HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240206

Exploring the Dynamic Nature of Trust Using Interventions in a Human-AI Collaborative Task

Sachini WEERAWARDHANA [a,1] and Michael AKINTUNDE [a] and Luc MOREAU [a]
[a] *Department of Informatics, King's College London, London, United Kingdom*
ORCiD ID: Sachini Weerawardhana https://orcid.org/0000-0002-8043-0507, Michael Akintunde https://orcid.org/0000-0002-5031-8813, Luc Moreau https://orcid.org/0000-0002-3494-120X

Abstract. People are increasingly interacting with machines embedded with intelligent decision aids, sometimes in high-stakes environments. When a human user comes into contact with a decision-making agent for the first time, it is likely that the agent's behaviour or decisions do not precisely align with the human user's goals. This phenomenon, known as *goal alignment*, has been recognised as a critical concern for human-machine teams. Prior work has focused on the effect of automation's behavioural properties, such as predictability and reliability, on trust in human-machine interaction scenarios. However, little is known about situations where automation's capabilities are misaligned with humans' expectations and its impact on trust. Even less is known about the effect of environmental factors on trust. We study the relationship between intervention behaviours and trust in a simulated navigation task where the human user collaborates with an agent with misaligned goals. We evaluate trust quantitatively using intervention frequency as a behavioural measure and qualitatively using self-reports. By advancing the understanding and measurement of trust in collaborative settings, this research contributes to the development of trustworthy and symbiotic human-AI systems.

Keywords. interventions, trust, uncertainty, human-agent interaction, goal alignment

1. Introduction

There is a notable increase in interactions between people and machines with integrated intelligent decision aids. Investors use trading agents to manage trades and make their money work smartly. Vehicles embedded with intelligent technologies process and communicate real-time information to the driver. When a human user initially comes into contact with a decision-making agent, most likely programmed by someone else with a different understanding of how the agent should function, the agent's behaviour or decisions may not fully align with what the user wants to accomplish. This problem, known as *goal (mis)alignment*, has been recognised as a critical concern for human-machine teams [1]. An example of goal misalignment is when Google Maps suggests an alterna-

[1] Corresponding Author: Sachini Weerawardhana, sachini.weerawardhana@kcl.ac.uk

tive shorter route while driving along a longer but scenic route. When the human recognises the goal misalignment, for instance, by noticing the agent making recommendations that are not fully aligned with the human's interests, human interventions occur. Considering the Google Maps-assisted driving example, such an act of intervention is when the human, preferring to experience the scenic drive, rejects the alternative shorter route. In the scope of this paper, we define interventions as a human-initiated action to alter the agent's behaviour, specifically, rejecting the decision and suggesting an alternative to modifying the decision-making process to better align with the human's goals. In automated driving, intervention behaviour is commonly phrased as "takeover behaviour" and is a widely researched topic [2–4]. Considering the behavioural opposite of intervention—compliance, where the human accepts the agent's decisions—misaligned goals can have disastrous outcomes, especially in safety-critical situations. For instance, in a recent incident, tourists following GPS directions were led straight into a harbour in Kailua-Kona, Hawaii [5]. Although the drivers were unharmed, tow crews had to pull the fully submerged SUV out of the water. The "flash crash" incident in 2010 [6], where humans relied on complex autonomous systems consisting of agents responsible for trading decisions, led to large monetary losses and has significantly affected regulations for US equity markets.

Compliance is often used as a behavioural demonstration of trust in human-machine interactions [7–9]. Muir [10] laid the foundation for understanding the relationship between trust and interventions in a supervisory control setting, arguing that human interactions with automation should not be viewed as a "*once and for all*" activity but rather a "*dynamic process*" where the human decides to either intervene or leave the system running under automation. The study found that automation's unreliability affects performance, trust, and self-confidence, impacting the human operator's decision to switch between automatic and manual modes. Prior work has focused on the automation's properties, such as predictability [11, 12] and reliability [13, 14]. However, little is known about situations where automation's capabilities are misaligned with humans' expectations and its impact on trust. For example, while on a trip, the driver uses a GPS-enabled system to find places of interest nearby. The system, programmed to present the closest locations first, will give the driver some options. If the driver chooses the closest attraction, it is not guaranteed that they will enjoy the attraction. If the user frequently rejects a system's suggestions because the results are not exactly what the user wants, this can lead to a loss of trust. Even less is known about the impact of environmental factors, which we define as external events occurring outside the control of the agent and the human. Environmental factors are important to consider in systems deployed in real-life conditions. For example, an unplanned road closure may happen while finding the places of interest using the GPS-enabled navigation system. We argue that such events may further impact the trust when working with agents with misaligned goals. Stated differently, our study finds evidence for miscalibrated trust [15] in environments where the human's and agent's goals are misaligned, evidenced through interventions. Because miscalibrated trust leads to systems' misuse and disuse, our study warrants further investigation into trust calibration when the human's and the agent's goals are misaligned. Meta-reviews on the trust in human-automation interaction literature [16, 17] highlighted the limited number of studies exploring environmental factors in trust development. In this paper, we address that gap and aim to understand how the agent's capability to act as a decision aid and the uncertainty in the environment impact trust.

