CHAPTER 5

Towards Understanding the Cognitive Aspects of Transparency in Human-Autonomy Teaming

August 27, 2022

MATT CABANAG AND CHRISTOPHER J STANTON

ABSTRACT

This study explores transparency in a command and control (C2) context, using a low-fidelity air traffic control game, which is real-time, dynamic, and time constrained. Autonomous agent performance, anthropomorphism, and other factors have been a major focus in studying trust in human-autonomy teaming (HAT). We propose that agent predictability may be an important area of investigation. Where autonomy is imperfect, increasing its predictability may reduce the incidence of mistrust and dis- use. Indeed, we suggest that predictability is a quintessential indicator of agent transparency, which we propose to encapsulate in a model of trust that is based on predictability. We speculate that cognitive fit and cognitive fit theory may have a large role to play in enabling predictability. This has implications for transparency design in self driving cars, domestic household robots, as well as other industrial applications where autonomous systems and agents are used.

KEYWORDS:

serious games; command and control; human autonomy teaming; trust in au- tonomous agents; cognitive load; cognitive fit

1 BACKGROUND

Anthropomorphism has long been a focal point in human-autonomy teaming (HAT), with researchers investigating etiquette (Parasuraman & Miller, 2004), apologies (M. C. Cohen, Demir, Chiou, & Cooke, 2021; Galdon & Wang, 2019), and compensation behaviours (De Visser, Pak, & Shaw, 2018; Rebensky et al., 2021). While these studies attempt to investigate if humans attribute human moral qualities to their autonomous team-mates, there may be another key factor which could be interesting. Embedded in these studies is the implicit notion of predictability being beneficial for trust.

For example, Parasuraman's etiquette paper concludes that expected interruptions from autonomy (i.e. polite error messages) are beneficial for trust (Parasuraman & Miller, 2004). De Visser's transactional model for trust violation and repair suggests that trust violations can come from unexpected behaviours of autonomy, even if the unexpected behaviour had some benefit for the human team-mate (De Visser et al., 2018). This accords with the well established notion that transparency can mitigate the effects of imperfect autonomy (Hoff & Bashir, 2015; O'Neill, McNeese, Barron, & Schelble, 2022)

2 EXTENDED BACKGROUND

2.1 TRUST IN AUTONOMY

Parasuaraman et. al. set out the broad issues surrounding humans and autonomy (Para- suraman & Riley, 1997). Although they were describing tool-like automation, these issues are still relevant for teammate-like autonomy. Trust in autonomy is a key factor in determining how a human will utilise autonomous technologies, so it is a very important focus in the field of HAT. The most obvious, and the primary moderator of trust is the reliability and performance of the autonomy itself (Baker, Phillips, Ullman,

& Keebler, 2018; Chen, Barnes, Selkowitz, & Stowers, 2016; M. S. Cohen, Parasuraman, & Freeman, 1998; Endsley, 2017; Hancock et al., 2011; Hoff & Bashir, 2015; Ososky, Schuster, Phillips, & Jentsch, 2013; Parasuraman & Riley, 1997; Schaefer, Chen, Szalma, & Hancock, 2016). However, the very nature of technology means that we will often be dealing with imperfect autonomy.

There are also many other factors that influence trust, including anthropomorphism, group membership and organisational factors (Baker et al., 2018; Hoff & Bashir, 2015). These are important and should be mentioned, however they are not the focus of our study.

2.2 TRUST CALIBRATION

When dealing with imperfect autonomy, it is important to be able to know how much it can be trusted or it will be used improperly (Parasuraman, 1997; Cohen, 1998; Dzindolet, 2003; Lee, 2004; McBride, 2010). Broadly, this kind of managed trust in automation is termed as “trust calibration”. If the autonomy is visibly unreliable, the user may simply disuse it and do the entire task manually themselves (Freedy, DeVisser, Weltman, & Coeyman, 2007; Parasuraman & Riley, 1997). This is undesirable as it may lead to lower overall task performance because the user discards all advantages provided by the automation (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Lee & See, 2004; Wright, Chen, Barnes, & Boyce, 2015). Lee & See succinctly describe trust calibration as being the correct assignment of trust levels given the capabilities of the automation (Lee & See, 2004).

For clarity, ‘misuse’ is defined as over-reliance on automation or overtrust. ‘Disuse’ is defined as an underutilisation of automation and an under-reliance on automation or undertrust. Disuse and misuse had specifically been identified by Parasuraman (Parasuraman & Riley, 1997) and these definitions have been broadly adopted by the HAT community. Both ‘disuse’ and ‘misuse’ are undesirable outcomes, and transparency has been shown to mitigate them by enabling trust calibration. A clear example of this was demonstrated by Furukawa and Parasuraman in their experiments with aviation automation (Furukawa & Parasuraman, 2003). When they deliberately introduced delays or errors in the automation’s notification capabilities, pilots were able to predict or identify imminent engine failures simply by using the provided transparency information. Seong & Bisantz (2008) also found similar results in their aircraft identification decision aid study; showing performance benefits for unreliable automation that have transparency information available. Similarly, Wang et. al (2016) found that users were able to correctly reject an automated teammate's recommendation based on the teammate's observation explanations alone.

There are many more experimentally demonstrated examples of transparency being successfully used for trust calibration, as cited by previous survey works (Hancock et al., 2011; Lee & See, 2004; McBride & Morgan, 2010; Seong & Bisantz, 2008; Westin, Borst, & Hilburn, 2016). The overarching paradigm here is not to increase trust in automation overall, but to give the user enough context to be able to assign the correct level of trust given the dynamic circumstances (Cohen et al., 1998; Lee & See, 2004). Imperfect autonomy is assumed.

2.3 TRANSPARENCY

Transparency is a critical component in forming trust in human autonomy teaming (HAT) (Baker et al., 2018; Endsley, 2017; Freedy et al., 2007; Hancock et al., 2011; Hoff & Bashir, 2015; Lyons, 2013; Ososky et al., 2013; Parasuraman & Miller, 2004; Parasuraman & Riley, 1997; Schaefer et al., 2016). By allowing the user some insight into the autonomy's reasoning and "state of mind", transparency holds the key to mitigating or correcting errors caused by the imperfect autonomy (Baker et al., 2018; Chen et al., 2014; Lyons, 2013; Ososky et al., 2013; Schaefer et al., 2016; Selkowitz, Lakhmani, & Chen, 2017; Wright et al., 2015). Even when there are no errors, but the behaviour is unexpected or unintuitive to the user, transparency has an important role in forming a user's trust in the machine (De Visser et al., 2018; Endsley, 2017; McBride & Morgan, 2010; Parasuraman & Riley, 1997).

