Maintaining Efficient Collaboration with Trust-Seeking Robots

Anqi Xu[1] and Gregory Dudek[1]

***Abstract*— In this work, we grant robot agents the capacity to sense and react to their human supervisor's changing trust state, as a means to maintain the efficiency of their collaboration. We propose the novel formulation of Trust-Aware Conservative Control (TACtiC), in which the agent alters its behaviors momentarily whenever the human loses trust. This trust-seeking robot framework builds upon an online trust inference engine and also incorporates an interactive behavior adaptation technique. We present end-to-end instantiations of trust-seeking robots for distinct task domains of aerial terrain coverage and interactive autonomous driving. Empirical assessments comprise a large-scale controlled interaction study and its extension into field evaluations with an autonomous car. These assessments substantiate the efficiency gains that trust-seeking agents bring to asymmetric human-robot teams.**

I. INTRODUCTION

Trust – one's belief in another's competence and reliability – is the cornerstone of all long-lasting collaborations, both within the human workplace as well as for partnerships between humans and robots. The degree of trust that a human supervisor has in an autonomous robot agent is strongly correlated with the team's performance, and also impacts the quality of their interactions [1]–[3]. Supervisor-worker teams that foster high levels of trust often demonstrate great synergy, where the human's high-level decision skills complement the robot agent's exhaustive planning and control capabilities. In contrast, low trust can cause teams to break down, where the human hesitates to delegate tasks to the robot agent or even disables it altogether [4].

The goal of our research is to improve the efficiency of collaborations between mobile robots and their human supervisors. Efficiency is a multi-faceted construct and is captured as a combination of objective metrics such as performance and workload, as well as subjective assessments including perceived collaborative efforts and trustworthiness.

To maintain strong team efficiency, we propose the framework of trust-seeking robots, namely autonomous robot agents that can infer and react to the human's evolving trust states. A key framework component comprises the novel formulation of Trust-Aware Conservative Control (TACtiC), in which the agent momentarily behaves in a conservative manner following salient trust losses. Conservative behavior alterations are designed both to elicit the human's assistance and to limit adverse effects of various causes of trust loss, such as erroneous motions or noisy task conditions. The trust-seeking robot framework further incorporates an online *performance-centric* trust modeling module [5] and an interactive adaptation technique [6] for learning from occasional human control inputs. Together, these components enable robot agents to actively strive to prevent breakdowns in teamwork and improve their performance over time, thus maintaining an efficient collaboration.

[1]The authors are with the School of Computer Science, McGill University, Montreal, Canada. {anqixu,dudek}@cim.mcgill.ca

Fig. 1. A human supervising a trust-seeking smart car as it drives along a road. Inset: camera view overlaid with tracked boundary, robot's steering command (blue arrow), and human's intervening command (green arrow).

Our contribution of the trust-seeking robot framework is the first-ever realization of robot agents that *react to a human's changing trust states explicitly*. We present two end-to-end implementations of trust-seeking robots for distinct task contexts of collaborative aerial coverage and interactive autonomous driving (see Fig. 1). Our developments entail instantiating the novel concept of Trust-Aware Conservative Control, as well as integration of an online trust inference engine and an interactive behavior adaptation method together into a sophisticated robot agent. We assess the efficiency gains of these trust-seeking agents both onboard aerial drones during a large-scale controlled user study and onboard a self-driving car during real-world interactions.

II. BACKGROUND AND RELATED WORK

This section elaborates on the foundations of the trust-seeking robot framework, including the supervisor-worker style asymmetric team structure, a robot control agent for visual navigation tasks with interactive adaptation abilities, and an online engine for inferring the human's trust state. We also highlight related studies on conservative robot behaviors.

A. Supervisor-Worker Team

Our research focuses on an asymmetric human-robot team structure in which the autonomous agent ("worker") onboard a mobile robot is chiefly responsible for handling a given task while the human takes on the role of the "supervisor". The supervisor can intervene at any time by overriding the agent's commands and assuming control over the robot vehicle. Ideally, interventions should occur only when necessary, such

as when correcting the agent's mistakes or switching to a new task target. The human can also make positive and negative critiques of the robot agent's task performance, as well as provide feedback on their current trust state.

To effectively collaborate with the supervisor, the autonomous agent should make use of the human's various interaction signals, including occasional intervening commands, periodic relative-scale critiques, and infrequent absolute-scale trust feedback. The trust-seeking robot framework incorporates all of these factors into a unified estimate of the human's trust state at each moment during interactions. The agent then uses this trust signal to induce temporary behavior alterations, towards coping with trust losses and preventing team breakdown. Concurrently, the agent also learns from the human's intermittent commands and adapts its own motions to establish a common task intent.

Our robot agents adapt to human critiques indirectly in a generalized trust-centric manner, which contrasts with approaches that learn directly from human-generated rewards [7], [8]. Also, the behavior adaptation technique used in this work closely relates to methods for learning from human demonstrations [9], [10], as further elaborated in [6].

B. Collaborative Visual Navigation

We instantiate the trust-seeking robot framework within a visual navigation task domain, in which the human supervises the mobile robot as it tracks various terrain boundaries. This work builds upon a generalized autonomous system used to steer aerial drones along highways and shorelines as well as to drive cars along roads and trails. Visual navigation tasks are appealing since humans innately excel at them whereas the complexity in autonomous solutions (e.g. [2], [6]) warrants the need for trust.

At the core of our autonomous robot agent lies a vision-based boundary tracking algorithm [6]. This system segments out and tracks a salient terrain boundary in the robot's camera frames using various traits such as hue or brightness. It also discriminates edge boundaries such as forest contours from road-like strip boundaries. Image-space features of the boundary are mapped into a linear control law to produce steering commands, while operating at a fixed speed.