We designed the human-agent interaction use case to allow the researcher to sample trust at regular intervals during the task while carefully balancing the workload of the human user Past studies have shown that a high workload may force a human user to rely on automation [18], creating an illusion of trustworthy behaviour. This poses a challenge to capturing trust using behavioural measures. In particular, it is recommended that for a measurement like an intervention to be feasible, the workload must be low [19]. Our interaction use case involves the human completing a navigation task travelling from point A to B with the help of an automated route planning agent. The agent finds a path to the destination and reveals only the next direction from the current position. When the human intervenes and proposes a new direction, the agent automatically recalculates a new path, including the human's suggestion, and the task continues until the destination is reached or the budgets have run out. Route planning tasks have commonly been used to evaluate trust with respect to automation reliance behaviours in human-robot [20–22] and human-agent interactions [23, 24].

Contributions. We explore how the number and frequency of interventions as a behavioural measure of trust vary according to an agent-related factor, i.e., the agent's capability to act as a decision aid, and an environment-related factor, i.e., the uncertainty.

We model *capability* in our route planning use case by assuming that the human needs can be modelled by a set of criteria (e.g., distance and time) in the route planning task. Specifically, the human wants to drive the shortest distance as quickly as possible. In our design, the agent can only optimise for one criterion (either distance or time), defined as the agent's capability. Thus, the capability is a partial alignment of what the agent optimises for and the goal of the human. Note that this differs from the agent giving faulty advice, as is much evaluated in past works on reliability and trust in human-agent interactions. The agent's suggestions are "correct" based on the agent's understanding of the world model, and the agent always proposes a direction from the optimal route considering the current state.

We model *uncertainty* as a *non-determinism in the environment*, where the movement of the placement marker indicating the vehicle's current position does not solely depend on the agent's or the human's choices. In the GPS-assisted driving example, our uncertainty modelling is similar to a situation where the driver turns into a by-road, a traffic officer approaches, informs about a sudden road closure and forces the driver to take a direction different from the one suggested by the agent or intended by the human. We display this visually in the interactive interface, where the moving action does not always take the placement marker in the intended direction. Note that this is different from the uncertainty arising from the agent itself, such as failures or mistakes. Given this setup, we answer the following research questions:

RQ1: What effect do the agent's capability and the uncertainty in the environment have on the intervention frequency measured at regular intervals during the task?

RQ2: What is the relationship between intervention frequency measured during the task and the user's confidence in agent capability and environment uncertainty?

We hypothesise that the agent's capability and uncertainty in the environment affect intervention frequency and self-reported trust and that users' confidence in automation is lower when there is environmental uncertainty. Further, we expect a human to intervene more and have less confidence in an agent where the consequences of goal misalignment are more severe. Rather than providing a *snapshot* view of trust measured at a specific point (e.g., at the end of the interaction), our study provides evidence for its variability during the task. We lay a foundation for understanding the relationship between self-reported trust and human interventions when the human and agent's goals are misaligned and when the agent's operating environment is uncertain. In the long term, such an understanding will lead to more effective human-agent partnerships.

2. Literature Review

The navigation task we designed aligns with the collaboration principles advanced in [25]. The agent is available throughout the task and provides direction suggestions to the human (responsiveness). Whenever the human suggests an alternative direction, the agent reruns the path planning algorithm and generates a new path that includes the direction the human suggested (joint activity). The human can freely decide to use the agent's recommendation. Trust can lead to cooperative behaviour in human-agent collaborations [11, 26] in situations where the human trustor will be at risk if not for the cooperation.

Antecedents of Trust. We adopt the trust definition advanced by Lee and See [27] for this work. Antecedents of trust are classified as automation-related, operator-related and environment-related [16]. Automation-related covers automated machines as well as embodied and virtual agents. We present capability as automation-related and uncertainty as environment-related factors.

Automation-related. In interpersonal settings, the trustee's ability signals trustworthy perceptions to the trustor [15]. In human-automation interaction contexts, the automation's ability is reflected in reliability, faults, predictability, transparency and automation level, which have been shown to impact trust. Reliability has been shown to positively affect trust [13, 14, 26]. Closely related to reliability, automation faults negatively affect trust. Fault occurrence frequency [28], timing of the occurrence [29], and the magnitude [30] impact trust in different ways. Automation predictability [11] positively impacts trust. Similarly, transparency (i.e., explaining the reasoning process) [31] positively impacts trust. The level of automation, mostly evaluated in automated driving scenarios [32], impacts human trust.

The automation-related factor we introduce in this study, i.e. *capability*, is a construct related to the agent's ability as defined by Mayer et al. [15]. However, we advance a nuanced take on the grounded definition of the ability, which asks whether or not the agent is capable of fulfilling its commitment [33]. Capability is derived from goal alignment definition in [34]; the degree to which the human's goal matches the AI's programmed goal. Specifically, the agent can only plan the route by optimising for one of the two criteria the human needs. The agent's suggestions are "correct" by its understanding of the world model, and it always provides the best recommendation given the current state and the goal that the agent optimises; the agent is neither faulty nor unreliable. A different representation of goal misalignment is presented in [35], where they consider the alignment of reward functions between collaborating agents.

Environment-related. Hancock et al. identified *team collaboration* and *tasking* as environmental factors [17]. Team collaboration refer to in-group membership, culture, communication and shared mental models. Tasking consider task type, complexity, multitasking requirement and physical environment. The study found moderate effects of environmental factors on trust development. They highlight a strong need for future empirical work to study the relationships between environmental factors and trust, citing the limited number of studies available.