A quick and simple example of transparency would be displaying the intended pathway of a self-navigating vehicle. Another would be a symbolic representation of what a computer vision (CV) agent is seeing, where detected objects are clearly annotated. In either example, a user will be able to see the agent's intentions and any flaws in its reasoning. If the agent plots a course through a dangerous obstacle, or if the agent wrongly classifies a critical object, the user will have the opportunity to enact appropriate countermeasures to avoid failures. In the case where failures do occur, the user will have some level of explainability as to why it occurred and avoid such occurrences in the future.

It must be noted that transparency is not limited to graphical formats, as in the previous examples. Verbal and written natural language and audio are also common channels of communication (Hoff & Bashir, 2015; Lee & See, 2004). Haptic feedback is not as prominent but exists in some specialised contexts such as telerobotics (Brown & Farkhatdinov, 2021; Preusche & Hirzinger, 2007).

Generally, transparency is the mechanism by which we understand the actions of the autonomous agent. However, the term should not be misconstrued as simply the mass of information given to the human. Too much information will actually result in less transparency, as the human reaches their cognitive load limits, impeding their performance.

2.4 COGNITIVE CONSIDERATIONS IN HAT

Whenever the machine needs to send information to the user, they will inevitably incur a cognitive cost for receiving it. Be it reading a dial, or interpreting a graph, or even listening to voice output, the user will need to spend some of their cognitive capacity to absorb this information. This has been the motivating factor in Ecological Interface Design (EID) theory (Furukawa & Parasuraman, 2003; Westin et al., 2016) but has not been the focus of HAT research. However, there are some relevant studies.

Cummings et. al. were interested in this area and demonstrated the upper bounds of human cognitive capacity in monitoring tasks (Cummings & Guerlain, 2007; Cummings & Mitchell, 2008). Interestingly, when using a "utilization" metric (percentage of busy time), they found that task performance stopped improving at around 50-60%, as shown in Figure 1. This suggests that at lower utilisation rates, the user simply was not sufficiently engaged in the task to achieve their best results. Interestingly, it also clearly demonstrates that user performance is significantly affected as the user approaches their cognitive load limit.

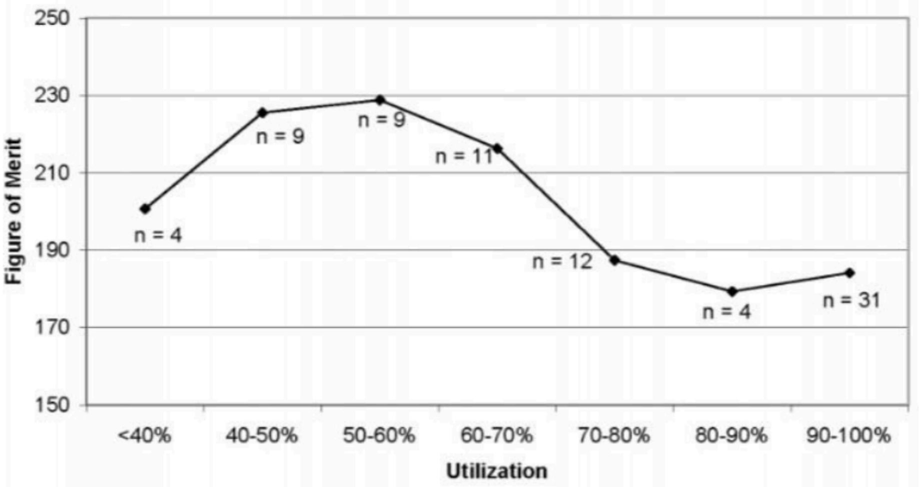


Figure 1: Cummings and Guerlain demonstrate a "cognitive hump", where performance peaks at 60% utilisation (Cummings & Guerlain, 2007). Human performance degrades after reaching a cognitive limit.

Chen et. al. have also found a similar “hump” for situational awareness (SA) and trust. In their experiments, each additional transparency layer resulted in improved trust and SA, until they reached SA1 2 3 U (Chen et al., 2016; 2014). Once the uncertainty information was added to the mix, SA and trust was comparable to when transparency was at SA1 2 only. It is reasonable to attribute this to the user reaching their cognitive limit, given the results of Cummings et. al. (Cummings & Guerlain, 2007; Cummings & Mitchell, 2008) and the general motivations of Ecological Interface Design (EID) theory.

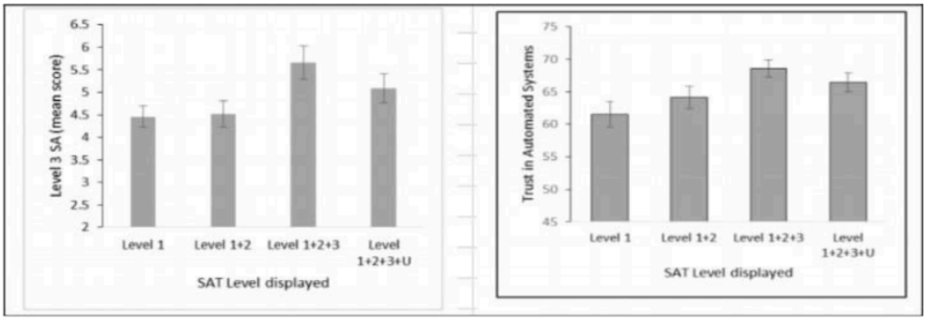


Figure 2: Situational Awareness and trust fall after SAT 1 + 2 + 3 is presented to the user (Chen et al., 2016, 2014). Human performance degrades after reaching a cognitive limit

Westin, et. al. (2016, p.202) support this explanation stating: “Increasing transparency by providing more information, can be a potential issue if the amount of information exceeds what the operator is capable to process within a certain amount of time”. Seong, & Bizantz (2008) have a further insight, suggesting that the type of transparency may have an effect on performance. They state: “It is feasible that a configural display of critical information can be better understood compared to an alphanumeric based display, which eventually may lead to better calibration of human trust”(Seong, & Bizantz, 2008, p. 624). Taken together, this suggests that different types of transparency will have varying cognitive costs, and there may be optimal types of transparency for certain types of information.

2.4.1 POSSIBLE ROLE FOR COGNITIVE FIT THEORY (CFT)

”Cognitive Fit Theory” (Vessey, 1991; Vessey & Galletta, 1991; Moody, 2009) offers a possible mitigation for the issue of cognitive overload. It is indeed possible to optimise data presentation for humans to better suit given tasks. More recently, Nuamah et. al. (2020) tested the user’s accuracy and speed in performing a judgment task with graphical and text/tabular information and found that the graphical mode of presentation performed best. Cognitive Fit Theory has so far only been investigated in static, non-dynamic, and time-insensitive contexts. That is, it has only been tested ”in the office”. We speculate that it may have applications in dynamic and time sensitive contexts as well. That is, CFT may have applications in the command and control (C2) domain.