The diverse parameters of this vision-based controller can be tuned to track a wide variety of terrain boundaries. We automate the parameter tuning process using the Adaptive Parameter EXploration (APEX) algorithm [6], as illustrated in Fig. 2. APEX is an anytime optimization method that incrementally adapts system parameters so as to ensure that its control outputs resemble the human's intermittent steering commands. This work integrates the original APEX implementation along with optimized hyper-parameter settings, notably using learning rates of $\alpha = 0.2$ and $\gamma = 0.7$.

C. Human-Robot Trust

Our research is inspired by the multi-disciplinary literature on trust, observing its vital role within human relationships. The concept of trust is subject to a multitude of interpretations in different contexts such as within a society, an organization, or a mutual collaboration [11]. For the

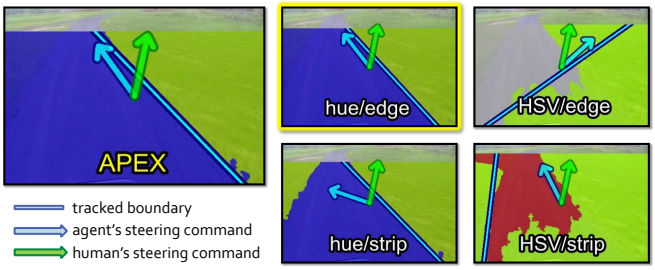

Fig. 2. The Adaptive Parameter EXploration (APEX) algorithm [6] learns to track a road by maintaining and optimizing hypotheses of continuous (e.g. horizon cutoff, control gains) and discrete (e.g. image feature, boundary type) parameters (right). On each frame, the hypothesis most consistent with recent human commands is integrated into the agent (left).

specific context of a single-human single-robot team, trust encompasses two major elements:

- the *degree of trust*: a quantifiable subjective assessment towards another individual;
- the *act of trust*: the decision and behavior of relying upon another individual's abilities or services.

Our trust-seeking robot framework integrates prior research on quantifying the degree of trust. This work also "closes the loop" by encouraging the human to adopt the act of trust with the use of behavior adaptation and alteration strategies.

An important aspect in quantifying trust is the diverse set of factors upon which the degree of trust can be based [1]. Factors applicable to a human-robot team broadly fall under two categories: those based on the robot agent's task performance and competence, as well as those related to its intentions and integrity. As is common in human-robot trust research (e.g. [2], [12]), *this work adheres to a performance-centric trust definition*, and we explicitly inform users that our robot agents are well-motivated, obedient, and non-deceptive. The performance-centric assumption is naturally suited to supervisor-worker teams, since the robot agent does not have personal motives for deception as it collaborates with the human towards a unified task goal.

1) Temporal Trust Modeling: Various representations have been proposed to quantify the degree of trust in a robot or a piece of automation. These include binary [13] and continuous [12] measures that causally attribute the robot's trustworthiness based on its performance, as well as ordinal scales [14], [15] used to elicit a person's actual trust state.

Many studies have described human-robot trust through correlations with interaction experiences and attitudes, although few models can predict the trust state over time. Lee and Moray [12] presented a model that relates trust assessments to performance factors within a human-automation context. Our prior work [3] similarly predicted changes in trust by relating to factors such as the robot agent's task failures and the frequency of human interventions. Desai and Yanco [2] proposed the Area Under Trust Curve measure, which accumulates an operator's positive and negative critique signals towards an autonomous robot.

Most recently, we developed the Online Probabilistic Trust Inference Model (OPTIMo) [5], which combines existing approaches of performance-based causal attribution

and evidential grounding to interaction factors and subjective trust assessments. OPTIMo boasts superior trust prediction accuracies, and its ability to infer trust updates every few seconds enables trust-seeking agents to react promptly. We next present an overview of this real-time trust model.

2) Online Probabilistic Trust Inference Model (OPTIMo): As shown in Fig. 3, OPTIMo is represented by a Dynamic Bayesian Network, which discretizes continuous-time interactions into W-second time blocks. This graphical model maintains a probabilistic belief over the trust state $t_k \in [0, 1]$ at each time step k. The trust belief $bel(t_k)$ is propagated as a linear function of the robot's current and recent task performance estimates $p_k, p_{k-1} \in [0, 1]$. For our boundary tracking agent, performance is quantified as the ratio of frames in each time block with successful boundary detections. This trust estimate is then calibrated to match the human's latest intervention state $i_k \in \{0, 1\}$. The likelihood of manual control is modeled against factors such as the predisposition to micromanage, effects of low trust and loss of trust, as well as extraneous motives $e_k \in \{0, 1\}$ when training the agent to follow a new boundary target. Trust beliefs are further grounded by the human supervisor's critiques $c_k \in \{+1, 0, -1\}$ and by infrequent trust feedback $f_k \in [0, 1]$.

Our implementation of the OPTIMo network uses a histogram-based inference engine, in which beliefs over the continuous trust space $[0, 1]$ are represented as probability masses for B equally-sized bins. Separate interaction datasets are collected to train personalized OPTIMo instances for each user in order to capture their unique trust tendencies. This procedure employs Expectation-Maximization with random restarts, and is cross-validated to minimize prediction error on trust feedback data f_k. Our OPTIMo implementation uses optimized hyper-parameter settings from prior work, notably with $W = 3$ seconds and $B = 300$ bins.

