Impact of Agent Reliability and Predictability on Trust in Real Time Human-Agent Collaboration

Sylvain Daronnat
University of Strathclyde, Glasgow
sylvain.daronnat@strath.ac.uk

Leif Azzopardi
University of Strathclyde, Glasgow
leif.azzopardi@strath.ac.uk

Martin Halvey
University of Strathclyde, Glasgow
martin.halvey@strath.ac.uk

Mateusz Dubiel
University of Strathclyde, Glasgow
mateusz.dubiel@strath.ac.uk

ABSTRACT

Trust is a prerequisite for effective human-agent collaboration. While past work has studied how trust relates to an agent's reliability, it has been mainly carried out in turn based scenarios, rather than during real-time ones. Previous research identified the performance of an agent as a key factor influencing trust. In this work, we posit that an agent's predictability also plays an important role in the trust relationship, which may be observed based on users' interactions. We designed a 2x2 within-groups experiment with two baseline conditions: (1) no agent (users' individual performance), and (2) near-flawless agent (upper bound). Participants took part in an interactive aiming task where they had to collaborate with different agents that varied in terms of their predictability, and were controlled in terms of their performance. Our results show that agents whose behaviours are easier to predict have a more positive impact on task performance, reliance and trust while reducing cognitive workload. In addition, we modelled the human-agent trust relationship and demonstrated that it is possible to reliably predict users' trust ratings using real-time interaction data. This work seeks to pave the way for the development of trust-aware agents capable of adapting and responding more appropriately to users.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Collaborative interaction**; *User studies*; *Laboratory experiments.*

KEYWORDS

human-virtual agent interaction; HAI experimental methods; gaming and serious games

ACM Reference Format:
Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2020. Impact of Agent Reliability and Predictability on Trust in Real Time Human-Agent Collaboration. In *Proceedings of the 8th International Conference on Human-Agent Interaction (HAI '20), November 10–13, 2020, Virtual Event, NSW, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3406499.3415063

1 INTRODUCTION

With advances in Artificial Intelligence, more and more intelligent agents are being deployed to aid human operators in completing tasks more effectively and efficiently [4]. Human-Agent Collaboration (HAC) often requires users to validate or invalidate agents' decisions in Command and Control (C2) scenarios, such as X-ray luggage screening [3, 30], collaborative bomb disposal robots [8] or intensive care monitoring agents [16]. In these safety critical scenarios, collaborative agents facilitate the completion of tasks by aiding in the decision-making process [11].

Recently, there has been a renewed focus on developing intelligent collaborative agents able to work with human operators as teammates. In most situations where human-agent collaboration occurs, decisions need to be made in real-time, as interactions between agents and operators are continuous. For instance, rather than having a user validate discrete decisions made by an agent, such as whether to give a patient insulin or not [34], the operator needs to actively work with agents to make decisions, such as monitoring and directing autonomous vehicles [5]. Trust represents an important component of any scenario involving collaborative decision-making, as the perceived trustworthiness of an agent will dictate how an user will interact with it [17, 26, 28, 42].

In past work, it has been shown that an agent's performance (in terms of reliability) as well as an agent's behaviour (in terms of predictability) are positively correlated with trust [9, 33]. However, such studies have largely been conducted in turn-based settings [7, 34] where operators and agents interact asynchronously. Human-agent teams often work together in real-time scenarios where the trust relationship evolves over time and is affected by various factors such as task performance and agents' behaviours [20]. Currently, there is a limited amount of work exploring the relationship between performance, predictability and trust when agents and humans work together in real-time collaborative settings.

1.1 Research Questions and Hypotheses

We ground this study in Human Factor research, where human-agent relationships are analysed to better understand users behaviours. In this work, we explore the relationship between users' perceived trust and reliance on agents who exhibit different levels of predictability and reliability. Specifically, we attempt to address

HAI '20, November 10–13, 2020, Virtual Event, NSW, Australia
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8054-6/20/11…$15.00
https://doi.org/10.1145/3406499.3415063

the following research questions: at the same level of agent's reliability (performance), how do changes in the agent's predictability (behaviour) affect:

(a) the users' reliance on the agent?
(b) the users' workload when interacting with the agent?
(c) the users' perceived trust in the agent?

As previous work has shown that more reliable and more predictable agents tend to be trusted more by users in turn-based settings [27, 33], we hypothesise that, at the same level of agent's reliability (performance), agents exhibiting systematically biased behaviours (i.e. errors committed in a more predictable and consistent fashion) will be trusted more than agents exhibiting randomly varied behaviours (i.e. errors that are unpredictable and committed in an inconsistent way). We further hypothesise that it is possible to use behavioural data from human-agent interactions to model and infer user's perceived trust in agents. The main contribution of our work lies in testing the impact of different degrees of agents' reliability on the human-agent trust relationships in real-time scenarios. We then use interaction data to model and determine how accurately reliance, agents' reliability and performance can predict trust in automation.

2 RELATED WORK

There has been a substantial amount of research on the measurement of trust in automation (see [35] for a thorough review), which has typically been conducted using turn-based scenarios and survey instruments. Less attention, however, has been paid examining the effects of agent's reliability and predictability in real-time human-agent collaborative tasks.

2.1 Trust in automation

While there are many interpretations of trust, we chose to use Lee and See's definition: "*the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability [...] an agent can either be an automated system or another person that actively interacts with the environment on behalf of the person*" [29, p.2]. This definition is of particular relevance as it highlights that trust, as a concept, (*i*) does not differs between team members nor differentiate whether they are human or not, (*ii*) involves collaboration and cooperation between team members, (*iii*) is task dependent, and (*iv*) evolves over time and through interactions. Trust is difficult to measure, monitor [21] and especially hard to assess in a real-time manner, as it is often too disruptive to interrupt and ask users to report trust ratings during the course of an interaction. Measuring and monitoring trust, however, is paramount to the success of human-agent teaming [31]. When trust in agents is too high, users tend to have a more complacent attitude, whereas when trust is too low, users tend to overlook or ignore agents' inputs. Both complacency and distrust are undesirable as they negatively impact task performance [39]. In the context of human-agent interaction, inadequate trust in automated systems can be a factor leading to incidents, such as the ones related to the Boeing 737 MACS system [36]. Thus, a better understanding of trust calibration could help inform the design of future interactive systems [23].