Hoff and Bashir's trust model [36] classified environmental factors as an antecedent to situational trust in human-automation interactions. They define risk and benefits of using automation as environmental factors. A similar link has been established in [37], suggesting that risk impacts trust and reliance behaviours in human-automation interactions. In a high-risk driving scenario, the participants trusted and used the GPS driving advice less [38]. A study by Hoesterey and Onnasch manipulated risk by altitude and measured trust attitude and behaviour in a decision automation task [39]. Results showed that trust attitude was not affected by risk. However, trust behaviour was higher and increased during the experiment for the automation-supported group. Conflicting results reported in these prior studies further highlight the empirical necessity to improve the understanding of the relationship between trust and environmental factors. Our study takes a step towards addressing this gap.

Adopting the definition advanced by Hoff and Bashir [36], in our study, we use the *uncertainty* factor to manipulate the risk in the environment. Our definition of uncertainty differs from the agent-related factor predictability in Mayer's trust model [15]. We model uncertainty as non-determinism in the environment, where external elements prevent movement in a direction suggested by the agent or human. Uncertainty in the environment does not cause the agent to replan. The agent perceives the unexpected landing position and recommends the best direction to take from that position based on the policy it has already generated.

2.1. Measuring Trust

Trust is typically measured as behavioural measurements, self-reports, and physiological measurements. We focus our literature review on behavioural measurements and self-reports because, in this study, we use interventions as a behavioural measure and self-reported trust as a post hoc measurement.

While Muir's study affirmed intervention as a behavioural indicator for trust (or lack thereof), a recent review on measurement of trust in automation [19] recognises that behavioural measures are capable of sampling trust at a much higher rate than self-reports, which are typically administered before and after the interaction in experimental conditions. Frequent sampling generates a more accurate measure of trust because it allows the researcher to capture "the area under the trust curve" as defined in [40]. Frequent sampling allows the researchers to capture the temporal nature of trust, yet another under-explored antecedent to trust [41]. In [34], to measure the temporal trust dynamic, trajectory epistemic network analysis is used to show the evolution of trust in human-AI conversations. We demonstrate the dynamic variation of trust, measured quantitatively with interventions using a higher rate of sampling measured periodically during the task.

Behavioural measures can be active (e.g., compliance with instructions or recommendations issued by automation, operator intervenes by taking control over from automation), passive (e.g., reliance) or engaging in risk-taking in the relationship [19]. An advantage of using a behavioural measurement is that it generates a continuous metric instead of a one-time measurement at the end of the task [37]. Our study uses intervention frequency during the task as an active behavioural measurement. A model predicting the trust in automated parking features consisted of the proportion of trials in which the driver intervened in addition to operator-related and automation-related constructs [2].

Prior work has recognised the contentious relationship between behavioural measurements and trust, claiming that behavioural measurements can be influenced by factors other than trust, such as workload [42] and risk [39]. Kohn et al. recommended that experiments be designed to confirm that behavioural trust measures correlate with other trust measures, such as using validated self-reported trust measurements [19]. In [43], self-reported trust (using a validated trust in automation scale in [44]) was measured periodically during the task alongside task performance-related metrics.

3. Methods

We now describe the navigation use case, how the independent variables, capability and uncertainty were operationalised, and the experimental setup.

3.1. The Navigation Simulator

We created a web-based interactive simulator where a virtual agent assisted a human user in planning a route. A valid route took the user from a starting location (*home*) to the destination (*hospital*), passing through three locations identified as landmarks (a *grocery store*, a *construction site* and a *school*). Although a human could enforce their own order of visiting the landmarks, the plans generated by the algorithms enforced an implicit temporal ordering for landmark visits: starting from the grocery store to the school, to the construction site, and finally the hospital. The agent used two route planning algorithms: the A^* search algorithm with the admissible heuristic landmark-cut [45] and Q-Learning [46] to generate a route to the destination, going through all un-visited landmarks. The generated route was incrementally revealed to the user one action at a time as direction suggestions. The location map is laid on a grid. Therefore, the direction suggestions were UP, DOWN, LEFT and RIGHT. We introduced risk and vulnerability to the user by constraining the route planning task to be within a time (t) and distance (d) budgets. We ensured that the budgets were sufficient to complete the route planning task in pilot studies. We also simulated critical events during the route planning task. An accident occurred halfway through the trip (i.e., less than half of the distance budget was remaining) where two cells on the grid became inaccessible, and the agent had to replan as its position was pushed to an adjacent cell. The interaction scenario was presented as a cover story where the participant was asked to imagine a situation where they were driving a friend to the hospital.

We designed a 10×10 grid world to keep the participants' workload at a minimum to reduce over-relying on the agent and encourage collaborative behaviour. Every trip started from the *home* position (bottom left of the grid) and ended at the *hospital*, the final position (top right). Each cell was associated with two costs: distance and time. Distance cost was uniformly distributed (cost=1) and indicated at the top-right corner of the cell. The default time cost was 1. However, some areas have a higher time cost (cost=3), indicated in red and described as "unsafe" in the cover story. The remaining budgets, visited milestones and the agent's suggested direction were displayed to the participant. Clicking the Go button indicated that the participant accepted the suggestion, thus moving the placement marker in that direction and updating the budgets. The participant could intervene by clicking the Interrupt button, at which point a web pop-up screen was displayed requesting the participant to input a new direction. Then the agent replans the route with the user's suggestion and incrementally reveals it as before[2].