D. Conservative Robot Behaviors

Our formulation of conservative control alterations shares commonalities with several studies on hesitant behaviors of social robots. Breazeal *et al.* [16] showed that non-verbal cues for humanoid robots, such as shrugging to convey hesitation, improved a human collaborator's mental model and resulted in superior teamwork and robustness. Kazuaki *et al.* [17] found that adding appropriate robot motion delays to convey hesitation accelerated the robot agent's learning process and improved its perceived teachability. Moon *et al.* [18] formalized hesitation gestures for planning trajectories of a robot manipulator. These motions were favorably perceived as less dominant and more animate, but did not improve performance in comparison to plain obstacle-avoidance paths.

III. TRUST-AWARE CONSERVATIVE CONTROL

A primary means for maintaining efficient human-robot collaborations is to rectify situations when the human loses trust in the robot. These may arise when the agent misbehaves, for instance when the tracker steers aimlessly after failing to detect boundaries. Trust can also be lost when task conditions are noisy, for example when a blurry boundary causes inaccurate and jittery controls. In both cases, the robot agent should elicit the human to provide assistance, and also attempt itself to limit the severity of its sub-optimal motions.

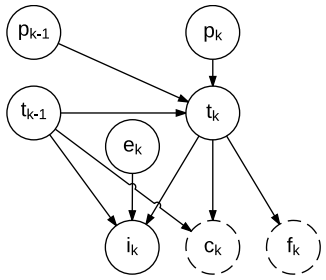

Fig. 3. The Online Probabilistic Trust Inference Model uses a Bayesian network in which some factors (dashed) are not present on all time steps k.

Towards these ends, we propose the strategy of Trust-Aware Conservative Control (TACtiC), in which the robot agent acts in a conservative manner to remedy trust losses. These conservative motions serve to both demonstrate hesitation towards eliciting the supervisor's help, and reduce the aggressiveness of its motions.

The TACtiC strategy constitutes two parts: how can robots display conservative behaviors, and when to engage and disengage these temporary alterations? The former is addressed via a "conservative state" of operations that alters motion commands issued by the agent. Focusing on the latter, we propose trigger conditions based on detecting *salient losses and gains* in the supervisor's trust, and discuss a statistical selection method for these personalized trust triggers.

A. Conservative Control Alterations

The general notion of "conservative behavior" refers to actions that purposefully aim to preserve existing conditions and limit change. We instantiate the "conservative state" $\mathcal{C}$ for vehicular steering by using both a speed reduction gain Γ and by smoothing the agent's steering signal ω_k. Speed reduction is useful for demonstrating uncertainty when the agent fails to detect any boundaries. Separately, we apply an exponential filter to transform the agent's steering command ω_k at each time step k into a smoothed version $\overline{\omega}_k$:

$$\overline{\omega}_k = a \cdot \overline{\omega}_{k-1} + (1 - a) \cdot \omega_k \quad \text{where} \quad a = e^{-\Delta K/\tau} \qquad (1)$$

The time constant τ introduces a temporal delay and signal smoothing, both in proportion to the command interval ΔK. These effects enforce predictable motions and help to attenuate oscillatory controls induced by blurry or otherwise noisy visual inputs. Sec. IV will investigate the impacts of different settings for the Γ and τ parameters.

B. Trust Triggers

The conservative state $\mathcal{C}$ is designed for situations where the human experiences salient trust loss, and thus should not be engaged at all times so as not to interfere with regular operations. Inspired by well-studied biases in human decision making that favor extreme and recent events [19], we define the "trust shift" δt_k within a recent window of $K \cdot W$ seconds as the difference between the latest expected trust state $\widehat{t}_k \triangleq E\left[bel\left(t_k\right)\right]$ and the most recent supremum or infimum:

