White and White [5], adapted for a discrete environment. To achieve this, an additional table `q_hist` is maintained, tracking the N most recently learned q-function values for each state, in our case the most recent 20 values. Those values are smoothed by uniformly sampling groups of 3 values with replacement and averaging them, placing the resultant averages in a new list.

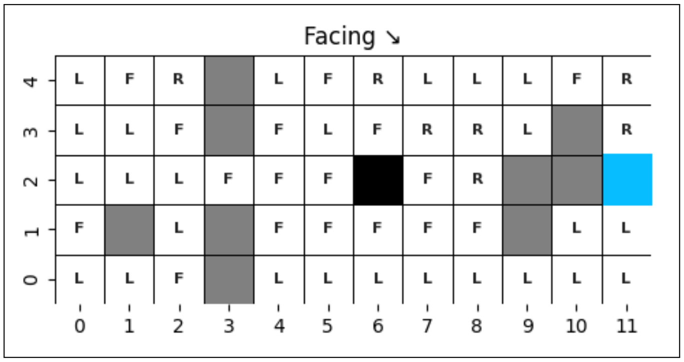
To determine the upper confidence bound, that list of samples is sorted largest to smallest, with the 90th percentile value treated as the upper confidence bound. The confidence score for this metric is the absolute difference between the current estimated q-value and the upper confidence bound for that q-value, with smaller values indicating higher confidence.

D. Handover Trigger Correctness Measures

Using the same notion of accuracy defined previously (an action is accurate if it leads to a goal state within 30 turns), we define a set of three correctness measures to compare the suitability of the four proposed confidence metrics as handover triggers. The first measure is the percentage of accurate actions regarded by the robot as confident (analogous to the true positive rate). A robot deciding not to handover when its planned actions are already suitable is a desirable outcome, as it avoids unnecessary interruption for the human supervisor.

Relatedly, the second measure is the percentage of inaccurate actions regarded by the robot as unconfident (analogous to the concept of true negative rate). A robot deciding it is unconfident and handing over to a human supervisor when it would have ended up taking suboptimal actions on its own is a similarly desirable outcome that will improve robot performance.

Blending these measures, we use the F_1 score as the third measure, defined as the harmonic mean of the trigger's **precision** (the number of accurate actions regarded as confident divided by the total number of actions regarded as confident) and its **recall** (the number of accurate actions regarded as confident divided by the total number of accurate actions) $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$. This measure, also on a scale of 0 to 1, provides a single number describing the correctness of the handover trigger.



	0	1	2	3	4	5	6	7	8	9	10	11
4	L	F	R		L	F	R	L	L	L	F	R
3	L	L	F		F	L	F	R	R	L		R
2	L	L	L	F	F	F		F	R			
1	F		L		F	F	F	F	F		L	L
0	L	L	F		L	L	L	L	L	L	L	L

Fig. 4. A visual representation of a robot policy when rotation is fixed to the south-east direction (**F**: go forward, **L**: turn left, **R**: turn right). The black square is the trap, the blue is the goal, and grey are obstacles. Note that the state at x=2, y=0 has a sub-optimal learned policy: the robot is facing a corner, and moving forward would simply run into the wall. The robot should ideally recognize this state as unconfident and request a handover to a human supervisor.

E. Results

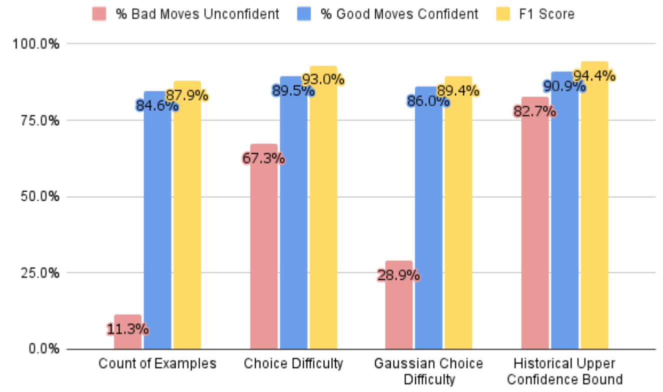


Fig. 5. The evaluation of the four confidence metrics averaged over 200 learned policies on randomly configured environments. Higher values are better in each case.