Through repeated interaction with agents, it has been shown that users' trust evolves depending on the agent's reliability [31]. This process is called *trust calibration* [15]. As trust is dynamic and task dependent, new methods are required to infer or predict a person's trust in an agent, over time, given their interactions, rather than using post-hoc questionnaires to elicit trust. Our paper aims at determining the impact of agents reliability and predictability on trust and performance via interaction data *and* questionnaires, and whether it is possible to use these information to predict trust.

2.2 Performance and Reliability

Performance is often understood as an outcome measure in cognitive tasks [44], while reliance is synonymous of trust and indicates the propensity of a user to take into account an agents' inputs in human-agent collaboration (HAC) scenarios. Past work has shown that an agent's *reliability* and its task *performance* heavily influences users' disposition to trust it [19, 37]. In HAC scenarios, agents are generally introduced to reduce users' cognitive workload, while trying to improve users' situational awareness and overall task performance [10, 13, 25, 41].

Fan et al. [14] tested different levels of agents' variability (using systematic biases) in a turn-based C2 threat assessment task. They found that informing participants of the agent's errors helps users calibrate their trust accordingly, which leads to higher task performance. However, too much information regarding the agent's errors can quickly overload users. In related work, Chavaillaz et al. [3] investigated different levels of agents' reliability on trust, reliance and overall task performance in a turn-based X-ray scanning scenario. Their results showed that, as agents reliability decreased, trust in the agents also decreased. Furthermore, they found that perceived reliability (i.e. how much a person is willing to rely on the agents' inputs) is also affected by the capabilities of the automated system. In their studies, users' perception of the reliability of agents was more accurate when interacting with low performing agents, compared to high performing ones.

In addition to studies focusing on different degrees of reliability, the work of Shirado and Christakis [38] explored turn-based coordination problems and found that error-prone agents (up to 30% loss in accuracy) could actually be beneficial to collaborative performance as it reduces the chances of the user being complacent while interacting with the agent. Given the evidence of past research, it is clear that the performance of an agent (its reliability) as well as how the agent behaves (its predictability) impact trust.

As previously mentioned, most studies in the area of trust in agents have been performed using turn-based scenarios, where the agent provides options that users either accept or reject. These scenarios usually offer users more time to assess a situation and react accordingly. However, agents are being integrated in more complex environment, where decisions have to be made in real-time. It is then increasingly important to study the dynamics of trust relationships in real-time scenarios and whether trust can be predicted given past interactions. In this paper, we focus on exploring how agents' reliability and predictability influence users in terms of trust, reliance and cognitive workload as well as the resulting impact on task performance in a real-time human-agent collaborative scenario.

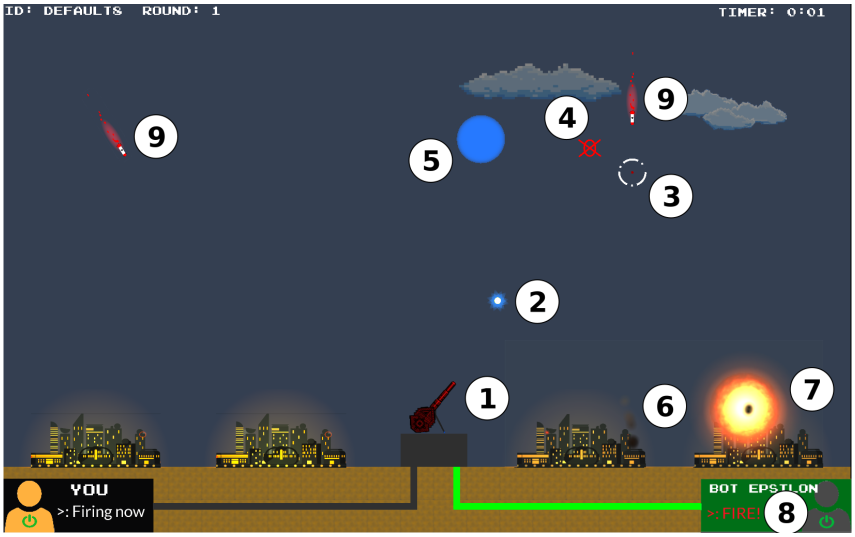

Figure 1: Annotated screen-capture of the missile command scenario.

3 METHOD

To answer our research questions and test our hypotheses, we designed a 2x2 within groups study, where participants undertook a Command and Control (C2) task with agents having different levels of reliability (low and high) and predictability (systematically biased or randomly varied when targeting). We also added two baseline conditions, where users play without any agent and with a flawless (completely reliable) agent. The experiment was undertaken in the context of a collaborative missile command scenario where participants and agents need to work together to defend cities from incoming enemy missiles. Ethics approval for this study was obtained from the University of Strathclyde's Department of Computer and Information Sciences (Approval No. 793).