2.5 TRUST VIOLATION AND REPAIR (TVR)

It is possible to degrade or violate a user's trust in the automation during the course of their interaction. In the context of HAT, the violation of trust centres around unexpected behaviour from the autonomous system (Baker et al., 2018; Cohen et al., 1998; De Visser et al., 2018; Dzindolet et al., 2003; Lee & See, 2004; Yang, Unhelkar, Li, & Shah, 2017). This encompasses blatant errors, total system failures, and more subtle, counter-intuitive behaviour, where there may not necessarily have been any errors. Over and above good trust calibration, it is also important to consider trust repair strategies when trust in au- tonomy is violated. Trust violation and repair (TVR) is a well researched topic in human to human (HH) relationships (Galdon & Wang, 2019), however there is fertile ground for exploration in human to machine (HM) relationships.

DeVisser et. al. proposes a transactional model of trust repair (De Visser et al., 2018). Human trust levels are increased or decreased when the autonomy engages in "relation- ship acts" and "relationship regulation acts". Cost acts can be roughly equated to trust violations and reduces trust. Beneficial acts are the opposite. Repair acts lessen the impact of cost acts, whereas dampening acts lessen the impact of beneficial acts. The combined impacts of these can result in a "net victim effect", which results in the degradation of trust.

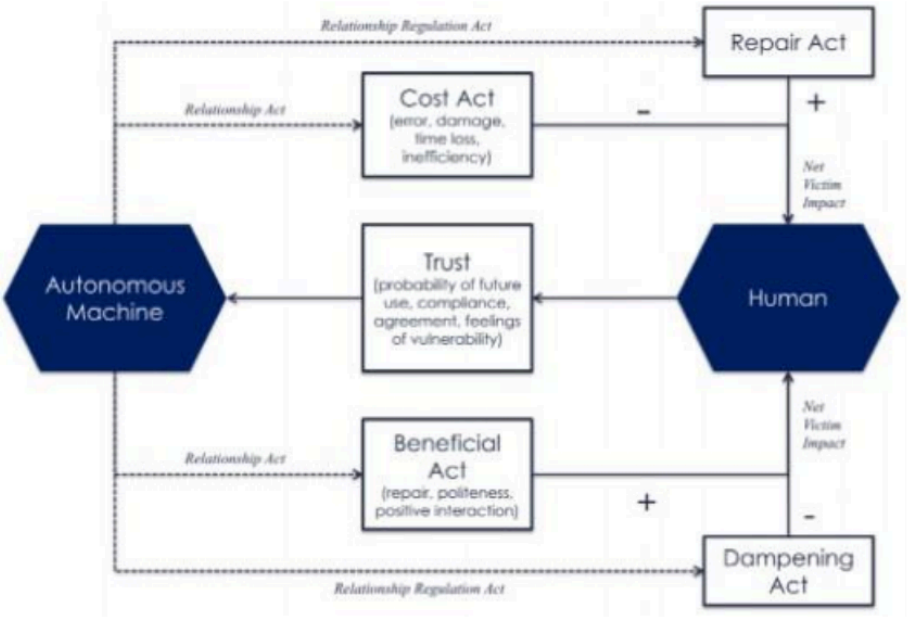

Relationship Regulation Act
Repair Act
Relationship Act
Cost Act (error, damage, time loss, inefficiency)
-
+
Net Victim Impact
Autonomous Machine
Trust (probability of future use, compliance, agreement, feelings of vulnerability)
Human
Net Victim Impact
Relationship Act
Beneficial Act (repair, politeness, positive interaction)
+
-
Dampening Act
Relationship Regulation Act

Figure 3: DeVisser's transactional model of trust violation and repair (De Visser et al., 2018)

In one illustrative example, the "net victim effect" is demonstrated in a situation where an automated personal assistant purchases a movie for the user without their explicit permission. In this instance, trust was violated because the user incurred an unexpected expense. However, this personal assistant was set to monitor the user's stress levels and act to alleviate their stress in various ways. The agent chose to purchase the movie because it was a type that the user liked and reasoned that having it available to watch when the user returned home would reduce their stress levels.

The trust repair behaviour in this example was for 1) the agent to explain its own reasoning and 2) offer a remedial action, which in this case was to initiate a refund. Once the user heard the agent's explanation of its behaviour and knew they had a remedial option, trust in the autonomy was repaired. This is of course only an illustrative example, and actual experimental data was not obtained. However, this is a guide for future research.

Although DeVisser's arguments are theoretical, Dzindolet et al. (2003) have experimentally demonstrated that providing explanations for the automation's errors improves trust. Indeed, a key finding that they have made is that knowing why an automation may make a mistake increases trust and reliance, even if the automation is ac- tually unreliable; another manifestation of automation bias. However, these studies were done in the context of trust calibration and not necessarily about active TVR behaviours. The results were also presented as overall summaries of trials in different conditions and do not show variability of trust over time.

To better understand TVR in HAT, we need to be able to see how trust is affected by trust repair behaviours, and this requires visibility of trust levels over time. DeVisser's paper provides for this by their example graphs (De Visser et al., 2018) but actual data from experimental studies is still difficult to find. Yang et. al (2017) have identified this gap and have run a study providing temporal data, however, once again, this study was focused on trust calibration, not TVR.

Trust repair and trust calibration are closely related in the sense that they both require transparency. The defining difference between them lies in the responsibility of the user in the interaction. For trust calibration, the user is actively assessing how much they can trust the automation and then acting accordingly. However, in TVR, it is the automation that is actively trying to regain the trust of the user. We advocate bypassing the TVR process in favour of promoting agent predictability.

3 PROPOSING A PREDICTABILITY BASED MODEL

The motivation behind this is to simplify more complex existing models such as the ones described by Hoff & Bashir (2015), and De Visser et al. (2018). The model described in Figure 4 shows the flow of possible interactions between the autonomous team-mate and the human in one interaction cycle. It also indicates the proportions of these interactions in a desirable scenario. The biggest arrow shows the most desired, and the smallest arrow shows the least desired.