Through training policies and confidence measures over 200 environment setups, we obtained cumulative results on the correctness of each confidence measure-based handover trigger, as shown in Fig. 5.

On all measures, the best performing confidence metric was *Historical Upper Confidence Bound*, with 82.7% of incorrect actions leading to handover trigger (far higher than the next leading candidate at 67.3%), 90.9% of correct actions avoiding a handover trigger, and an F_1 score of 94.4%.

Choice Difficulty also performed well, ranking second on all measures, but struggled whenever facing a choice between two similarly valued actions which both led to optimal solutions, underrating that state's confidence. *Gaussian Choice Difficulty* performed notably worse, with only 28.9% of sub-optimal moves triggering handover, a poor true negative rate. *Count of Examples* performed worst of the four metrics, largely due to the high concentration of most visited states around the agent's starting state, which led to those states always being rated as confident, and far away states unconfident, in spite of relative the policy optimality in those areas.

Although further optimization could be achieved by a more principled choice of confidence threshold for each method which reduces false positive and false negative rates, our simulation evaluation serves to highlight the relative suitability of *Historical Upper Confidence Bound* as an automated trigger for robot-to-human handover by reinforcement learning agents. For this reason, that is the method we implemented for use in our human-subjects trial.

IV. Human-Subjects Evaluation

We designed and conducted a small IRB approved between-subjects human-subjects study (n = 7) to determine the relative effectiveness of human-initiated handover vs. robot-initiated handover using the *Historical Upper Confidence Bound* method. In the study, a human supervised the same robot simulation task described in Section III-A, while attending to their own task simultaneously.

A. Hypotheses

H1: Robot-initiated handover will lead to higher robot task performance compared to human-initiated handover.

H2: Robot-initiated handover will lead to higher human task performance compared to human-initiated handover.

H3: Robot-initiated handover will be rated as more usable compared to human-initiated handover.

H4: Human-initiated handover will be rated as requiring more workload compared to robot-initiated handover.

B. Experimental Setup

Participants played the role of a human supervisor for the simulation environment described above. On one monitor, participants see a game window showing the robot and its environment (Fig. 2). On another, participants see a picture card memory-matching game. The goals of each experimental exercise are to maximize the number of matches found by the human participant in the memory-matching game, and to maximize the number of tennis balls found by the autonomous robot within a set time window. Most of the time, the robot operates autonomously, following its policy to locate the tennis ball, and respawning in a random location to head towards the tennis ball again once it is reached. This leaves the human available to attempt to find and clear matching cards from the screen with their mouse.

The method of control handover from robot to human forms the difference between the two experimental conditions. In **human-initated handover**, the human decides to take control of the robot by hitting the spacebar and entering actions manually using keyboard control, until they are satisfied the robot will operate successfully from that point onward, relinquishing control to the robot by hitting the spacebar again. Presumably, the human will decide to take control when they witness the robot behaving sub-optimally (for instance, if they are stuck on an obstacle).

In **robot-initiated handover**, the robot will decide when the human should take control, triggering a handover when it reaches one of the 15% lowest confidence states according to the *Historical Upper Confidence Bound* method. To alert the human, the robot sounds an audible handover alarm (a series of sharp beeps). The robot then waits for the human to input an action via the keyboard, returning to autonomous control if the new state has sufficient confidence, or sounding the handover alarm again if not, requesting another human-provided action.

C. Protocol

Though participants see both conditions in the experiment, their ordering is randomized and counterbalanced. To begin, participants are given descriptions of their memory-matching and robot-supervision tasks as well as opportunities to play the memory-matching game and enter actions for the robot to reduce possible learning effects. After onboarding, participants play the first round for 2 minutes, with the number of successful matches in the memory-matching game and the number of tennis balls found by the robot recorded. After the round concludes, participants are administered two brief surveys: the NASA TLX [13] to determine subjective workload and the SUS [14] to determine the subjective usability of the condition they just saw.