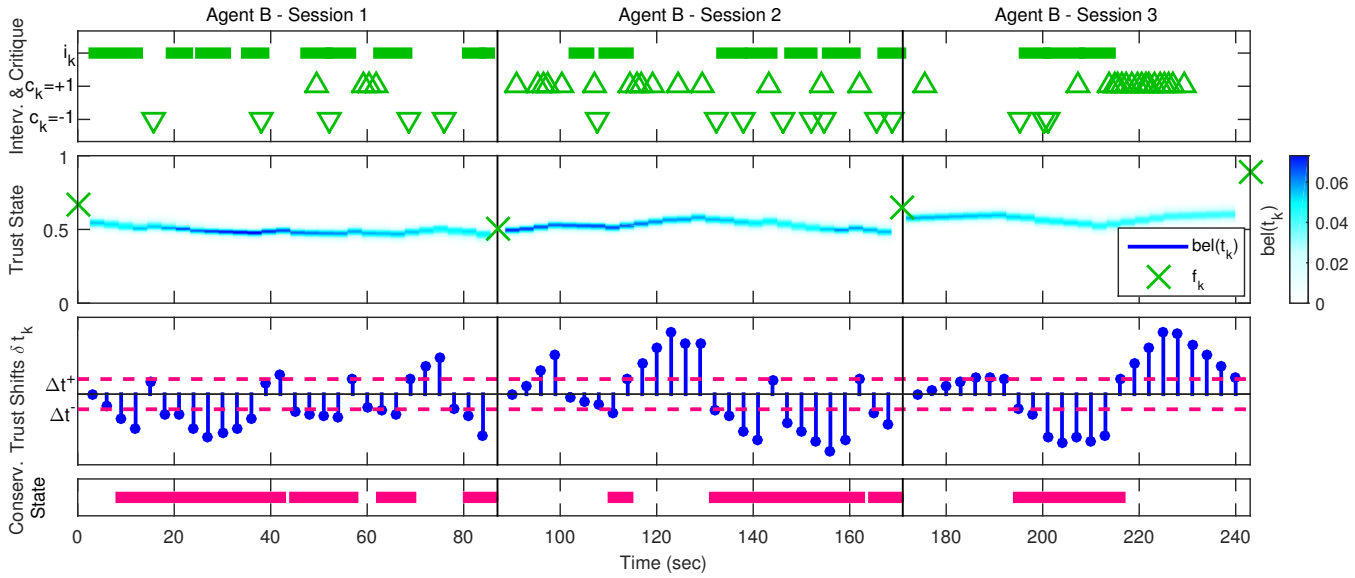

Fig. 4. Sample interaction dataset with the mild trust-aware conservative agent (B) from the user study. Interventions i_k and trust change critiques c_k are integrated into beliefs of the trust state $bel(t_k)$. The trust shift signal δt_k is computed from the expected trust state at each time step k. The agent engages and disengages conservative control as the user's trust shift surpasses the trust loss Δt^- and trust gain Δt^+ thresholds (purple dashed lines).

$$\delta t_k = \begin{cases} \widehat{t}_k - \inf\left\{\widehat{t}_{k-K}, ..., \widehat{t}_{k-1}\right\}, & \widehat{t}_k > \widehat{t}_{k-1} \\ \widehat{t}_k - \sup\left\{\widehat{t}_{k-K}, ..., \widehat{t}_{k-1}\right\}, & \widehat{t}_k < \widehat{t}_{k-1} \\ \delta t_{k-1}, & \widehat{t}_k = \widehat{t}_{k-1} \end{cases} \quad (2)$$

Fig. 4 depicts a sample trust belief sequence and its corresponding trust shift signal.

We engage the conservative state $\mathcal{C}$ when the trust shift signal falls below Δt^-, and disengage it when the signal surpasses Δt^+. These thresholds are computed by analyzing the statistical properties of each user's behaviors on the same dataset used to train our trust model. Specifically, we assume that the human will intervene each time the agent misbehaves or encounters a noisy task condition. After the agent successfully adapts to these interventions, we also assume that the supervisor will issue a trust gain critique ($c_k = +1$) as explicit approval. We hence scan each user's interaction experiences and identify the largest negative trust shifts during each intervention period, as well as the largest positive values during follow-up periods before a trust gain critique. The trust loss Δt^- and trust gain Δt^+ thresholds are then computed as the medians of these two sets of extreme values.

A benefit of this statistical selection method is that it does not need to identify each intervention's cause. Also, interventions due to non-trust related causes such as when switching boundary task targets have minimal impact on these thresholds. These factors are appropriately modeled by OPTIMo's intervention likelihood and have also been shown to not cause significant trust loss [3].

C. Algorithm Summary

Algorithm 1 summarizes the TACtiC behavior alteration strategy. This strategy post-processes every steering command from the autonomous agent before sending it to the robot's low-level control interface.

The simplicity of this strategy reflects the intuitive nature of human reactions to criticisms and trust loss at the workspace, namely by eliciting actionable feedback and adapting their behaviors. Another benefit of prompting the supervisor to help out the robot agent is the ability to cope with arbitrary misbehaviors without requiring separate detectors for different types of agent failures or noisy task conditions. Finally, it is important to acknowledge such trust-induced reactions are contingent on the ability to infer the human's evolving trust states in real time, and therefore is only enabled by our recent advances in online human-robot trust modeling [5].

IV. CONTROLLED INTERACTION STUDY

We conducted a large-scale interaction study with 46 participants to evaluate the potential efficiency gains of TACtiC and trust-seeking robots. In this study, users teamed up with boundary tracking agents to carry out aerial coverage tasks. An aerial drone simulator was used to ensure controlled and

Algorithm 1 Trust-Aware Conservative Control (TACtiC)

Input: recent expected trust states $\widehat{t}_{k-K}, ..., \widehat{t}_k$; conservative state $\mathcal{C}_{k-1}$; nominal speed cmd. ν_k; agent's steering cmd. ω_k; previous altered steering cmd. $\overline{\omega}_{k-1}$

$\delta t_k \leftarrow$ `computeTrustShift` $\left(\widehat{t}_{k-K}, ..., \widehat{t}_k\right)$ // Eqn. 2
$\mathcal{C}_k \leftarrow \mathcal{C}_{k-1}$
if $\mathcal{C}_{k-1} = 0$ **and** $\delta t_k \leq \Delta t^-$ **then** $\mathcal{C}_k \leftarrow 1$
else if $\mathcal{C}_{k-1} = 1$ **and** $\delta t_k \geq \Delta t^+$ **then** $\mathcal{C}_k \leftarrow 0$
$\overline{\nu}_k, \overline{\omega}_k \leftarrow \nu_k, \omega_k$
if $m_k = 1$ **then**
    $\overline{\nu}_k \leftarrow \nu_k \cdot \Gamma$
    $\overline{\omega}_k \leftarrow$ `runExpFilter`$(\overline{\omega}_{k-1}, \omega_k)$ // Eqn. 1
return $\mathcal{C}_k, \overline{\nu}_k, \overline{\omega}_k$

repeatable conditions, although frames from its downward-facing camera were synthesized from real satellite imagery.

A. Infrastructure and Interface

This study featured several task scenarios in which the user trained a boundary tracking agent to steer the aerial drone along designated terrain contours. The agent was not provided with these boundary targets directly and thus required human intervention when switching task goals. The drone operated at a fixed altitude and constant speed, although agents with trust-seeking abilities could reduce speed during conservative control. The boundary tracker processed camera frames at 10 Hz, while the OPTIMo trust model updated the human's inferred trust state at a $W = 3$-second interval.