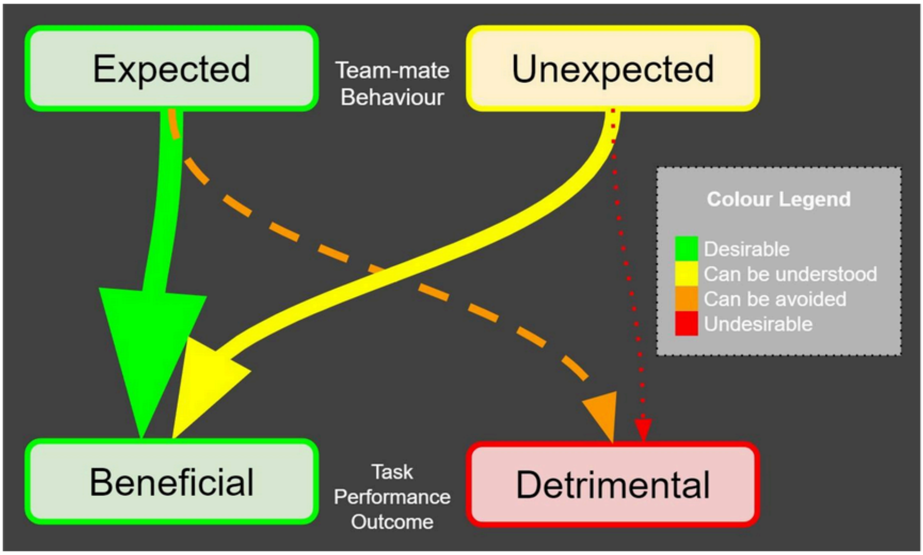


Figure 4: Human-autonomy teaming trust model based on predictability

While of course, the optimal scenario is to have all agent interactions be expected and beneficial, this is not always possible. When this model is viewed as part of a continuous interaction cycle, we can see that it is desirable to change all other agent interaction flows to one that is expected and beneficial. Where it is possible, all effort must be made to make this transition.

All possible interventions to make this happen are not explicitly shown in the model, however it does attempt to encapsulate them. These interventions can of course include, but are not limited to:

1. Reallocation of team-mate tasks

- e.g. the human will take over tasks that the agent performs badly

2. Adjustment of team-mate performance parameters

- e.g. the human will only allow the agent to perform automation to a level that the human is confident that the autonomy can perform well

3. Adjustment in transparency schemes to be more suitable for the current context

3.1 BEHAVIOURAL CONTRACTS

Recall that in Section 2.5, De Visser et al. (2018) described an illustrative example where an autonomous agent wrongly anticipated the desire of its human counterpart to purchase a DVD. This resulted in what they referred to as a ”Net Victim Effect”. We note that one cause of the net victim effect could have been avoided if the agent had only asked permission in the first place. There was a point where the user had a complete lack of transparency of the process; a Human out-of-the-loop (OOTL) situation as Endsley (2017) would describe it. However, in time-critical settings, asking for permission may not always be a practical interruption, as the user might be concentrating on higher priority tasks. It would be interesting to see if violation of consent is involved here, and if a mechanism for pre-consent can mitigate or even eliminate the net victim effect.

We suggest that a mechanism for pre-consent could come in the form of behavioural contracts. That is, the human teammate and the autonomy negotiates pre-agreed parameters of behaviour to avoid a human out-of-the-loop situation as shown in the illustrative example. The artificial agent would not need to ask permission at the point of decision making because pre-authorisation has already been given. This avoids the need for a prompt that might potentially be intrusive, and also avoids unexpected behaviour from the agent. The idea of behavioural contracts is embedded in the proposed model, however it is not tested in this study. We will study this directly in future work.

4 RESEARCH QUESTIONS

Considering the exposition that was given in the background sections, and our proposed predictability model, we aimed to answer the following questions:

- What is the relationship between trust and agent predictability?

- What is the relationship between trust and user workload?
- Can cognitive fit theory be applied to the command and control (C2) context?

5 METHODOLOGY

An air traffic control task was given to 70 undergraduate psychology students in exchange for course credit. This was conducted in multiple Zoom sessions where participants did the task on a web browser. Participants were exposed to several trials in which they were paired with an autonomous team mate which they could intervene with if they thought appropriate. Each session was strictly timed to 30 minutes, and the number of trials participants were exposed to varied depending on how fast they completed the surveys, and/or if they required extra time in the training phase.

This paradigm was chosen because of its ability to quickly increase a user's cognitive workload (Cummings & Guerlain, 2007), forcing them to rely on their autonomous team- mate. Users were also discouraged from intervening unless they deemed it necessary as an extra incentive to rely on their autonomous team-mate.

The online context was deliberately chosen to avoid the logistical challenges and the recruitment difficulties involved in face-to-face studies. For example, Cabanag et al. (2012) only had eight participants due to these challenges. Finally, given the continuing pandemic situation, it is preferable to minimise face- to-face contacts where possible and practical.

5.1 TASK (GAME)

The goal of the game was to safely land as many aeroplanes as possible. Each aeroplane carried a cargo value, which would be added to the participant's score when it had safely landed. If an aeroplane crashed for any reason, the value of their cargo would be deducted from the player's score.

To discourage users from simply micromanaging the aeroplanes, all interventions would deduct one point from their score. This prevented them from indiscriminately making interventions, and interventions would only be advantageous if it resulted in saving aeroplanes from crashing.

5.1.1 Collisions

It was possible for aeroplanes to collide with each other, and this is obviously an undesirable outcome. The participant would be able to detect this by observing the direction of aeroplanes and their indicated heights. The player's display is a typical, two dimensional, top-down view, akin to a real air-traffic controller's display.

A subtle situation that must be noted is what we refer to as a ”feigned collision course”. This is where aeroplanes appear to be on a collision course on the 2D, top-down display, but in fact indicate different heights. In this case, no collision will occur as the aeroplanes will safely overfly each other.

Players are informed that they should intervene if they detect a collision course, but not intervene otherwise. A key error that we observe in this study is when participants unnecessarily divert the aforementioned ”feigned collision course.”

5.1.2 Danger Zones

Aeroplanes could also be endangered by simply flying over designated areas, which would be indicated by a translucent red box. When aeroplanes made incursions into these danger zones, they would slowly take damage until they finally crashed when their health reached zero. In the obvious case, participants are encouraged to avoid the danger zones.

Again, there was a subtle situation that must be noted, which we refer to as a ”safe danger zone incursion”. Danger zones have a minimum safe speed. If aeroplanes travel at a speed above this level, they will not take damage while they are in the danger zone. Traveling safely through danger zones has an advantage in shortening flight distance to the goal. Players were informed of this fact and were encouraged to allow safe danger zone incursions.

5.2 VARYING VISUALISATIONS

Participants were exposed to three different types of data visualisations. They were:

- Text: Shown in Figure: 9
- Graphical: Shown in Figure: 10
- Text Graphical: Shown in Figure: 11

Please see Figures 9, 10 and 11 at the end of this document

5.3 TEAM ROLES, AUTONOMY AND HUMAN

Both the human and the autonomous team-mate could instruct the aeroplanes to make diversions at any time. The autonomous team-mate was ostensibly actively ensuring the safe passage of all aeroplanes, while the role of the human was to oversee the autonomous team-mate's decisions.