Modelling Agent Capability. We model agent capability as the partial alignment of what the agent optimises for and what the participant was required to accomplish in the route planning task. The participant was instructed to complete the route-planning task without overrunning both the time and the distance budgets, that is, travel the shortest distance while avoiding unsafe areas. However, the agent can only plan a route to travel the shortest distance or avoid unsafe areas, but not both. The agent's capability was clearly described to the participant before the task. The distance-optimising agent aimed to take the shortest path, even if this involved passing through red cells with a higher time cost. The time-optimising agent aimed to avoid red cells as much as possible, which resulted in a lengthier route.

The task introduces Assistant Dede, which optimises for distance, and the Assistant Cece, which optimises for time. We hereafter refer to a_d and a_t for the distance-optimising (Dede) and time-optimising (Cece) agents, respectively. The two agents differ in the consequence of goal misalignment to the human. If the agent-human collaboration results in a distance budget overrun, more likely when working with Cece, the route planning task immediately terminates. If the time budget overruns, more likely when working with Dede, the task does not terminate; however, alert messages indicating the passenger's deteriorating health condition (as per the cover story) will be displayed repeatedly. The goal alignment problem and the respective consequences were explained to the participant before the experiment.

Modelling Environment Uncertainty. We model uncertainty as non-determinism in the environment, where the moving action does not land the placement marker in the intended direction at all times; the direction of the movement itself is uncertain and does not solely rely on the agent's choice. This takes a similar form to Open AI's FrozenLake-v0 environment [47]. We consider two levels: with and without non-determinism. Without uncertainty, the moving action lands the marker in the expected direction. With uncertainty, we assume an 80% probability of moving in the intended direction and a 10% probability of moving in either of the directions perpendicular to that intended. These two levels are referred to here as a *deterministic* environment without uncertainty and a *non-deterministic* environment with uncertainty. We use a planner that implements the A^* algorithm for the deterministic environment and reinforcement learning (RL) for the non-deterministic environment.

3.2. Experimental Design

We adopted a between-group mixed design described in [48] to resolve RQ1 and RQ2. The study received approval from the university's Ethics Review Board. The participants were recruited via Prolific [49]. After completing informed consent online, participants were randomly assigned to the deterministic or non-deterministic group. We ensured that

[2] The source code for the web-based navigation simulator is publicly available with the Creative Commons Universal License at `https://github.com/sachinisw/HHAI24-Navigation-Simulator.git`

participants who took part in one uncertainty condition were excluded from the other. Before the tasks, participants were asked to read the cover story carefully. Instructions to use the web application were provided in writing and via a tutorial video. Then, the participants in each group completed the route planning task once with the distance-optimising agent and once with the time-optimising agent. The agents were presented in random order.

For each landmark visited, we collected the number of times the participant rejected the agent's suggestion. We use the number of rejections as a measure of intervention frequency. Further, similar to the process followed in [43], the user was periodically polled at each landmark point to rate their confidence in the agent's ability to help complete the route planning task on a 10-point Likert scale. When the trip was forcibly terminated by distance budget run out, or when the participant successfully reached the hospital, they filled out a 4-question survey commonly used to measure trust in automation and has been empirically validated developed by Muir [30].

3.3. Participants

The non-deterministic group consisted of 44 participants (52% female, 42% age 44 or above). The majority had completed either a Bachelor's degree or school-level compulsory education (32% each). The deterministic group consisted of 47 participants (35% ages 20-25, 67% female), with a 33% being college-educated with Bachelor's degrees.

4. Evaluation

In this section, we discuss the experiments we conducted through user studies to answer the research questions (RQs) outlined in Section 1.

4.1. RQ1: Effect of Agent Capability and Environment Uncertainty on Intervention Frequency

We use the frequency of the participants rejecting the suggestions by the navigation agent to answer RQ1. We refer to this quantity as *interrupts*, a measure of intervention frequency. We hereafter refer to the case of a deterministic environment with an agent optimising for distance and time as DET-a_d and DET-a_t respectively, and for a non-deterministic environment, NONDET-a_d and NONDET-a_t respectively.

We report both the raw (non-cumulative) number of interrupts observed for each milestone in isolation and the cumulative number of interrupts, which for each milestone is the sum of the total number of interrupts for previous milestones reached and the number of interrupts at the current milestone. When observing the cumulative number of interrupts, there appears to be a greater amount in the non-deterministic environment than that of the deterministic one. Interrupts appeared to happen initially for the deterministic environment and then stop thereafter. In contrast, more interrupts appear to occur for all milestones travelled through for the non-deterministic environment, hence the increasing trend seen in Figure 2. More interrupts were observed overall for the distance-optimising agent than the time-optimising agent, which we identify the cause being due to the distance-optimising agent passing through red blocks and the participants recognising and responding to the goal misalignment.