3.1 Missile Command Scenario

The goal of our real-time interactive task consists in aiming at and destroying missiles appearing from the top of the screen in order to protect cities positioned at the bottom of the screen. To do so, participants can freely move a crosshair across the screen and fire projectiles in the direction of their choosing. In most of the scenarios we designed, participants can collaborate with agents capable of aiding with the aiming process. Agents can help participants by taking care of the aiming process and moving the crosshair automatically. At any moment, however, participants can chose to override the agents' inputs and manually move the crosshair. In all scenarios, only participants can fire projectiles to destroy incoming missiles (this was designed in order to lessen the likelihood of complacent behaviours from the users). Game-based frameworks are often used to study human-agent interactions due to their immersive and easy-to-access nature [1, 43]. Similarly to previous studies on trust [7, 40], this scenario provides a controlled environment where human-agents interactions can be monitored and recorded. Figure 1 shows our interaction scenario in action, where the main elements are numbered and described as follows:

(1) **Gun-turret**: controlled by either the participant or the agent in order to aim and target incoming missiles. All of the projectiles are fired from the turret.
(2) **Projectile**: fired by the participant, it travels at a fixed speed until it explodes in a small circular area. If a missile lies within this area, it is destroyed.
(3) **Crosshair**: provides a visual indication of where the participant or agent is aiming. The crosshair changes colour depending on who is controlling it (yellow for the participant, white for the agent, and dark-grey for neither.)
(4) **Red Indicator Area**: appears when a projectile is fired to show participants the area where the projectile will explode.
(5) **Projectile's explosion/halo**: In order to destroy missiles, they have to enter within the radius of such explosion.
(6) **City**: Assets that the participants are tasked to protect.
(7) **Missile Impact**: when a missile reaches a city, it produces an orange/red explosion with smoke emanating from the city.
(8) **User and Agent panels**: The participant's panel (on the bottom left of the screen) and the agent's panel (on the bottom right) light up in green when one of them is moving the crosshair.
(9) **Enemy missile**: progress at a fixed speed and angle depending on the task difficulty. At the end of a session (with or without an agent), participants are shown how many missiles they hit and/or missed. All missiles missed eventually hit a city.

3.2 Agents: Reliability and Predictability

Participants interacted with five different agents. Each agent performed different types of targeting, which was controlled to create different levels of reliability and predictability. This was achieved by adjusting two variables: (*i*) the systematic bias (biased or not), and (*ii*) the random variance (low and high). Depending on the combination of (i) and (ii), these elements were controlled to result in different levels of performance. Figure 3 shows the different combinations of agents used, which we refer to as: Alpha, Beta, Gamma, and Delta (A,B,C,D). Agents names were introduced to make it easier for participant to refer to any particular agent. *Alpha* and *Beta* were biased with, respectively, low and high variance. Agents *Gamma* and *Delta* were not biased with, respectively, low and high variance.

All agents had a certain amount of variance in their accuracy such that, for a given target, a certain amount of error would be applied to the targeting. The greater the variance, the less accurate the agent's aim, leading to worse task performance (see Figure 2).

In addition to variance, agents Alpha and Beta had their aiming systematically biased in a particular direction: (i) always above and to the right of their target, (ii) always below and to the left, (iii) always above and to the left, (iv) always below and to the right. The direction of the systematic bias was randomly selected at the beginning of the experiment, per participant, and kept constant during the condition. By randomly selecting the direction, we ensured that our findings were not constrained by a specific type of systematic bias. This systematic bias impacted the agents' targeting behaviours, but *not* their performance, which were only impacted by random variance.

Agents' performance was calibrated using simulations where the agents completed the task by themselves. We then ensured that the accuracy of respectively low and high performance agents was not significantly different using t-tests. Agents Beta and Delta were tuned to have high performance (approx. 70% accuracy), while agents Alpha and Gamma were tuned to have low performance (approx. 30% accuracy).

By controlling agents' performance and predictability, we could test our main hypothesis using a 2x2 design. In addition to the aforementioned agents, we also included a *perfect* agent: *Epsilon* which exhibited no bias and no variance – and thus had the highest reliability and predictability out of all of the other agents (effectively serving as an upper bound).

3.3 Rounds & Difficulty

During each interaction with a particular agent, participants went through three rounds which lasted for 90 seconds each. This duration was set so that participants had enough time to familiarise themselves and adapt to the agents, while ensuring that the entirety of the experiment could be completed within an hour (lessening participants' fatigue). Each round increased in difficulty (going through"Easy", "Medium" and "Hard" difficulty levels). In "Easy" difficulty, missiles spawned every 4 seconds at a speed of 100 pixels per second, for "Medium" difficulty, missiles spawned every 2 seconds with a speed of 150 pixels per second, and finally for "Hard" difficulty, missiles spawned every second with a speed of 200 pixels per second. These settings were calibrated during pilot testing with ten participants to make sure that changes in difficulty were noticeable without completely overwhelming participants (see Section 3.4).

3.4 Piloting

Before conducting the study presented in this paper, a formal pilot experiment was created. Ten participants were recruited from our local Computer Science department. This pilot focused on calibrating the single player (no agent) experience as well as core gameplay elements such as the controls, visuals and overall difficulty.

To evaluate participants' performance, we used F1 scores detailed in Section 3.5. F1 scores varied between 0.88 for the "Easy", 0.77 for the "Medium" and 0.46 for the "Hard" difficulty levels. We then decided to increase the speed of missiles in the "Medium" difficulty level to intensify its complexity.

During unstructured post-hoc interview, The radius of the projectiles' explosions was found to be too big, we then decided to reduce it from 60 to 45 pixels. The speed at which participants were able to move the cross-hair was perceived to be too slow, we decided to increase it from 600 to 800 pixels per second. During further informal pilots, participants gave additional feedback on which colours were the most clearly distinguishable when either the users or the agents are taking over the controls. We then chose to associate the agent with yellow and the user with white.