Errors and other specific behaviours were deliberately executed by the autonomous team-mate in some trials. The participant's reaction to these errors and specific behaviours were observed and measured.

5.4 BEHAVIOURAL MEASURES

We considered a higher number of interventions to be an indication of mistrust of the agent. Conversely, we considered a lower number of interventions to be an indication of trust.

We also monitored specific successful and erroneous behaviours. There were two distinct tasks which were presented to the participant, and they were:

- Managing Collisions
- Managing Danger Zones

As the human player and the autonomous-agent interacted, we counted the occurrences of these successful and erroneous behaviours. While there were numerous sub-behaviours involved in here, these were amalgamated into the following key measures:

- Collision Management:

– Successes

– Errors

- Danger Zone Management:

– Successes

– Errors

Please refer to the Appendix for a full listing of all behavioural measures, including the behaviour codes that were used for data gathering. Behaviour codes are relevant for reading the data summary in Figure 8.

5.5 SELF REPORT SURVEYS

After each trial, participants were given two surveys to measure perceived workload and agent predictability:

- The NASA Task Load Index (TLX) (Hart, 2006) to measure the perceived workload after each trial.
- A ”Competence and Predictability” survey to measure the perceived performance of the autonomous agent. The 1-5 Likert scale questions are:

– This autonomous team-mate contributed to successfully performing the overall task

– The autonomous team-mate made a lot of mistakes

– I knew what the autonomous team-mate was going to do

Please note that this survey was abbreviated in the interests of experimental time, and we acknowledge that the lack of co-verification questions is a limitation.

6 ANALYSIS OF RESULTS

6.1 PREDICTABILITY & TRUST

Results show that the most predictable agents were also the ones that were given the least amount of interventions, as shown in Figures 5 and 6.

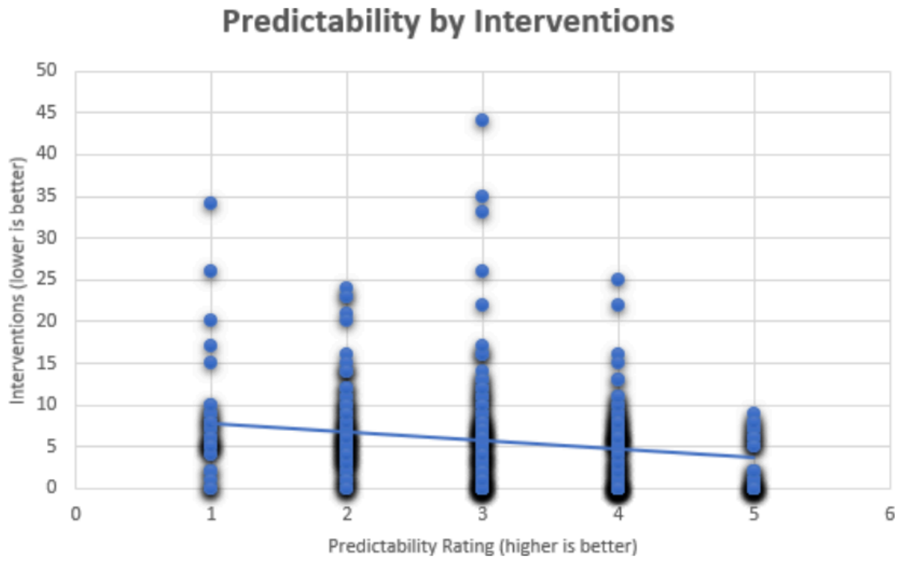


Figure 5: Scatter graph of intervention count grouped by predictability rating. Note the downward trend-line of interventions as predictability rating increases.

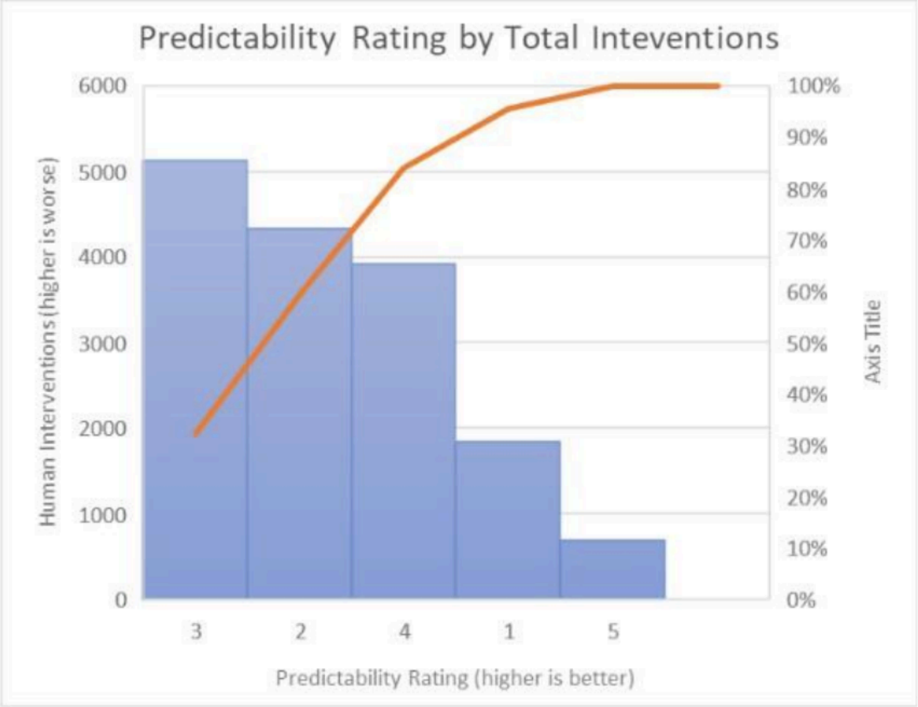


Figure 6: Pareto line chart showing that the highest predictability rating has the lowest number of human interventions.

Although we have not yet determined the key factors that affected the number of interventions, the Pareto line chart in Figure 6 indicates that a high level of predictability in the agent may be conducive to a high level of trust. It is strange that the lowest predictability rating resulted in the second lowest number of interventions, as we would have expected it to have the highest number of interventions. However, it is interesting to note that the highest predictability rating had over 2.5 times the number of interventions than what was observed for the highest predictability rating. This is indicative that agent predictability is indeed a desirable trait for trust in an autonomous team-mate.

6.2 PREDICTABILITY & WORKLOAD

Our results show that the most predictable agents were also the ones that had the least perceived workloads, as shown in Figure 7. It appears that high agent predictability is also conducive to low human workload.