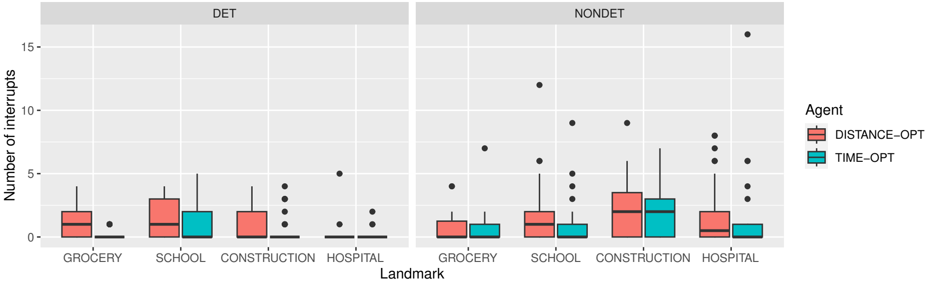

Figure 1. Distribution of the raw interrupt frequencies for each landmark visited. TIME-OPT and DISTANCE-OPT refer to the time-optimising (a_t) and distance-optimising (a_d) agents, respectively.

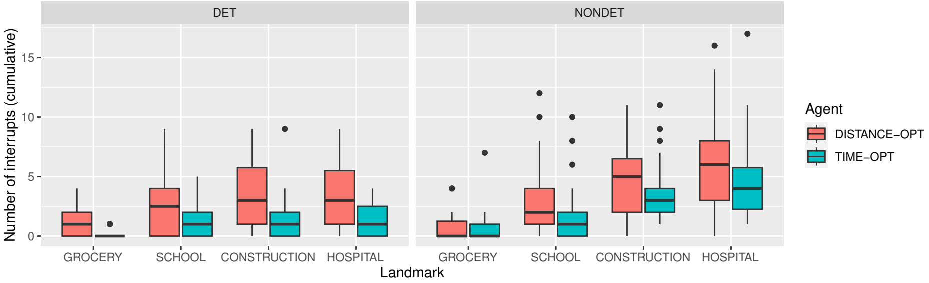

Figure 2. Distribution of the cumulative number of interrupts for each landmark visited. TIME-OPT and DISTANCE-OPT refer to the time-optimising (a_t) and distance-optimising (a_d) agents, respectively.

The median (mean) cumulative number of interrupts observed at the last landmark for the distance- and time-optimising agents, respectively, were 4 and 2 (4.53 and 2.71); the distributions of the cumulative number of interrupts between the groups differed significantly (Wilcoxon rank sum test[3] ($W = 67790$, $p < 0.01$, one-tailed)). The median cumulative number of interrupts overall in the non-deterministic and deterministic settings observed at the last landmark were 5 and 2 (5.29 and 2.41), respectively; the distributions in the two groups also differed significantly (Wilcoxon rank sum test ($W = 60771$, $p < 0.01$, one-tailed)). The distribution of the raw interrupt data throughout the visited landmarks is displayed in Figure 1. The median number of raw interrupts observed overall (across all landmarks) for the distance- and time-optimising agents were 1.22 and 0.75 respectively. The median raw number of interrupts observed overall in the non-deterministic and deterministic settings were 1.38 and 0.67 respectively. The median number of raw interrupts was 1 overall for the non-deterministic setting. The medians for all other settings were zero due to the zero-skewed data.

4.2. RQ2: Relationship Between Confidence Ratings and Intervention Frequency

At each milestone, participants were repeatedly asked to rate their confidence in the agent's ability to help complete the navigation task on a 1 (not at all) – 10 (extremely)

[3] We remark that non-parametric tests were used in our analyses where the data followed a non-normal distribution.

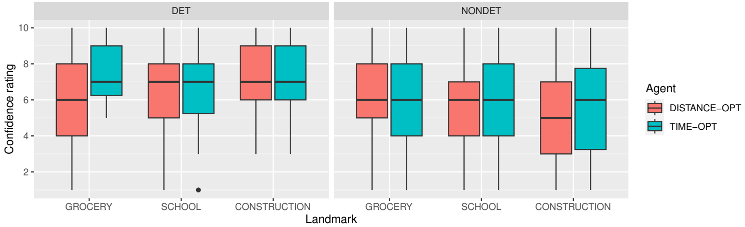

Figure 3. Distribution of confidence ratings for each landmark visited. TIME-OPT and DISTANCE-OPT refer to the time-optimising (a_t) and distance-optimising (a_d) agents respectively.

Likert scale. The agents' planning algorithms imposed a strict ordering of the milestones visited: grocery first, followed by the school, followed by the construction site and finally, the hospital. This allows us to observe the temporal dynamics of interventions. We find that the DET-a_t case has consistently the highest median confidence rating, NONDET-a_d has the lowest ratings, which decrease at a rate faster than NONDET-a_t over time. That for DET-a_d increases over time to the level of DET-a_t. Using a Kruskal-Wallis rank sum test, we found that there was not a statistically significant difference in confidence scores between the visited landmarks overall ($\chi^2 = 0.204$, $p = 0.9$), although there was a statistically significant difference between the number of cumulative number of interrupts between the landmarks ($\chi^2 = 125.36$, $p < 0.01$). A summary of the confidence ratings for each agent-environment setting is illustrated in Figure 3.

When looking at the confidence ratings overall, however, we are able to find that the median confidence rating for a_d and a_t respectively were 6 and 7, suggesting that **participants were less confident in the distance-optimising agent**; the distributions in the two groups differed significantly (Wilcoxon rank sum test ($W = 32229$, $p = 0.011$, one-tailed)). The median confidence rating in the non-deterministic and deterministic settings were also 6 and 7 respectively, suggesting that **participants were less confident in agents operating in the non-deterministic setting**; the distributions in the two groups also differed significantly (Wilcoxon rank sum test ($W = 26138$, $p < 0.01$, one-tailed)). From this, we can deduce that overall, a **higher number of interventions is linked to lower confidence in an agent.**

4.3. Posthoc Evaluation

We administered the validated questionnaire in [30] after the participants completed the navigation task to measure trust qualitatively.