Following this, participants play a second experimental round for 2 minutes, using the opposite handover control method from the first round. The performance is recorded, and the TLX and SUS are administered again. Lastly, a final survey is administered, asking demographic and comparative questions and soliciting open-ended feedback. Specifically, participants are asked to decide both which round they felt required the most mental effort, as well as which round they found the robot to behave the most intelligently.

D. Results

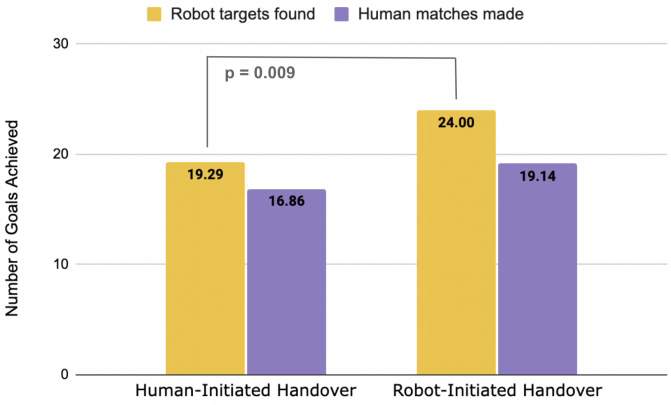


Fig. 6. Objective performance: number of robot targets found and human matches made by condition, with means and significance shown.

For robotic performance, we measured the average number of tennis ball targets found across all 2 minute rounds. Using a one-tailed t test, we measure whether participant intervention in the human-initiated or robot-initiated conditions led to better robot performance. Running the test, we found that the robot in the robot-initiated handover condition $(M = 24.00)$ found significantly more tennis balls than the human-initiated handover condition $(M = 19.29)$; $(t(13) = 2.80, p = 0.009)$. This result serves to **support H1**.

For human performance, we measured the average number of matches made in the memory-matching game across all 2 minute rounds. Again using a one-tailed t test, we compare the matches made across the two conditions. No significance was found between human performance in the robot-initiated handover condition $(M = 19.14)$ and in the human-initiated handover condition $(M = 16.86)$, though the mean is higher in the robot-initiated case. More data is required to demonstrate the effect more definitively. Due to the lack of significance, **H2 is inconclusive**.

Anecdotally, individual participant strategy appeared to play a large role in how human and robotic performance were affected in the human-supervised condition. Participants who frequently checked the other monitor to supervise the robot lost performance on the memory-matching task due to distraction, whereas participants who focused mainly on their own memory-matching task ignored the robot for longer when it required help, degrading robot performance. This represents a key tradeoff for attention which is somewhat remedied by the autonomous, robot-initiated handover paradigm.

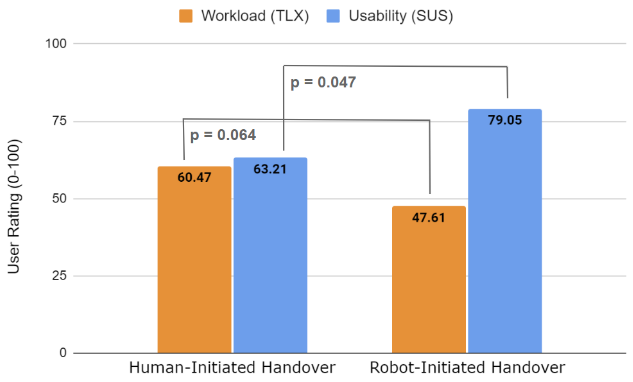


Fig. 7. Rated usability and workload by condition (0-100), with means and significance shown. It is preferable to have a low rating of workload and a high rating of usability.

Individual survey responses for the NASA TLX and SUS scales were coded and added together to form uniform scales from 0-100. Lower TLX scores are regarded as preferable (lower perceived workload), whereas higher SUS scores are preferable (greater usability). Via a one tailed t-test, we compared the two conditions on both subjective measures.