Fig. 5 shows the visual interface for this user study. During interaction sessions, this interface depicts a live camera feed from the drone as well as steering commands from both the agent and the human, drawn as blue and green arrows respectively. The agent's commands were visualized even during periods of manual intervention, to assist the human in deciding when to yield control appropriately. As boundary targets changed between sessions, these were conveyed through a text overlay and were also read aloud using a synthetic speech engine. The score overlay reflected the coverage progress for the current session, although score increments were downscaled during interventions to incentivize delegating control back to the agent. Furthermore, users were prompted periodically by a `t?` icon to provide critiques reflecting salient changes in their trust state.

Each participant was provided with a gamepad to interact with the boundary tracking agents. The user could intervene at any time by pushing and holding an analog stick in the desired steering direction. Critiques could also be freely issued by pressing buttons corresponding to "trust gain", "trust loss", or "no trust change" (i.e. $c_k = \{+1, -1, 0\}$). Furthermore, the gamepad rumbled to warn users whenever the agent failed to detect any boundaries.

The interaction experience was divided into short sessions lasting 1-2 minutes each. After every session, the user was

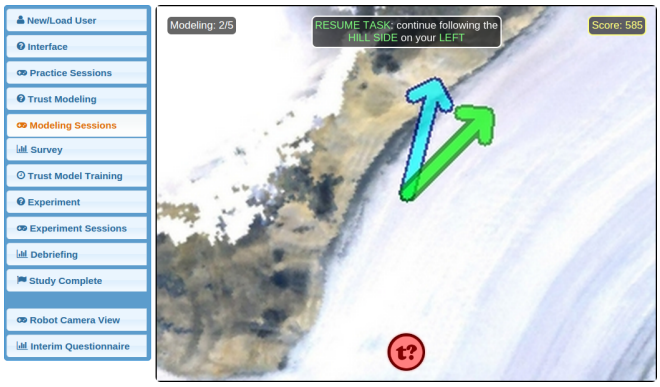

Fig. 5. Interface showing the aerial drone's live camera feed (right), and various study stages (left). The camera view is overlaid with the agent's steering command (blue arrow) and the human's interventions (green arrow). Other overlays provide information on session progress, current task goal, terrain coverage score, and prompts to issue trust change critiques (`t?`).

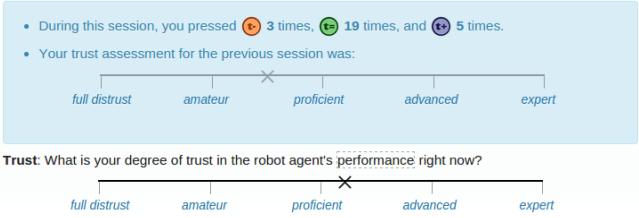

Fig. 6. Post-session trust feedback questionnaire.

asked to provide feedback about their current trust state, as seen in Fig. 6. The answer format used a continuous scale with multiple anchors, which was shown to facilitate repeated trust responses [5]. Both the previous trust response and trust critique counts were displayed in the questionnaire to remind users of their past behaviors and attitudes and help them provide consistent incremental trust feedback.

B. Experimental Setup

This study evaluated the impacts on team efficiency for the following autonomous agents:

- *Strongly Conservative* (`A`): agent with exaggerated trust-aware conservative control ($\Gamma = 60\%, \tau = 1.0$ sec);
- *Mildly Conservative* (`B`): agent with mild trust-aware conservative control ($\Gamma = 80\%, \tau = 0.5$ sec);
- *Baseline* (`C`): agent without conservative control.

The conservative motion parameters Γ, τ and the chosen memory window of $K \cdot W = 15$ seconds were empirically optimized during preliminary study testing. All three agents were equipped with the APEX module and hence could improve their task performance by adapting to occasional human interventions.

This study featured diverse scenarios with varying degrees of difficulty. These included tracking the smooth boundaries of highways and forest paths, following blurry snow-covered hillsides, and circumnavigating curvy coastlines populated with many inlets. The drone's operating speeds and altitudes were empirically tuned to ensure ample-paced task conditions with limited field of view. This design aimed to motivate users to allow the agent steer whenever possible, and especially when tracking rapidly-changing boundaries.

We quantified team efficiency through complementary sets of objective and subjective metrics. For instance, task performance was captured by the coverage area around the designated boundary path, whereas active workload was measured as the frequency of manual interventions. After interacting with each of the three agents, users were asked to assess their perception of the agent's active collaboration efforts. Along with post-session trust feedback, these assessments reflected the user's subjective attitude towards each agent.

These metrics quantify distinct aspects of team efficiency, and are all essential in forming a thorough evaluation of trust-seeking robots. We computed separate rankings of the three agents for each metric, session, and user. Mean rankings were then obtained by separately grouping objective (performance and workload) and subjective (collaboration and trustworthiness) aspects of team efficiency, and were analyzed statistically using Friedman test and post-hoc Nemenyi testing [20].

We further used the Kemeny-Young voting method [21] to corroborate these aggregate efficiency rankings. In summary, this study aimed to validate the following two hypotheses:

Hypothesis 1 *The mildly conservative trust-aware agent (B) yields greater **objective** efficiency ranking, compared to strongly conservative (A) and non-conservative (C) agents.*

Hypothesis 2 *The mildly conservative trust-aware agent (B) yields greater **subjective** efficiency ranking, compared to strongly conservative (A) and non-conservative (C) agents.*

This study was separated into three phases: coaching users, collecting interaction data to build a personalized trust model, and evaluating the three robot agents. The study began with an initial set of slides introducing the task and interface. Next, the user was led through an interactive tutorial and two practice sessions to familiarize with supervising the boundary tracking agent. One practice session purposefully featured an ambiguous forest-tracking task, where the robot often veered off due to narrow branching tree-lines. This setup allowed the user to practice issuing critiques and intervening, and further helped to calibrate their trust expectations. The second study phase consisted of 5 trust modeling sessions featuring diverse scenarios such as tracking highways, forest paths, hillsides, and coastlines. The trust model was subsequently trained while the participant filled out a demographics survey. In the third phase, the user partnered up with the trust-seeking and baseline agents in a counter-balanced repeated-subjects design. The study interface instructed users to treat agents independently, and also asked them to assess each agent's collaborative efforts at the end of 3 sessions. Both the modeling and evaluation phases entailed 20 minutes of interaction each, and the entire study lasted about 60 minutes.