The summary statistics (mean, median and standard deviation) of the four questions overall (O) and for a_d and a_t for the deterministic (D) and non-deterministic (N) settings are displayed in Table 4.3. Here we see a similar pattern in the post-hoc trust scores as we saw previously in the confidence ratings collected during the task. The median posthoc trust score for a_d and a_t respectively were 5 and 7, suggesting that **participants had less trust in the distance-optimising agent**; the distribution of scores between a_d and a_t differed significantly (Wilcoxon rank sum test ($W = 3445.5$, $p < 0.01$, one-tailed). The median trust score rating in the non-deterministic and deterministic setting were 4 and 7 respectively (highlighted in bold in Table 4.3, suggesting that **participants had less trust in goal misaligned agents operating in the non-deterministic setting**; the distribution of scores between the non-deterministic and deterministic settings differed signif-

Setting	**Question 1**			**Question 2**			**Question 3**			**Question 4**		
D-a_d	6.92	7.00	1.84	6.65	7.00	2.00	6.04	7.00	2.27	6.04	6.00	2.02
D-a_t	7.52	8.00	1.72	7.62	8.00	1.79	6.98	7.00	1.81	7.16	7.50	1.79
D-O	7.22	8.00	1.79	7.14	7.00	1.95	6.52	7.00	2.09	6.61	**7.00**	1.98
ND-a_d	4.76	4.00	2.04	4.37	4.00	2.40	4.11	4.00	2.36	4.11	5.00	2.24
ND-a_t	5.27	5.00	2.09	5.06	5.00	2.32	4.43	5.00	2.34	4.64	5.00	2.45
ND-O	5.01	5.00	2.07	4.71	5.00	2.36	4.27	4.00	2.34	4.37	**4.00**	2.35

Table 1. Summary statistics for post-hoc questionnaire responses overall (O), for distance-optimising (a_d) and time-optimising (a_t) agents under deterministic (D) and non-deterministic (ND) settings. For each question–setting combination, the mean, median and standard deviations are reported.

icantly (Wilcoxon rank sum test ($W = 2101$, $p < 0.01$, one-tailed). Through a Kruskal-Wallis rank sum test, there was a statistically significant difference (at the 5% level only) between the different posthoc survey questions ($\chi^2 = 10.90$, $p = 0.012$), confirming the validity of the Trust in Automation questionnaire [30] we used for this experiment.

5. Discussion

The quantitative results show that during the task the agent's capability and the uncertainty in the environment had an impact on intervention frequency and confidence. Specifically, participants intervened more frequently for the distance-optimising agent, and when there was uncertainty in the environment. When confidence ratings were looked at, participants had a pattern of lower confidence in the distance-optimising agent compared to the time-optimising agent, and lower confidence when there is uncertainty in the environment compared to when there is a lack thereof. The self-reported trust, measured after the task, was higher on average in the deterministic setting for both agents than in the non-deterministic setting. The time-optimising agent was associated with higher trust scores on average compared to the distance-optimising agent, within the deterministic and non-deterministic settings. Our observations support the hypothesis that humans would intervene more, have less confidence in, and consequently have less trust in an agent with goal misalignment when the misalignment can be quickly recognised from its behaviour during the task (e.g., distance optimising agent travelling through red zones). The consequence of goal misalignment when working with the distance-optimising agent was less severe compared to the time-optimising agent, which may also explain the inclination to interrupt more. Further, the results show evidence that intervention frequency (when used as a behavioural measurement for trust) is indicative of self-reported trust, thus agreeing with Muir et al. findings [30].

Since we collected data for each milestone visited in a specific order, we were also able to observe temporal trends in the number of interventions. The changes in the number of interventions between different landmarks were statistically significant.

There appeared to be a sharper increase in the number of interrupts over time for the non-deterministic environment compared to that for the deterministic environment. In other words, participants tend to interrupt more as the task progresses in the non-deterministic environment. We note that our data is limited in the sense that we only have four temporal points for interrupts and three for confidence ratings, so we do not perform a correlation test here.

Working with agents programmed by others, who may not necessarily take the present user's objectives in mind, leads to goal misalignment, which impacts interventions. We see that more uncertainty in the environment further exacerbates the problem, impacting confidence in the agent and trust. Combining the qualitative and quantitative data, we see a more detailed picture of trust dynamics in HAI. This supports Kohn et al. [19] recommendation to have both kinds of measurements in trust experiments.

Our results suggest that the goal misalignment in human-agent interaction scenarios negatively impacts trust. We observe that being aware of the goal misalignment at the start of the interaction did not maintain trust. This finding implies that such situations require the agent/automation to be built with mechanisms to build and foster trust continuously throughout the interaction, such as providing explanations or utilising design characteristics that signal trustworthy perceptions.

Further, Kohn et al. [19] suggest that behavioural metrics are affected by external aspects such as workload and risk, motivating the need to confirm that behavioural trust measures correlate with other trust measures. Using validated self-reported measures is one solution. We measured self-reported trust only at the end of the task. We can measure trust between the milestones and wish to perform this comparison in the future. Considering the limitations of our study, it would be ideal to have a finer-grained notion of the temporal dynamics of trust, as we only consider four milestones during the task and collect intervention frequencies and confidence scores. There may be human factors-related causes for the observed trends, which cannot be explained with our data.