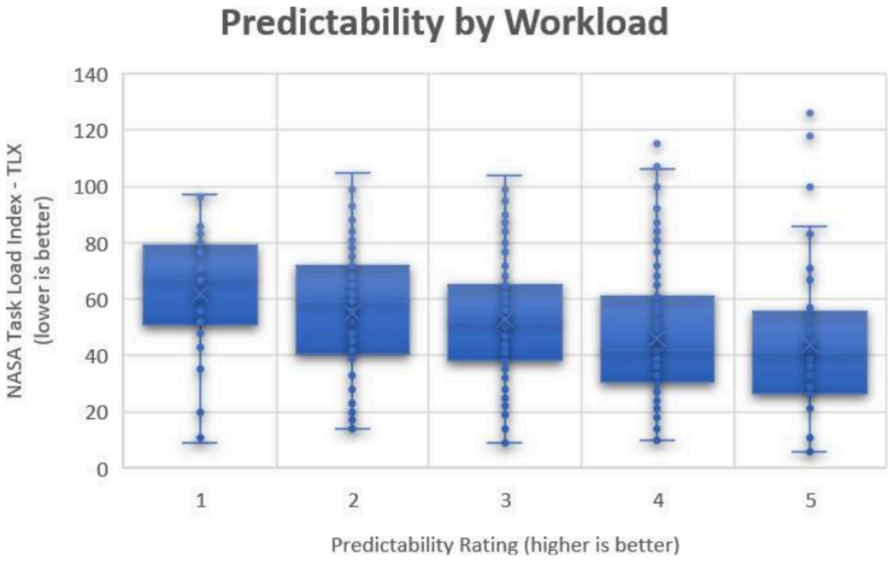
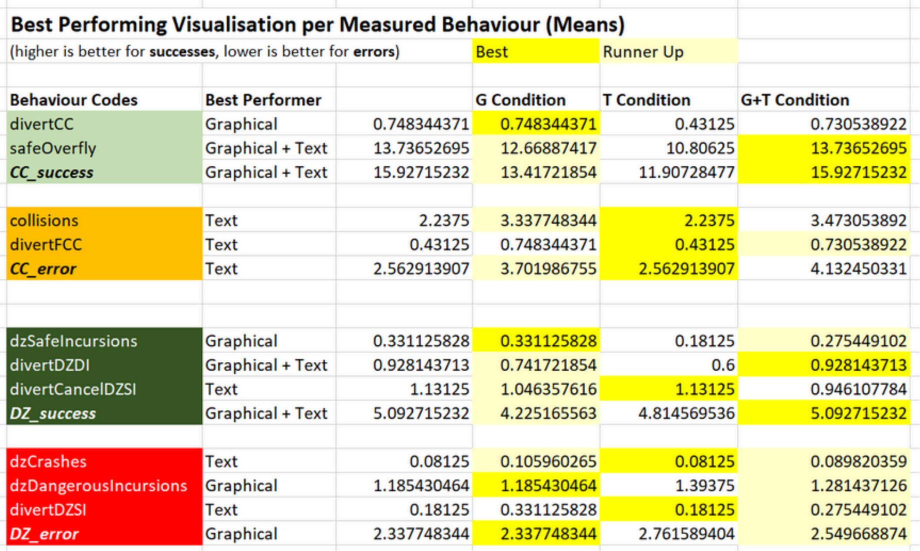


Figure 7: Boxplot of TLX rating grouped by predictability rating. Note the downward trend of workload as predictability increases

6.3 COGNITIVE FIT

Looking at Figure 8, we see that there is some clustering effect around some visualisa- tion types and task types. Specifically, it is interesting to note that errors for the Collision Course Management task dramatically fall in the Text visualisation condition. Similarly, graphical information seems to be a noticeable advantage for successful behaviours in the Collision Course Management task.

Still looking at Figure 8, we see that the Graphical Text visualisation does not au- tomatically result in the worst task performance, which suggests that the user is still able to use relevant available information (whilst filtering out the less relevant information) to perform their task. This is not necessarily surprising, but it would be interesting to see the effects of participants being close to their cognitive load limits.



Best Performing Visualisation per Measured Behaviour (Means)

(higher is better for **successes**, lower is better for **errors**) Best Runner Up

Behaviour Codes	Best Performer		G Condition	T Condition	G+T Condition
divertCC	Graphical	0.748344371	0.748344371	0.43125	0.730538922
safeOverfly	Graphical + Text	13.73652695	12.66887417	10.80625	13.73652695
CC_success	Graphical + Text	15.92715232	13.41721854	11.90728477	15.92715232
collisions	Text	2.2375	3.337748344	2.2375	3.473053892
divertFCC	Text	0.43125	0.748344371	0.43125	0.730538922
CC_error	Text	2.562913907	3.701986755	2.562913907	4.132450331
dzSafeIncursions	Graphical	0.331125828	0.331125828	0.18125	0.275449102
divertDZDI	Graphical + Text	0.928143713	0.741721854	0.6	0.928143713
divertCancelDZSI	Text	1.13125	1.046357616	1.13125	0.946107784
DZ_success	Graphical + Text	5.092715232	4.225165563	4.814569536	5.092715232
dzCrashes	Text	0.08125	0.105960265	0.08125	0.089820359
dzDangerousIncursions	Graphical	1.185430464	1.185430464	1.39375	1.281437126
divertDZSI	Text	0.18125	0.331125828	0.18125	0.275449102
DZ_error	Graphical	2.337748344	2.337748344	2.761589404	2.549668874

Figure 8: Some clustering effect is apparent when it comes to task and visualisation type. Note the CC error block resulting in notably lower error rates under the Text visualisation condition. Please see the appendix for descriptions of the behaviour codes.

7 LIMITATIONS

The results discussed here are preliminary and require deeper analysis. The simple anal- ysis of means needs to undergo significance testing to determine that there are indeed statistically significant effects. However, the clustering behaviour observed in Figure 8 seems to suggest that it might be statistically significant.

The survey based measurement of predictability is also a limitation. It would be more convincing to measure predictability behaviourally, i.e. we should try to see if the participant is successfully predicting their autonomous team-mate's behaviour, rather than just suggesting that they did.

Similarly, workload was also only measured using a self-report survey. Although the NASA Task Load Index or TLX (Hart, 2006) is often cited and used in our field, it may be of some benefit to measure workload behaviourally as well.

Additionally, the data sampled has unequal instances for each visualisation condition due to the strict requirement of ending the experiment after 30 minutes, even if the participant was not able to complete all trials. However, it is in the order of single digits across 478 trials. The exact numbers are: Graphical: 151, Text: 160, Graphical Text: 167. This small variation should not have a significant skewing effect, nonetheless we are noting it for the sake of reader transparency.