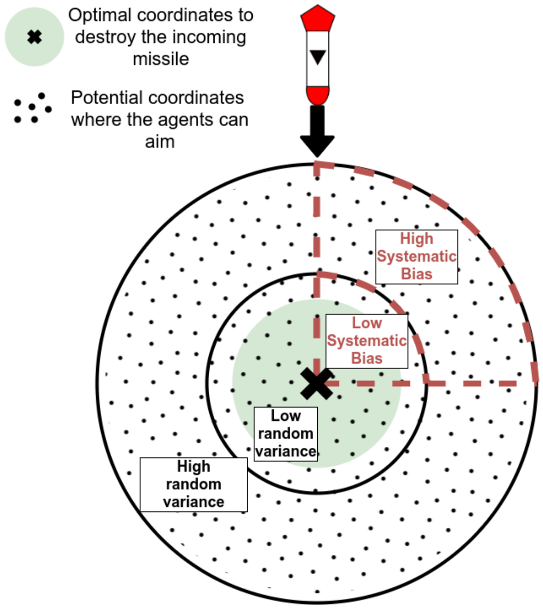
Optimal coordinates to destroy the incoming missile
Potential coordinates where the agents can aim
High Systematic Bias
Low Systematic Bias
Low random variance
High random variance

Figure 2: Visualisation of the different biases applied to the agents in the study (not to scale). The greater the bias, the lower the accuracy of the agent. For the systematic bias, a "quadrant" is randomly chosen for each participant at the beginning of a session. Low syst. bias and low random variance or high syst. bias and high random variance result in the same performance output.

3.5 Interactions and Performance Logging

Participants interactions were logged during each task. Logging included the number of shots fired, missiles destroyed, missiles on screen, amount of time the user controlled the crosshair (in seconds) and the distance the crosshair was moved for. Logging of these elements was completed for all scenarios, with or without agents. Using data collected during these interactions, we then calculated the following task performance measures:

$$\textbf{Precision} = \frac{\#MissilesDestroyed}{\#ShotsFired}$$

$$\textbf{Recall} = \frac{\#MissilesDestroyed}{\#IncomingMissiles}$$

$$\textbf{F1} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Higher precision indicates greater accuracy (fewer attempts to hit a target), while higher recall indicates greater task performance (more cities being protected). F1 is the harmonic mean of precision and recall which provides a combined measure of performance. The user control time was computed as the number of seconds for which participants were controlling the crosshair for each round (a greater user control time suggests less reliance on the agent).

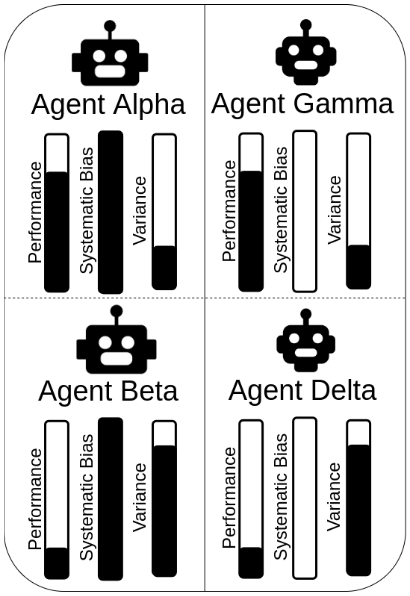

Figure 3: Agents with different degrees of predictability (behaviours) and reliability (performance) were created for this study. Systematic bias and random variance were used to respectively constrain how predictable and accurate the agents' accuracy was.

3.6 Questionnaires

Participants completed NASA TLX rating scales, which are 6 items survey instrument commonly used to measure cognitive workload [18]. In this study, RAW TLX [2] scores are reported. To measure trust in the agents, we used a single statement at the end of each round: "I can trust the agent" graded on a 11 points Likert scales from 1 (complete distrust in the agent) to 11 (total trust in the agent) adapted from the work of Jian et al. [24].

3.7 Dependant and independent variables

The independent variables in this study are:

- **Agent Behaviours**: *Systematic Bias* (bias or no bias) and *Random Variance* (no, low and high variance). These aspect determine Agent Performance (perfect, high and low).
- **Difficulty** per round (Easy, Medium, and Hard).

The dependent variables in this study were:

- **Time in Control**: The time participants and agents spent controlling the crosshair for each round.
- The number of **missiles destroyed**, the number of **projectiles fired** and the total number of **hits sustained by cities**, per round.
- **Distance travelled** by the crosshair when the user or an agent were in control of it.
- **NASA TLX** [32] ratings scales to measure participants' cognitive workload after having played with each agent.
- **Single Trust Question** [12, 24] answered at the end of each round. Higher ratings indicate higher reported trust.

3.8 Experimental Procedure

Participants were briefed on the experiment and asked to provide consent required to undertake the study. After completing a demographic questionnaire, participants were first given a short tutorial on how to play the game and interact with the agents. They were instructed that their goal was to work with the agents to protect cities by destroying all incoming missiles. They were informed that they could always correct the agents' aiming if they desired to do so. Following this briefing, participants completed a session without the assistance of an agent, to record their individual (no agent) performance. Participants then played with all of the other agents. The order in which participants interacted with each agent was randomised using a William Square design in order to mitigate possible learning effects [45]. During each session, participants worked through three rounds of low to high levels of difficulty. At the end of each round, participants were asked to rate their trust in the agents. At the end of each session, participants were asked to complete the NASA Task Load Index (TLX) questionnaire. At the end of the study, which lasted for approximately an hour, participants were compensated for their time with a shopping voucher worth £10.

3.9 Demographics

Participants were recruited through flyers and mailing lists on our local campus. We recruited a total of 30 participants (14M,16F) with ages ranging from 19 to 38 years old ($M = 27 \pm 5.19$). Most participants were enrolled as postgraduate students. Ratings from the Complacency Potential Rating Scale (CPRS) [22] were used to evaluate general attitude toward automation. CPRS scores ranged from 55.57 and 90.84 ($M = 72.55 \pm 9.3$) which denotes a population more likely to rely on automation than not [22]. Overall, the distribution of scores was homogeneous enough that it could *not* be divided in different group representing particular attitude toward automation.

Table 1: Metrics related to performance (Recall, Precision and F1, higher = better) and reliance (User control time (in seconds) higher = less reliance on the agent). Superscript letters next to the results indicate which agents yielded significantly worse scores ($p < 0.05$).