C. Results and Discussion

We recruited 46 participants (13 females) with varied age ($\mu = 28, \sigma = 6$) from McGill University's School of Computer Science. Users had diverse levels of prior robot experience and included 8 undergraduate students, 32 graduate students studying robotics, 4 professors involved in robotics research, and 2 robot engineers. We purposefully targeted participants with technical backgrounds, since these users will likely be among the first adopters of robot technologies as they become mainstream.

The cross-validated OPTIMo instances resulted in minimal training errors, as reflected by small Root Mean Square Errors (RMSE) for predicting trust feedback $RMSE_f = 0.08\,(\sigma = 0.04)$ and high accuracies in predicting interventions $acc_i = 78\%\,(\sigma = 9\%)$ and trust change critiques $acc_c = 63\%\,(\sigma = 11\%)$. Our success in personalized trust modeling was further substantiated by comparable test-phase prediction performance: $RMSE_f = 0.14\,(\sigma = 0.07)$, $acc_i = 73\%\,(\sigma = 14\%)$, $acc_c = 58\%\,(\sigma = 14\%)$.

Friedman test on the objective efficiency rankings revealed weakly significant differences among the agents ($\chi^2\,(2) = 5.609, p \leq 0.10$). A similar analysis found statistically significant differences among the subjective efficiency rankings ($\chi^2\,(2) = 11.783, p \leq 0.01$), and post-hoc analysis

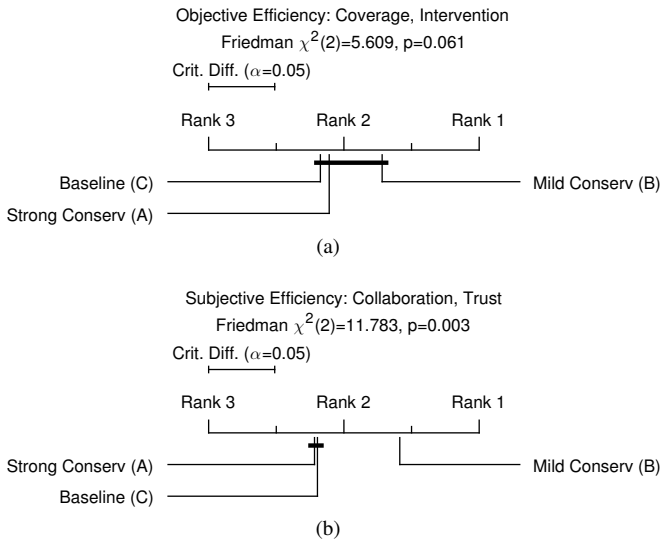

Fig. 7. Critical difference plots of mean agent ranks across users of the controlled study, separately for objective and subjective efficiency metrics.

depicted in Fig. 7b showed dominant preference for the mildly conservative agent (B). Both mean rank orderings were corroborated by identical across-user rankings computed using the Kemeny-Young method.

Compared to the baseline agent under each efficiency metric separately, both trust-seeking agents attained greater terrain coverage, required fewer interventions, and were perceived as more collaborative. The mildly conservative agent (B) was also more trusted than the baseline agent (C), whereas the strongly conservative agent (A) was the least trusted. This distrust arose from frustrations towards agent A's slow speed and delayed reactions, yet users nonetheless acknowledged its active efforts to maintain teamwork.

Per-metric efficiency results did not reveal any significant differences between the three agents. These numerical similarities are justified since the metrics were *cumulative* over long interaction sessions, whereas trust-seeking agents only exhibited different behaviors in *rare occurrences of notable trust loss*. Also, both the trust-seeking and baseline agents have behavior adaptation capabilities, which has been shown to yield significantly greater efficiency when compared to non-adaptive agents and to plain teleoperation [6].

This user study demonstrated that mildly conservative agent behaviors induced by the human's trust state consistently contributed to superior team efficiency. Also, the successful validations of Hypotheses 1 and 2 were contingent on several precursor conditions, including proper user coaching, accurate trust modeling, and suitable selection of trust triggers. These conditions were all met within the study's compact duration, thus reflecting the rigor of our iterated design. Furthermore, multiple users reported perceiving the trust-seeking agent to having great foresight, since its filtered steering commands smoothly tracked the curvy coastline without wastefully turning in and out of every little inlet. These perceptions further substantiated the benefits of trust-aware conservative agents, both at improving team performance and also attaining greater satisfaction.

V. ROBOT FIELD STUDY EXTENSION

We extended the previous study by deploying trust-seeking agents onboard a smart car, and invited 12 passengers to supervise and train these agents to drive along a challenging gravel course. These field trials aim to re-assess the efficiency gains of TACtiC-equipped agents during *real-world interactions*, while re-using previously trained trust models.