The strict ending time also affected the survey results, as participants were instructed to stop the experiment midway through. 14 trials did not have any TLX survey data, and were excluded from any analysis involving TLX.

8 IMPLICATIONS & FUTURE WORK

Our results support key notions that:

A : High agent predictability is tied with high levels of trust.

B : Reduced human workload is tied with increased agent predictability.

This accords well with the predictability based model that we proposed earlier in Section 3. Therefore, it can be said that strategies that increase agent predictability, and reduce human workload will indeed result in improved trust in autonomy. We believe further in- vestigation is warranted to study the validity of the proposed predictability based model. TVR behaviours, as described by De Visser et al. (2018) could be encapsulated into this scheme. However, we advocate explicitly investigating the concept of behavioural contracts, and the notion of pre-consent between the human and autonomous team-mates. This would bypass the need for TVR behaviours all together.

Our results are also suggestive of the phenomenon described by ”Cognitive Fit The- ory” (Moody, 2009; Vessey, 1991; Vessey & Galletta, 1991), whereby specific tasks are ob- served to have optimal visualisations. This study seems to demonstrate a counterexample to the findings by Nuamah et. al. (2020), which showed that the graphical condition was the best performing visualisation. As shown in Figure 8, the Text Condition for minimising errors in the Collision Management task was the highest performing visualisation type in this study.

This is not to say that Nuamah et. al. claimed that the graphical condition would be best generally. However, it does reinforce the idea that data visualisations are closely coupled to the specific task that they are aimed at.

This is a more nuanced approach that is taken by contemporary schemes such as Situational Awareness Transparency (SAT), which puts visualisations on different levels along the same axis. Generally, there is a prevalent practice of equating the term transparency with information given to the user (Westin et al., 2016). Adopting cognitive fit as a guiding principle allows us to specify that transparency refers to the understandability of a system, not simply the amount of visualisations presented to the user. In fact, taking this approach is prone to exceeding the human cognitive load limit and can inadvertently reduce transparency instead of increasing it (Westin et al., 2016).

Instead of pursuing a grand theory for cognitively efficient information displays, more pragmatic approaches may be required to meet the current needs of C2 operators. It may be more practical to use heuristic based approaches which will be able to readily accept and apply domain expert knowledge within the immediate context in which they operate. As these approaches proliferate, a clearer picture will emerge and would help inform a grand theory for cognitively efficient information displays.
Future studies will address some of the previously discussed limitations. The tutorial phase and other administrative parts of the trial will be shortened so as to allow enough time for participants to complete all exposures. A behavioural measure of predictability will be used, in conjunction with a survey based one. It would also be interesting to see how cognitive fit interacts with a wide range of workload conditions. This study did not manipulate workload, but only measured it using TLX. We would be specially interested in looking at workload conditions at users' upper cognitive capacity.

REFERENCES

Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(4), 1–30.

Brown, J., & Farkhatdinov, I. (2021). A soft, vibrotactile, shape-changing joystick for telerobotics. In *2021 IEEE World Haptics Conference (WHC)* (pp. 1158–1158).

Cabanag, M., Richards, D., & Hitchens, M. (2012). A novel agent based control scheme for rts games. In *Proceedings of the 8th Australasian Conference on Interactive Entertainment: Playing the system* (pp. 1–9).

Chen, J. Y., Barnes, M. J., Selkowitz, A. R., & Stowers, K. (2016). Effects of agent trans- parency on human-autonomy teaming effectiveness. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 001838–001843).

Chen, J. Y., Procci, K., Boyce, M., Wright, J. L., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency* (Tech. Rep.). Army Research Lab Aberdeen Proving Ground Maryland, University of Central Florida

Cohen, M. C., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). Anthropomorphism and trust in human-autonomy team communication dynamics. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 65, pp. 1056–1056).

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. In P*roceedings, Command and Control Research and Technology Symposium.*

Cummings, M. L., & Guerlain, S. (2007). Developing operator capacity estimates for supervisory control of autonomous vehicles. *Human Factors*, *49*(1), 1–15.

Cummings, M. L., & Mitchell, P. J. (2008). Predicting controller capacity in supervisory control of multiple uavs. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *38*(2), 451–460.

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation 'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human Factors*, *59*(1), 5–27.

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In *2007 International Symposium on Collaborative Technologies and Systems* (pp. 106–114).

Furukawa, H., & Parasuraman, R. (2003). Supporting system-centered view of operators through ecological interface design: Two experiments on human-centered automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 47, pp. 567–571).

Galdon, F., & Wang, S. J. (2019). From apology to compensation: A multi-level taxonomy of trust reparation for highly automated virtual assistants. In *International Conference on Human Interaction and Emerging Technologies* (pp. 42–46).

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, pp. 904–908).

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series.*

McBride, M., & Morgan, S. (2010). Trust calibration for automated decision aids. *Institute for Homeland Security Solutions*, 1–11.

Moody, D. (2009). The “physics” of notations: toward a scientific basis for constructing visual notations in software engineering. *IEEE Transactions on Software Engineering*, *35*(6), 756–779.

Nuamah, J. K., Seong, Y., Jiang, S., Park, E., & Mountjoy, D. (2020). Evaluating effectiveness of information visualizations using cognitive fit theory: A neuroergonomics approach. *Applied Ergonomics*, *88*, 103173.

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013). Building appropriate trust in human-robot teams. In *2013 AAAI Spring Symposium Series.*

O’Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, *64*(5), 904–938.

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Preusche, C., & Hirzinger, G. (2007). Haptics in telerobotics. *The Visual Computer*, *23*(4), 273–284.
Rebensky, S., Carmody, K., Ficke, C., Nguyen, D., Carroll, M., Wildman, J., & Thayer, A. (2021). Whoops! something went wrong: Errors, trust, and trust repair strategies in human agent teaming. In *International Conference on Human-Computer Interaction* (pp. 95–106).

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation:

Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400.

Selkowitz, A. R., Lakhmani, S. G., & Chen, J. Y. (2017). Using agent transparency to support situation awareness of the autonomous squad member. *Cognitive Systems Research*, *46*, 13–25.

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, *38*(7-8), 608–625.

Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, *22*(2), 219–240.

Vessey, I., & Galletta, D. (1991). Cognitive fit: An empirical study of information acquisi- tion. *Information Systems Research*, *2*(1), 63–84.

Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 109–116).

Westin, C., Borst, C., & Hilburn, B. (2016). Automation transparency and personalized decision support: Air traffic controller interaction with a resolution advisory system. *IFAC-PapersOnLine*, *49*(19), 201–206.

Wright, J. L., Chen, J. Y., Barnes, M. J., & Boyce, M. W. (2015). The effects of information level on human-agent interaction for route planning. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 59, pp. 811–815).

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 408–416).