	No Agent	Agent Alpha bias/low var.	Agent Beta bias/high var.	Agent Gamma no bias/low var.	Agent Delta no bias/high var.	Agent Epsilon no bias/no var.
Recall	0.64 ± 0.03^{D}	0.82 ± 0.02^{NDBG}	0.60 ± 0.03	0.72 ± 0.02^{NDB}	0.58 ± 0.03	0.98 ± 0.01^{NDBAG}
Precision	0.57 ± 0.02^{DB}	0.60 ± 0.02^{DBG}	0.50 ± 0.03	0.53 ± 0.02^{D}	0.47 ± 0.02	0.86 ± 0.01^{NDBAG}
F1	0.60 ± 0.02^{DB}	0.68 ± 0.02^{NDBG}	0.54 ± 0.03	0.60 ± 0.02^{DB}	0.51 ± 0.03	0.91 ± 0.01^{NDBAG}
User Ctrl Time	25.12 ± 0.96^{EAG}	5.34 ± 0.83^{E}	24.18 ± 1.16^{EAG}	10.61 ± 1.09^{EA}	27.68 ± 1.29^{BEAG}	1.02 ± 0.43

Table 2: Metrics related to cognitive load and trust ratings. Superscript letters next to the results indicates which agents yielded significantly worse scores ($p < 0.05$).

	Agent Alpha bias/low var.	Agent Beta bias/high var.	Agent Gamma no bias/low var.	Agent Delta no bias/high var.	Agent Epsilon no bias/no var.
Raw TLX	9.64 ± 0.34^{E}	14.62 ± 0.38^{EAG}	11.57 ± 0.30^{EA}	15.47 ± 0.31^{BEAG}	4.79 ± 0.36
Trust Ratings	7.82 ± 0.26^{DBG}	2.16 ± 0.16	6.28 ± 0.28^{DB}	2.17 ± 0.16	10.61 ± 0.13^{DBAG}

4 RESULTS

In this section, we present our main results regarding task performance, users' reliance on agents, users' workload and users' reported trust in the agents. Then, we model and predict trust ratings using aforementioned performance and user behaviour metrics. To compare performance between conditions, we first used ANOVAs (for which we are always reporting p and F values) and then performed follow-up pairwise comparisons using T-tests, if statistically significant results were found ($p < 0.05$). Bonferroni corrections were applied to determine which conditions were significantly different. For T-tests, we always report p-values as well as the effect size using Cohen's d values[1]. In Tables 1 and 2, if the score in a given condition was significantly better than in other conditions, we denote it by using superscripts (N for no agent and A,B,G,D and E for each agents).

4.1 Performance

Table 1 shows the average task performance achieved by participants in each condition. These scores are averages over all three levels of difficulty. As expected, participants performed the best with agent Epsilon (no bias or variance) compared to any of the other conditions across all measures. When using Alpha and Gamma, participants were able to achieve higher precision scores than by themselves, but performed worse with Beta and Delta (across Recall, Precision and F1 scores). ANOVA testing yielded significant results for Precision scores ($p < 0.0001$, $F = 3.55$), Recall scores ($p < 0.0001$, $F = 37.47$), and F1 scores ($p = 0.0002$, $F = 9.65$). Follow-up pairwise comparisons showed statistically significant results between Alpha (syst. bias, low variance) and Gamma (no bias, low variance) for Precision $p = 0.0001$, $d = 0.54$, Recall $p < 0.0001$, $d = 0.86$ and F1 $p < 0.0001$, $d = 0.66$ scores.

4.2 Reliance

Table 1 shows the amount of time (in seconds) participants spent in control of the crosshair (denoted as *User Ctrl Time*). If participants controlled the crosshair for a longer period of time, it suggests that they relied on the agents less (and vice versa). As expected, we observed that participants spent less time controlling the crosshair when working with Epsilon compared to any of the other conditions, with or without agents. In addition, participants spent significantly more time controlling the crosshair ($p < 0.0001$) when collaborating with low performance, high variance agents (Beta and Delta) compared to high performance, low variance agents (Alpha and Gamma). ANOVA testing yielded statistically significant results ($p < 0.0001$, $F = 22.70$) when comparing overall user control time, but follow up pair-wise comparisons showed that these differences were only significant between Alpha (biased, low variance) and Gamma (not biased, low variance) with $p < 0.0001$ and a large effect size $d = 0.81$.

4.3 Cognitive Load

We observe on Table 2 that participants reported much lower cognitive load (NASA TLX scores) when interacting with agent Epsilon (no bias or variance) compared to any of the other agents. In addition, participants reported much higher cognitive load when interacting with low performance agents (Beta and Delta) compared to high performance ones (Alpha and Gamma). When comparing overall Raw Nasa TLX scores, an ANOVA yielded significant results ($F = 8.73$, $p = 0.006$). While performing pairwise comparisons, we found that participants perceived the low systematic bias agent (Alpha) as significantly less cognitively taxing than the high variance agent (Gamma) with $p = 0.0061$, $d = 0.623$. In addition, participants found the agent with high variance (Delta) as being significantly more cognitively taxing than the agent with high systematic bias (Beta) with $p = 0.0473$, $d = 0.26$.

[1] Note that $0.5 < d < 0.8$ is considered a medium effect size, whereas $d > 0.8$ is a high effect size [6]

Table 3: Linear regression results when predicting participants trust ratings from using contextual (difficulty) and behavioural measures (performance and reliance). Only the most important results are presented. A higher R^2 value indicates more accurate predictions.