A. Infrastructure and Interface

In collaboration with the Canadian Space Agency, we deployed boundary tracking agents onboard their SL-Commander all-terrain vehicle, as seen in Fig. 1. This electric car can travel at up to 50 km/h in both paved and off-road conditions. It features a drive-by-wire system allowing programmable control of the accelerator pedal and steering wheel. This vehicle is equipped with a high-resolution camera along with a pan-tilt unit mounted on its front hood.

Most of the infrastructure was directly transplanted from the controlled study, including identical settings for the APEX interactive adaptation method and the OPTIMo trust inference engine. The boundary tracking algorithm was altered to steer along ground-plane boundaries based on a front-facing slanted camera view [6]. The camera was set to a fixed pose of 8° right-side pan and 20° downward tilt.

Each participant sat in the passenger seat while supervising the boundary tracking agents. A similar visual interface featuring a live camera feed with overlaid steering commands was displayed on a tablet placed below the windshield on the passenger's side. Since some users preferred to keep their eyes on the road at all times, an audio cue was used to prompt for trust change critiques every 5 seconds.

Participants interacted with autonomous agents using the same gamepad interface for issuing interventions and critiques. Users could also press buttons to increase or decrease the car's nominal speed by ±1 km/h, between 0 km/h initially and a maximum of 20 km/h. During preliminary trials, we found that speed changes consistently correlated to changes in trust, where one would decrease speed in response to abrupt motions, and increase speed after appropriate training. We thus configured speed change commands to automatically generate "trust gain" and "trust lost" critiques as well.

B. Experimental Setup

Extending the findings from the user study, we again contrasted the mildly conservative trust-aware agent (B) against the non-conservative agent (C). The task scenario, shown in Fig. 8, was composed of a 1 km gravel circuit that was surrounded by tall grass and various hurdles impacting agent behavior and user comfort. We re-validated Hypotheses 1 and 2 using a near-identical set of objective and subjective efficiency metrics, with the sole exception of quantifying performance via each session's elapsed duration.

Each trial run entailed 3 interaction sessions along the test course. The first session paired up the user with the baseline agent to practice with the control interface and course layout. The following two sessions featured the mildly conservative agent (B) and the baseline agent (C) in a randomized order.

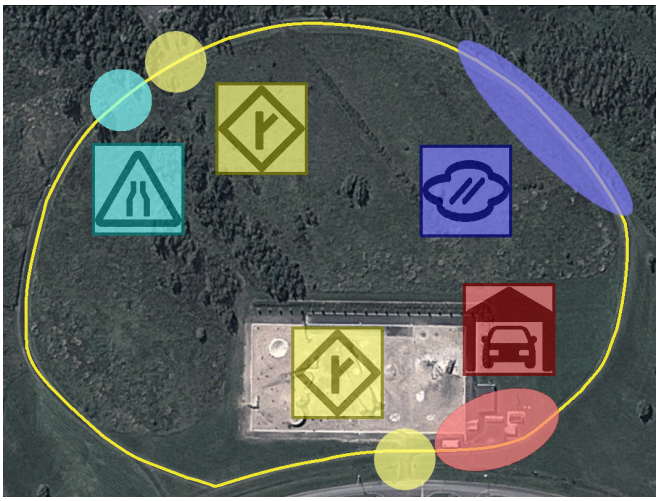

Fig. 8. The test site at the Canadian Space Agency features a 1 km gravel circuit (yellow) with various hurdles, including a narrow overpass (cyan), intersections (yellow), a watery ditch (blue), and nearby parked cars (red).

C. Results and Discussion

We invited all participants from the user study to attend these field trials. Nevertheless, 12 users (1 female) satisfied the security requirements needed to enter the test site at the Canadian Space Agency. These users were all actively engaged in robotics research, and comprised of 8 graduate students, 2 professors, and 2 robot engineers.

We evaluated trust prediction accuracies of the previously-trained models on each user's field study dataset: $RMSE_f = 0.21\,(\sigma = 0.18)$, $acc_i = 69\%\,(\sigma = 5\%)$, $acc_c = 50\%\,(\sigma = 15\%)$. These results were comparable to the previous study's model performance, although there was slightly larger error when predicting trust feedback. We suspect that users now had vested interest in their physical well-being while sitting in the self-driving car. Thus, they may have adopted more cautious attitudes towards the robot's performance.

Aggregate rankings in Fig. 9 show that the conservative

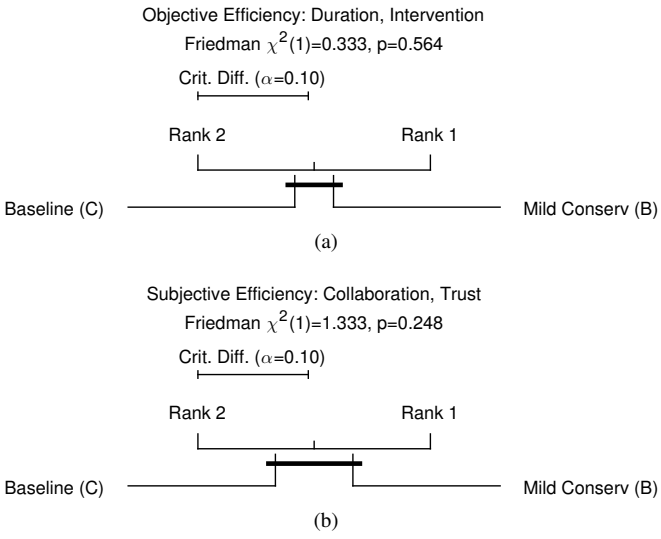

Fig. 9. Critical difference plots of mean agent ranks across users of the field study, separately for objective and subjective efficiency metrics.

agent was favored against the baseline agent in both objective and subjective efficiency facets. These preferences were supported by identical orderings computed using the Kemeny-Young scheme. Nevertheless, neither the objective efficiency results (Wilcoxon signed-rank: $p = 0.774$; Friedman: $\chi^2(1) = 0.333$, $p = 0.564$) nor the subjective rankings (Wilcoxon signed-rank: $p = 0.388$; Friedman: $\chi^2(1) = 1.333$, $p = 0.248$) were statistically significant. Individual efficiency measures were also not statistically different between the trust-seeking and baseline agents. These likely resulted from the low number of users and study sessions, both of which were due to severe operational constraints at the test site.