6. Conclusion & Future Work

In this work, we examined what interventions reveal as a behavioural measure of trust in an agent-assisted collaborative environment. We show that the agent's capability and the uncertainty in the environment had an impact on intervention frequency and the human's confidence in the agent during the task. We also found that the goal misalignment between a human and an agent impacts trust. Similarly, trust decreases even further in non-deterministic settings. In real-life settings where it is infeasible for the human to avoid uncertainty or goal misalignment, the agent/automation needs to be embedded with features or capabilities that reinforce trust. Our study assumed that the human's need are to optimise for distance and time, while the agent can optimise only for one, which remains static throughout the interaction. In the future, we aim to explore how trust evolves when the agent can adapt to human preferences. We also wish to expand the analysis into other human factors, such as self-confidence and affinity for technology, which have been shown to impact trust in automation [50].

Acknowledgements

This research was funded by the UK Research and Innovation Trustworthy Autonomous Systems Hub (EP/V026801/2). We thank Rita Borgo and Luca Viganò for their guidance in designing the study and the anonymous reviewers for their feedback.

References

[1] Russell S. Human-compatible artificial intelligence. Human-like machine intelligence. 2021:3-23.

[2] Tenhundfeld NL, de Visser EJ, Ries AJ, Finomore VS, Tossell CC. Trust and Distrust of Automated Parking in a Tesla Model X. Human Factors. 2020;62(2):194-210.

[3] Li H, Zhao H, Li C, Wang Q, Zhao X. Takeover behavior patterns for autonomous driving in crash scenarios. Journal of Transportation Safety & Security. 2023;15(11):1087-115.

[4] Brandenburg S, Roche F. Behavioral changes to repeated takeovers in automated driving: The drivers' ability to transfer knowledge and the effects of takeover request process. Transportation Research Part F: Traffic Psychology and Behaviour. 2020;73:15-28.

[5] Compton NB. Tourists follow GPS, drive car into Hawaii harbor; 2023. Accessed: 2024-04-12. https://www.washingtonpost.com/travel/2023/05/02/hawaii-tourists-car-sink-harbor/.

[6] Sommerville I, Cliff D, Calinescu R, Keen J, Kelly T, Kwiatkowska M, et al. Large-scale complex IT systems. Communications of the ACM. 2012;55(7):71-7.

[7] Chancey ET, Bliss JP, Yamani Y, Handley HA. Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. Human factors. 2017;59(3):333-45.

[8] Natarajan M, Gombolay M. Effects of anthropomorphism and accountability on trust in human robot interaction. In: Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction. New York, NY, USA: Association for Computing Machinery; 2020. p. 33-42.

[9] Karli UB, Cao S, Huang CM. "What If It Is Wrong": Effects of Power Dynamics and Trust Repair Strategy on Trust and Compliance in HRI. In: Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. HRI '23. New York, NY, USA: Association for Computing Machinery; 2023. p. 271–280.

[10] Muir BM. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics. 1994;37(11):1905-22.

[11] Daronnat S, Azzopardi L, Halvey M, Dubiel M. Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration. Frontiers in Robotics and AI. 2021;8.

[12] Schadenberg BR, Reidsma D, Heylen DKJ, Evers V. "I See What You Did There"': Understanding People's Social Perception of a Robot and Its Predictability. ACM Transactions on Human-Robot Interaction. 2021 jul;10(3). Available from: https://doi.org/10.1145/3461534.

[13] Rodriguez SS, Zaroukian E, Hoye J, Asher DE. Mediating Agent Reliability with Human Trust, Situation Awareness, and Performance in Autonomously-Collaborative Human-Agent Teams. Journal of Cognitive Engineering and Decision Making. 2023;17(1):3-25.

[14] Wright JL, Chen JYC, Lakhmani SG. Agent Transparency and Reliability in Human–Robot Interaction: The Influence on User Confidence and Perceived Reliability. IEEE Transactions on Human-Machine Systems. 2020;50(3):254-63.

[15] Mayer RC, Davis JH, Schoorman FD. An Integrative Model of Organizational Trust. The Academy of Management Review. 1995;20(3):709-34.

[16] Schaefer KE, Chen JYC, Szalma JL, Hancock PA. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. Human Factors. 2016;58(3):377-400. PMID: 27005902.

[17] Hancock PA, Billings DR, Schaefer KE, Chen JYC, de Visser EJ, Parasuraman R. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors. 2011;53(5):517-27. PMID: 22046724.

[18] Biros DP, Daly M, Gunsch G. The Influence of Task Load and Automation Trust on Deception Detection. Group Decision and Negotiation. 2004 March;13:173-89.

[19] Kohn SC, de Visser EJ, Wiese E, Lee YC, Shaw TH. Measurement of Trust in Automation: A Narrative Review and Reference Guide. Frontiers in Psychology. 2021;12.

[20] McKenna PE, Romeo M, Pimentel J, Diab M, Moujahid M, Hastie H, et al. Theory of Mind and Trust in Human-Robot Navigation. In: Proceedings of the First International Symposium on Trustworthy Autonomous Systems. TAS '23. New York, NY, USA: Association for Computing Machinery; 2023. p. 1-5.