Parameters	Mean Square Error	Adjusted R^2
User Ctrl Time + Precision + Recall + F1 + Difficulty	3890.0	0.893
User Ctrl Time + Precision + Recall + F1	4717.9	0.867
User Ctrl Time + Precision + Recall	6222.2	0.858
Recall	17253.2	0.793
Raw TLX	7796.8	0.357
User Ctrl Time	1830.2	0.082

Table 4: Spearmans correlation tests between participants' behavioural metrics (performance and reliance) and reported trust ratings. A higher ρ scores indicates greater correlation.

Parameter 1	Parameter 2	ρ	p-value
User Ctrl Time	Trust Ratings	0.801	<0.001
Raw TLX	Trust Ratings	0.730	<0.001
Recall	Trust Ratings	0.614	<0.001
F1	Trust Ratings	0.552	<0.001
Precision	Trust Ratings	0.501	<0.001
Difficulty	Trust Ratings	0.012	0.790

4.4 Trust

By inspecting Table 2, it is clear that, on average, participants trusted agent Epsilon (no bias or variance) more than any of the other agents, which was expected. In addition, trust ratings of agents with high variance (Beta and Delta) were on average much lower than agents with low variance (Alpha and Gamma). When comparing answers pertaining to the trustworthiness of agents, an ANOVA yielded significant results ($F = 7.80$, $p = 0.0018$). While performing pairwise comparisons, we found that participants rated Alpha (syst. bias, low variance) significantly higher than Gamma (no syst. bias, low variance) with $p = 0.0011$, $d = 0.86$. Overall, no significant results were found when comparing Beta (syst. bias, low variance) to Delta (no bias, high variance). These results indicate that, at the same high level of agents' performance, participants were more trustful of a systematically biased agent (Alpha) than an agent with random variance only (Gamma).

4.5 Predicting Trust

To examine how different variables influence trust, we analysed correlations between Trust Ratings, task difficulty, reliance metric (User Control Times), cognitive workload (NASA TLX scores) and performance metrics (Precision, Recall and F1 scores). Table 4 reports the Spearmans' ρ (and the p-value), where we can see that participants reliance on the agents (as measured by User Control Time) led to the highest correlation ($\rho = 0.801$), whereas the performance metrics (Recall, F1 and Precision) ranged from 0.5 to 0.61. In addition, we created multiple linear regression models to determine which combinations of factors led to the best predictions of users' trust ratings. Table 3 shows the combination of factors, mean square error, and adjusted correlation coefficients for each models. Our results show that the best performance for predicting trust ratings ($R^2 = 0.893$) were achieved by combining measures related to reliance (user control time), performance (the number of shots fired, missiles destroyed and misses) and task complexity.

5 DISCUSSION

In this paper, we have explored how agents' predictability and reliability influence users' perception of agents in terms of cognitive workload and trust, as well as the resulting effects on task performance. As expected, we found that interacting with a near perfect agent (agent Epsilon) led participants to achieve higher performance while having an overall more positive outlook of the agent. When comparing the rest of the agents, however, clear differences in users' behaviours and perceptions were found.

With our first research question (see Section 1.1), we set out to explore how agents' predictability impacts reliance, workload and trust. When comparing the agents with low performance and systematic bias (Beta) to the agent with low performance and no bias (Delta), we noticed that both yielded poor overall task performance, even worse than when participants did not interact with any agent at all. These worst results were found across all performance indicators: F1, Recall and Precision. In addition, participants had to compensate more for the agents' inaccuracy, as is evidenced by higher user control times, greater reported workload and lower trust ratings. Nevertheless, when comparing agent Beta (syst. bias and high variance) to agent Delta (no bias and high variance), we found that participants performed slightly better with agent Beta, in addition to spending slightly less time correcting the agent and reporting significantly lower cognitive workloads. This suggests that when an agent's behaviour is more predictable by making errors in a systematic way, participants are able to compensate for its inaccuracy better.

When comparing agent Alpha (systematic bias, low variance) to agent Gamma (no bias, low variance), we found that participants achieved significantly higher performance with Alpha. They also corrected agent Alpha significantly less and reported significantly lower workload. These results further suggest that when an agents behaviour is more predictable, participants could not only better compensate for the agents' imprecision, but also *adapt* and *work* with the agent better, resulting in an overall better task performance.

Overall these findings suggest that, in the case of imperfect automation, systematic biases are preferable to random variance. When compared to randomly biased agents, at the same level of agents' performance, systematic biases allow users to adapt better and quicker to an agent's behaviour, which results in a higher reported trust in the agent, better task performance and reduced cognitive load.

We further hypothesised that it is possible to infer trust in an agent using information collected during human-agent interactions. To investigate this area, we first sought to determine which factors were the most important to predict participants' perceived trust in agents. Previous work hypothesised that performance is the most important predictive factor of users trust in agents [20]. Table 4 shows correlations between trust ratings and some variables monitored in our study. This shows that our different performance indicators (such as F1, Recall and Precision) are moderately correlated with trust ratings. In addition, our findings reveal that reliance, expressed in our study by the amount of time users spent correcting the agents, was correlated more strongly than performance to users' reported trust in the agents. To further study which combinations of factors could predict trust ratings best, we performed several multi-linear regressions. We achieved the best results (see Table 3 by using data related to users' reliance on the agents, performance scores and the difficultly of the task. These findings suggest that it is important to consider both performance and reliance metrics in order to infer users' trust in an agent more effectively. Moreover, we showed that it is possible to predict users trust ratings with a very high correlation.

Our study is a step forward for the development of trust-aware agents capable of using real-time interactions data to adapt to the users. However, additional tests on the variables that influence trust the most in human-agent interactions should be conducted in different contexts in order to further verify what components are the most important for the building and maintaining of the human-agent trust relationship. While in this work we only considered user control time as a measure of reliance, other behavioural measures could be included, such as the number of corrections issued by users, or the amount of time users spent monitoring the agents actions while not directly correcting them. Such measures could be used to further enhance the real-time prediction of trust in agents. The main benefit of being able to monitor this trust relationship in real-time resides in the ability to continuously monitor trusts relationships based on interactions, without needing to interrupt human operators.