These field trials qualitatively re-affirmed the interaction study's findings, thus re-substantiating the efficiency merits of mildly conservative trust-aware agents. All participants expressed content with the agents' performances and with the overall experience of collaborating with a smart car. One user commented: "Since I was pressing the trust change buttons constantly during autonomous control, it dawned on me pretty late that I was not actually in control; even then I felt unusually comfortable." Another stated: "Once I got used to teaching the agent quickly, I was more confident in the narrow parts of the course, and overall my trust was increased." These reports suggest that agents that can react to user's actions and attitudes have great utility and promise towards enabling autonomous cars and other future robots to achieve efficient and trusting collaboration with humans.

VI. CONCLUSION

In this work, we introduced the trust-seeking robot framework along with the novel formulation of Trust-Aware Conservative Control (TACtiC). Our framework is the first-ever realization of robot agents that react *in direct and explicit response* to changes in their human collaborator's trust state. We presented end-to-end instantiations of trust-seeking robot agents for distinct tasks of collaborative aerial coverage and interactive autonomous driving, and notably deployed and evaluated these agents onboard an actual manned vehicle.

Results from our large-scale controlled interaction study and field study extension have shown that the pervasive human notion of trust can be capitalized by autonomous agents to maintain efficient human-robot collaborations. The efficiency gains of these trust-seeking agents also reflect the successful and accurate modeling of each user's trust tendencies. Finally, while many robot learning methods can improve performance, few systems have the ability to cater to each human's individual preferences. Our trust-seeking robot framework fulfills both objectives by imbuing robot agents with the capacity to adapt to their human collaborator's actions *and* attitudes.

We are keen to study extensions of the trust-seeking robot framework to improve its robustness during longer-term interactions. In particular, we would like to update each user's trust model as they acclimate to the robot agent over time. We are also considering data-driven approaches for determining initial trust shift thresholds and adapting them over time, to reflect long-term evolutions in the supervisor's attitudes. Finally, TACtiC is currently applicable to a wide variety of locomotion-based robots, including autonomous cars, drones, and aquatic vehicles. Nevertheless, we are excited by the possibilities of expanding this strategy to cover even more robotic platforms, such as factory manipulators and personal assistance robots.

ACKNOWLEDGEMENT

The authors would like to thank all of the participants in the interaction study and field study extension, which were sanctioned by McGill University's Research Ethics Board (#183-1112). We are also very grateful to the Canadian Space Agency, and especially to Tom Lamarche and Erick Dupuis, for their support in realizing the robot field study. This work was proudly funded by the NSERC Canadian Field Robotics Network (NCFRN).

REFERENCES

[1] J. Lee and K. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, 2004.
[2] M. Desai, "Modeling trust to improve human-robot interaction," Ph.D. dissertation, University of Massachusetts Lowell, 2012.
[3] A. Xu and G. Dudek, "Towards modeling real-time trust in asymmetric human-robot collaborations," in *Int. Symp. on Rbt Rsrch (ISRR)*, 2013.
[4] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse." *H Factors*, vol. 1, 1997.
[5] A. Xu and G. Dudek, "OPTIMo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *Proc. of the ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2015.
[6] A. Xu, A. Kalmbach, and G. Dudek, "Adaptive Parameter EXploration (APEX): Adaptation of robot autonomy from human participation," in *Proc. of the IEEE Int. Conf. on Robotics and Auto. (ICRA)*, 2014.
[7] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6-7, 2008.
[8] W. B. Knox, "Learning from human-generated reward," Ph.D. dissertation, University of Texas at Austin, 2012.
[9] B. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, 2009.
[10] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Survey: Robot Programming by Demonstration," Handbook of Robotics, ch. 59, 2008.
[11] D. H. McKnight and N. L. Chervany, "The meanings of trust," University of Minnesota, Tech. Rep., 1996.
[12] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, 1992.
[13] R. J. Hall, "Trusting your assistant," in *Knowledge-Based Software Engineering Conf. (KBSE)*, 1996.
[14] B. M. Muir, "Operators' trust in and use of automatic controllers in a supervisory process control task," Ph.D. dissertation, University of Toronto, 1989.
[15] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. of Cognitive Ergonomics*, vol. 4, no. 1, 2000.
[16] C. Breazeal, C. Kidd, A. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2005.
[17] T. Kazuaki, O. Motoyuki, and O. Natsuki, "The hesitation of a robot: A delay in its motion increases learning efficiency and impresses humans as teachable," in *ACM Int. Conf. on Human-Robot Int. (HRI)*, 2010.
[18] A. Moon, C. A. Parker, E. A. Croft, and H. M. V. der Loos, "Design and impact of hesitation gestures during human-robot resource conflicts," *J. of Human-Robot Interaction*, vol. 2, no. 3, 2013.
[19] B. L. Fredrickson, "Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions," *Cognition and Emotion*, vol. 14, no. 4, 2000.
[20] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. of Machine Learning Research*, vol. 7, 2006.
[21] P. Young, "Optimal voting rules," *J. Econ. Perspectives*, vol. 9, 1995.