No significant effect was found on TLX score between the robot-initiated $(M = 47.61)$ and human-initiated conditions $(M = 60.47); (t(13) = -1.64, p = 0.064)$, though the initial results suggest that an effect indicating lower workload for robot-initiated handover may be revealed through further data collection. Significant differences were found on SUS score however, with the robot-initiated handover condition $(M = 79.05)$ rated as more usable than the human-initiated handover condition $(M = 63.21); (t(13) = 1.81, p = 0.047)$. This result serves to **support H3**.

Although the TLX measure alone does not conclusively address the hypothesis of lower workload in the robot-initiated condition, 7 out of 7 participants indicated in the post-experimental survey that the human-initiated condition required the higher mental effort of the two rounds, a significantly greater proportion than the expected random proportion of 50%, $p = 0.008$. This result does serve to **support H4**. All 7 participants also chose the robot in the robot-initiated round as being the most intelligent. These results combine to showcase the benefits of automated control handover using the *Historical Upper Confidence Bound* method.

V. Conclusion

Through our simulated evaluation of four RL confidence measures, we found that determining confidence by tracking the difference between the upper confidence bound and the current best estimate of the value for a given state-action pair (the *Historical Upper Confidence Bound* method) led to the highest correctness as a handover trigger. It led to the highest fraction of suboptimal actions triggering a handover to the human, as well as the lowest fraction of optimal actions interrupting the human unnecessarily.

We compared this method of handover against human-initiated handover in a human-subjects study. In the study, participants supervised a robot navigation task while conducting their own task simultaneously. We found that robot-initiated handover, based on the *Historical Upper Confidence Bound* measure, led to higher robot task performance, as well as higher subjective ratings of usability and lower workload. These results serve to highlight the potential of this confidence based handover technique for human-in-the-loop supervision.

Future work will focus on refining this technique, developing more principled ways to set the confidence threshold used to trigger handover and evaluating the method in larger, more complex domains. Additionally, future work will aim to close the training loop by using the human-provided actions following robot-to-human handovers as training data to improve the value estimates in the lowest confidence states through inverse reinforcement learning. This will aid in life-long learning, improving robot performance and automated confidence determination steadily as the human interacts with the robot, leading to a smaller and smaller fraction of states requiring handover over time.

References

[1] J. Y. C. Chen, E. C. Haas, and M. J. Barnes, “Human performance issues and user interface design for teleoperated robots,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*.

[2] D. Honeycutt, M. Nourani, and E. Ragan, “Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy,” in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, 2020, pp. 63–72.

[3] P. Abbeel and A. Y. Ng, “Apprenticeship learning via inverse reinforcement learning,” in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 1.

[4] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis, “Bias and variance in value function estimation,” in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 72.

[5] M. White and A. White, “Interval estimation for reinforcement-learning algorithms in continuous-state domains,” *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[6] M. Lebreton, K. Bacily, S. Palminteri, and J. B. Engelmann, “Contextual influence on confidence judgments in human reinforcement learning,” *PLoS computational biology*, vol. 15, no. 4, p. e1006973, 2019.

[7] M. Coggan, “Exploration and exploitation in reinforcement learning,” *Research supervised by Prof. Doina Precup, CRA-W DMP Project at McGill University*, 2004.

[8] M. G. Bellemare, W. Dabney, and R. Munos, “A distributional perspective on reinforcement learning,” in *International conference on machine learning*. PMLR, 2017, pp. 449–458.

[9] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka, “Nonparametric return distribution approximation for reinforcement learning,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 799–806.

[10] Z. Cao, S. Xu, H. Peng, D. Yang, and R. Zidek, “Confidence-aware reinforcement learning for self-driving cars,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7419–7430, 2021.

[11] C. J. C. H. Watkins, “Learning from delayed rewards,” 1989.

[12] B. Welford, “Note on a method for calculating corrected sums of squares and products,” *Technometrics*, vol. 4, no. 3, pp. 419–420, 1962.

[13] S. G. Hart and L. E. Staveland, “Development of nasa-tlx (task load index): Results of empirical and theoretical research,” in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[14] J. Brooke *et al.*, “Sus-a quick and dirty usability scale,” *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.