6 LIMITATIONS

It should be noted that our study is not without limitations. We have only explored how predictability and reliability influence trust in one kind of interactive scenario. In addition, even if initial pilots guided the design of the study, our framework is new and further work is needed to explore how our findings generalise to other real-time collaborative settings, and on other populations with different attitudes toward automation. In order to ensure the experiment could be completed within an hour, a number of constraints restricted how many agents and for how long interactions with the agents lasted. It is possible that more time spent working with the agents would help participants better calibrate their trust over time. Inversely, interactions that are too lengthy could lead to complacency or complete distrust. In our study, however, standard deviation of trust ratings between participants was very low, which indicates that the impact of our different agents on participants was fairly consistent throughout the experiment. Furthermore, while we controlled for performance and agents behaviours, we only tested four combinations. With more agents, different levels of performance and different amounts of predictability could have been explored to see how participants' perceptions of agents transitions from high to low trust, and less to more reliance. We would like to note that these limitations do not undermine the main findings of our study, but we acknowledge that additional investigations are required to understand more precisely the relationship between the different variables linked to trust in agents, as well as how other types of tasks influence this relationship. We leave these directions for future work.

7 CONCLUSION

In this study, we set out to explore the relationship between trust, agents' predictability and agents' reliability in a real-time collaborative scenario. To achieve this, we designed a within-groups study where participants completed an aiming task with the help of different collaborative agents. We found that, at the same level of performance, participants reported higher levels of trust in agents that were more predictable than less predictable agents. However, as the agents' reliability decreased, participants were less trustful of the agents, regardless of their predictability. In addition, participants achieved better performance and reported lower cognitive load with systematically biased agents compared to agents with more variance, especially at a high level of agents' performance. These findings further highlight the importance of predictability and consistency in the design of potentially error-prone agents, and how it impacts human-agent collaboration in real-time. Furthermore, our study investigated whether it was possible to infer trust ratings based on participants' interactions. Our findings show that while performance indicators are important, in the context of real-time collaboration, participants' reliance on agents is a better predictor of trust. These findings suggest that the development of methods that can monitor trust in automation over time is possible, and could be used by agents to better adapt to users. For instance, if under-relying on an agent leads to degrading performance, "trust repair mechanisms" could be deployed to improve trust and reliance in automation and hopefully lead to an increase in overall task performance. With this work, we further our understanding of how agents' behaviours are linked to trust, and which components influence the evolution of trust the most in real-time collaborative scenarios.

ACKNOWLEDGMENTS

We thank the Thales company for their on-going support in the project.

REFERENCES

[1] S. Almajdalawi, V. Pavlinek, M. Mrlik, Q. Cheng, and M. Sedlacik. 2016. The Influence of Virtual Agents on Player Experience and Performance. *CHI PLAY*

(2016). https://doi.org/10.1088/1742-6596/412/1/012003

[2] Alex Cao, Keshav K. Chintamani, Abhilash K. Pandya, and R. Darin Ellis. 2009. NASA TLX: Software for assessing subjective mental workload. *Behavior Research Methods* 41, 1 (feb 2009), 113–117.

[3] Alain Chavaillaz, Adrian Schwaninger, Stefan Michel, and Juergen Sauer. 2018. Automation in visual inspection tasks: X-ray luggage screening supported by a system of direct, indirect or adaptable cueing with low and high system reliability. *Ergonomics* 61, 10 (Oct. 2018), 1395–1408.

[4] Jessie Y. C. Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science* 19, 3 (feb 2018), 259–282.

[5] Ting Chen, Duncan Campbell, Luis Felipe Gonzalez, and Gilles Coppin. 2015. Increasing Autonomy Transparency through capability communication in multiple heterogeneous UAV management. In *IEEE International Conference on Intelligent Robots and Systems*. https://doi.org/10.1109/IROS.2015.7353707

[6] J Cohen. 1988. Statistical Power Analysis for the Behavioral Sciences. In *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition*. Routledge.

[7] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. 2018. Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 507–513.

[8] S Costo and R Molfino. 2004. A new robotic unit for onboard airplanes bomb disposal. In *35th International symposium on robotics ISR*, Vol. 2004. Citeseer, 23–26.

[9] Ewart J. de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman. 2012. The World is not Enough: Trust in Cognitive Agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1 (sep 2012), 263–267.

[10] Mustafa Demir, Nathan J. McNeese, and Nancy J. Cooke. 2017. Team situation awareness within the context of human-autonomy teaming. *Cognitive Systems Research* 46 (2017), 3–12. https://doi.org/10.1016/j.cogsys.2016.11.003

[11] Katharina Emmerich, Patrizia Ring, and Maic Masuch. 2018. I'm Glad You Are on My Side: How to Design Compelling Game Companions. *The Annual Symposium on Computer-Human Interaction in Play Extended Abstracts - CHI PLAY '18* (2018), 141–152.

[12] Elliot E. Entin and Daniel Serfaty. 2017. Sequential Revision of Belief, Trust Type, and the Order Effect. *Human Factors* 59, 3 (2017), 407–419. https://doi.org/10.1177/0018720816678322

[13] Xiaocong Fan, Michael McNeese, Bingjun Sun, Timothy Hanratty, Laurel Allender, and John Yen. 2010. Human-Agent Collaboration for Time-Stressed Multicontext Decision Making. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 40, 2 (mar 2010), 306–320.

[14] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R. Endsley. 2008. 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* (2008). ACM, 7.

[15] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems, CTS* (2007), 106–114.