[21] Rossi A, Garcia F, Maya AC, Dautenhahn K, Koay KL, Walters ML, et al. Investigating the Effects of Social Interactive Behaviours of a Robot on People's Trust During a Navigation Task. In: Althoefer K, Konstantinova J, Zhang K, editors. Towards Autonomous Robotic Systems. Cham: Springer International Publishing; 2019. p. 349-61.

[22] Mason E, Nagabandi A, Steinfeld A, Bruggeman C. Trust during robot-assisted navigation. In: 2013 AAAI Spring Symposium Series. AAAI Press; 2013. p. 54-9.

[23] Tokushige H, Narumi T, Ono S, Fuwamoto Y, Tanikawa T, Hirose M. Trust Lengthens Decision Time on Unexpected Recommendations in Human-agent Interaction. In: Proceedings of the 5th International Conference on Human Agent Interaction. HAI '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 245-52.

[24] Perelman B, Evans A, Schaefer K. Mental Model Consensus and Shifts During Navigation System-Assisted Route Planning. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2017 September;61:1183-7.

[25] Bratman ME. Shared Cooperative Activity. The Philosophical Review. 1992;101(2):327-41.

[26] Fan X, Oh S, McNeese M, Yen J, Cuevas H, Strater L, et al. The influence of agent reliability on trust in human-agent collaboration. In: Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction. ECCE '08. New York, NY, USA: Association for Computing Machinery; 2008. p. 1-8.

[27] Lee JD, See KA. Trust in Automation: Designing for Appropriate Reliance. Human Factors. 2004;46(1):50-80.

[28] Ye S, Neville G, Schrum M, Gombolay M, Chernova S, Howard A. Human Trust After Robot Mistakes: Study of the Effects of Different Forms of Robot Communication. In: 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN); 2019. p. 1-7.

[29] Fox JE, Boehm-Davis DA. Effects of Age and Congestion Information Accuracy of Advanced Traveler Information Systems on User Trust and Compliance. Transportation Research Record. 1998;1621(1):43-9.

[30] Muir BM, Moray N. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics. 1996;39(3):429-60.

[31] Fischer K, Weigelin HM, Bodenhagen L. Increasing trust in human–robot medical interactions: effects of transparency and adaptability. Paladyn, Journal of Behavioral Robotics. 2018;9(1):95-109.

[32] Miele D, Ferraro J, Mouloua M. Driver Confidence and Level of Automation Influencing Trust in Automated Driving Features. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2021;65(1):1312-6.

[33] Smith MJ, Desjardins M. Learning to trust in the competence and commitment of agents. Autonomous Agents and Multi-Agent Systems. 2009;18:36-82.

[34] Li M, Lee JD. Modeling Goal Alignment in Human-AI Teaming: A Dynamic Game Theory Approach. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2022;66(1):1538-42.

[35] Sanneman L, Shah JA. Validating metrics for reward alignment in human-autonomy teaming. Computers in Human Behavior. 2023;146:107809.

[36] Hoff KA, Bashir M. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Human Factors. 2015;57(3):407-34.

[37] Lewis M, Sycara K, Walker P. 8. In: Abbass HA, Scholz J, Reid DJ, editors. The Role of Trust in Human-Robot Interaction. Cham: Springer International Publishing; 2018. p. 135-59.

[38] Perkins L, Miller JE, Hashemi A, Burns G. Designing for Human-Centered Systems: Situational Risk as a Factor of Trust in Automation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2010;54(25):2130-4.

[39] Hoesterey S, Onnasch L. The effect of risk on trust attitude and trust behavior in interaction with information and decision automation. Cognition, Technology & Work. 2023;25(1):15-29.

[40] Yang XJ, Unhelkar VV, Li K, Shah JA. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. HRI '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 408–416.

[41] Kaplan AD, Kessler TT, Sanders TL, Cruit J, Brill JC, Hancock PA. Chapter 6 - A time to trust: Trust as a function of time in human-robot interaction. In: Nam CS, Lyons JB, editors. Trust in Human-Robot Interaction. Academic Press; 2021. p. 143 -157.

[42] McBride SE, Rogers WA, Fisk AD. Understanding the effect of workload on automation use for younger and older adults. Human factors. 2011;53(6):672-86.

[43] Clare AS, Cummings ML, Repenning NP. Influencing Trust for Human–Automation Collaborative Scheduling of Multiple Unmanned Vehicles. Human Factors. 2015;57(7):1208-18.

[44] Jiun-Yin Jian AMB, Drury CG. Foundations for an Empirically Determined Scale of Trust in Automated

Systems. International Journal of Cognitive Ergonomics. 2000;4(1):53-71.

[45] Helmert M, Domshlak C. Landmarks, Critical Paths and Abstractions: What's the Difference Anyway? Proceedings of the International Conference on Automated Planning and Scheduling. 2009;19(1):162-9.

[46] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.

[47] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, et al. Openai gym. arXiv preprint arXiv:160601540. 2016.

[48] Charness G, Gneezy U, Kuhn MA. Experimental methods: Between-subject and within-subject design. Journal of economic behavior & organization. 2012;81(1):1-8.

[49] Palan S, Schitter C. Prolific.ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance. 2018;17:22-7.

[50] De Vries P, Midden C, Bouwhuis D. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. International Journal of Human-Computer Studies. 2003;58(6):719-35.