[16] Behnood Gholami, Wassim M. Haddad, and James M. Bailey. 2018. AI in the ICU: In the intensive care unit, artificial intelligence can keep watch. *IEEE Spectrum* 55, 10 (oct 2018), 31–35. https://doi.org/10.1109/MSPEC.2018.8482421

[17] F. S. Grodzinsky, K. W. Miller, and M. J. Wolf. 2011. Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?". *Ethics and Information Technology* 13, 1 (2011), 17–27. https://doi.org/10.1007/s10676-010-9255-1

[18] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[19] Jean-Michel Hoc, Mark S. Young, and Jean-Marc Blosseville. 2009. Cooperation between drivers and automation: implications for safety. 10, 2 (2009), 135–160.

[20] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570

[21] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink. 2013. Trust in Automation. 28, 1 (2013), 84–88.

[22] Indrarnani L. Singh, Robert Molloy and Raja Parasuraman. 1993. Automation-Induced "Complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology* 3, October 2013 (1993), 37–41. https://doi.org/10.1207/s15327108ijap0302

[23] Theodore Jensen, Mohammad Maifi Hasan Khan, Yusuf Albayram, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. 2020. Anticipated Emotions in Initial Trust Evaluations of a Drone System Based on Performance and Process Information. *International Journal of Human–Computer Interaction* 36, 4 (2020), 316–325.

[24] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (mar 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[25] Daisuke Karikawa, Hisae Aoyama, Makoto Takahashi, Kazuo Furuta, Toshio Wakabayashi, and Masaharu Kitamura. 2013. A visualization tool of en route air traffic control tasks for describing controller's proactive management of traffic situations. 15, 2 (2013), 207–218.

[26] Da-jung Kim and Youn-kyung Lim. 2019. Co-Performing Agent. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–14. https://doi.org/10.1145/3290605.3300714

[27] G. Klien, David D. Woods, Jeffrey M. Bradshaw, Robert R. Hoffman, and Paul J. Feltovich. 2004. Ten challenges for making automation a" team player" in joint human-agent activity. 19, 6 (2004), 91–95.

[28] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2019. Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics* 62, 3 (2019), 345–360. https://doi.org/10.1080/00140139.2018.1547842

[29] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (jan 2004), 50–80.

[30] Stephanie M. Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. I Trust It, but I Don't Know Why. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 55, 3 (jun 2013), 520–534. https://doi.org/10.1177/0018720812465081

[31] Stephanie M. Merritt, Deborah Lee, Jennifer L. Unnerstall, and Kelli Huber. 2015. Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors* 57, 1 (2015), 34–47. https://doi.org/10.1177/0018720814561675

[32] NASA. 1986. TASK LOAD INDEX, Nasa TLX, v1.0. *NASA* (1986).

[33] Sherry Ogreten, Stephanie Lackey, and Denise Nicholson. 2010. Recommended roles for uninhabited team members within mixed-initiative combat teams. In *2010 International Symposium on Collaborative Technologies and Systems*. IEEE.

[34] Richard Pak, Nicole Fink, Margaux Price, Brock Bass, and Lindsay Sturre. 2012. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics* 55, 9 (July 2012), 1059–1072.

[35] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. 2016. A Meta-Analysis of Factors Influencing the Development of Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58, 3 (mar 2016), 377–400.

[36] Tommaso Sgobba. 2019. B-737 MAX and the crash of the regulatory system. *Journal of Space Safety Engineering* 6, 4 (2019), 299–303.

[37] TB Sheridan. 1989. Trustworthiness of command and control systems. In *Analysis, Design and Evaluation of Man–Machine Systems 1988*. Elsevier, 427–431.

[38] Hirokazu Shirado and Nicholas A. Christakis. 2017. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* 545, 7654 (2017), 370–374. https://doi.org/10.1038/nature22332

[39] Indramani L. Singh, Robert Molloy, and Raja Parasuraman. 1993. Automation-Induced "Complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology* 3, 2 (apr 1993), 111–122.

[40] Alessandro Sordoni, Jean-Pierre Briot, Isabelle Alvarez, Eurico Vasconcelos, Marta de Azevedo Irving, and Gustavo Melo. 2010. Design of a Participatory Decision Making Agent Architecture Based on Argumentation and Influence Function - Application to a Serious Game about Biodiversity Conservation. *RAIRO - Operations Research* 44, 4 (2010), 269–283. https://doi.org/10.1051/ro/2010024

[41] Kimberly Stowers, Nicholas Kasdaglis, Michael Rupp, Jessie Chen, Daniel Barber, and Michael Barnes. 2017. *Insights into human-agent teaming: Intelligent agent transparency and uncertainty*. Vol. 499. 149–160 pages. https://doi.org/10.1007/978-3-319-41959-6_13

[42] Trond A. Tjøstheim, Birger Johansson, and Christian Balkenius. 2019. A Computational Model of Trust-, Pupil-, and Motivation Dynamics. In *Proceedings of the 7th International Conference on Human-Agent Interaction* (Kyoto, Japan) *(HAI '19)*. Association for Computing Machinery, New York, NY, USA, 179–185. https://doi.org/10.1145/3349537.3351896

[43] Ning Wang, David V Pynadath, and Susan G Hill. 2015. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. *Aamas* Aamas (2015), 997–1005. http://delivery.acm.org.prx.library.gatech.edu/10.1145/2940000/2937071/p997-wang.pdf?ip=143.215.137.43{&}id=2937071{&}acc=ACTIVESERVICE{&}key=A79D83B43E50B5B8.5E2401E94B5C98E0.4D4702B0C3E38B35.4D4702B0C3E38B35{&}CFID=828239253{&}CFTOKEN=84023669{&}{_}{_}acm{_}{_}=1510434143

[44] Eric N. Wiebe, Allison Lamb, Megan Hardy, and David Sharek. 2014. Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. *Computers in Human Behavior* 32 (2014), 123–132. https://doi.org/10.1016/j.chb.2013.12.001

[45] EJ Williams. 1949. Experimental Designs Balanced for the Estimation of Residual Effects of Treatments. *Australian Journal of Chemistry* 2, 2 (1949), 149.