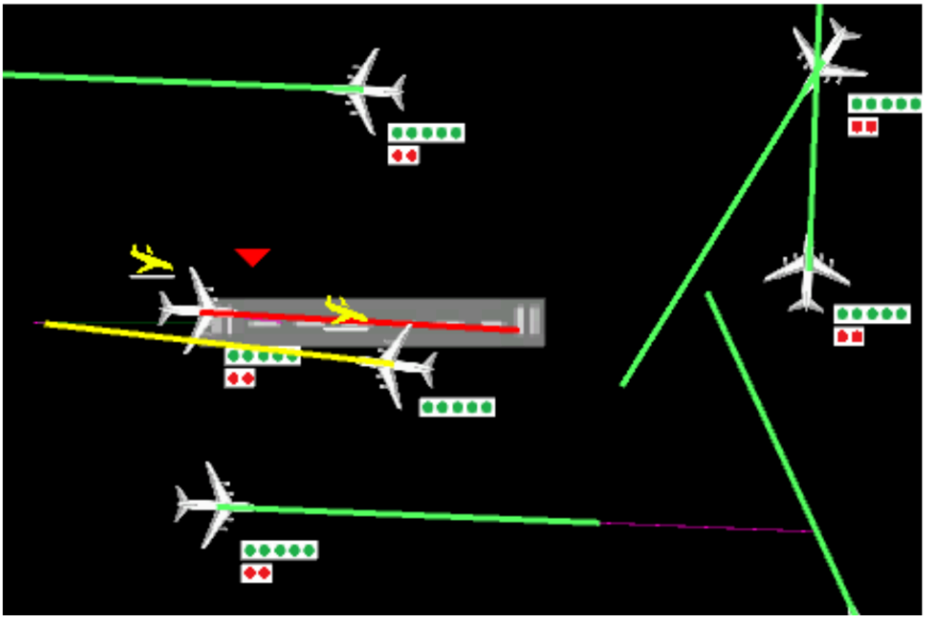


Figure 9: Text condition

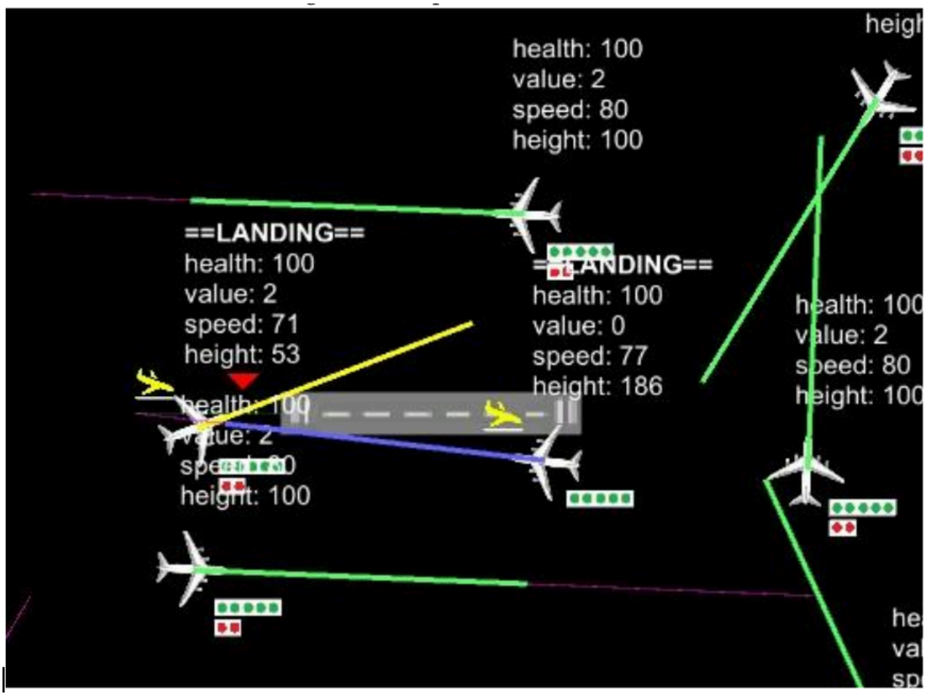


Figure 10: Graphical condition

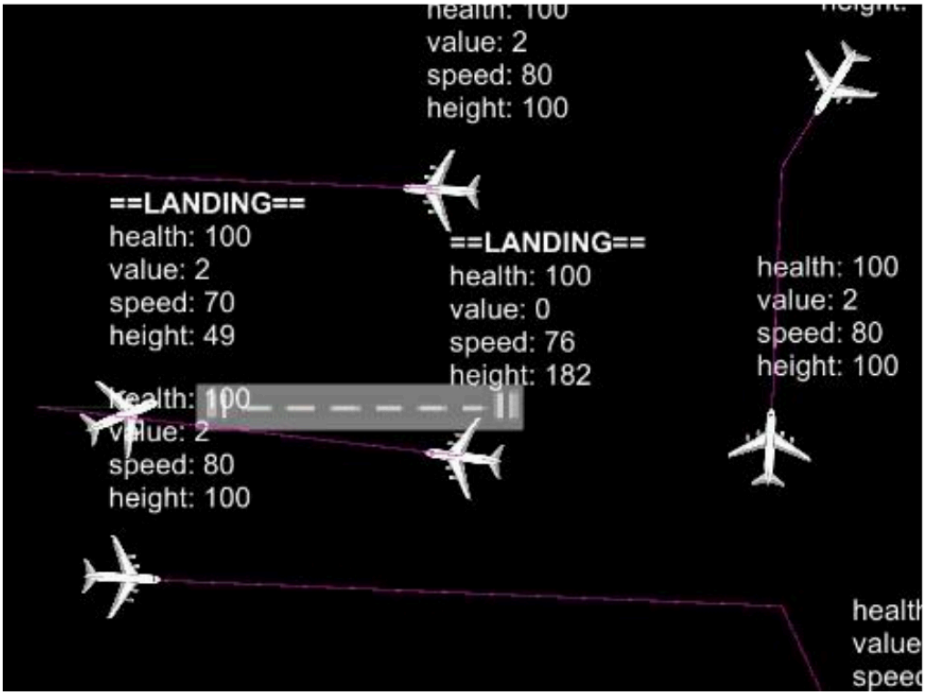


Figure 11: Text Graphical condition

A APPENDIX

A.1 FULL ENUMARATION OF BEHAVIOURAL MEASURES

For transparency to the reader, the conceptual tasks are further broken down into the following specific measures:

- Collision Management:
  – Successes (CC success)
  * Diverting aeroplane on collision course [divertCC]
  * Allowing safe overfly [safeOverfly]
  – Errors (CC error):
  * Aeroplanes collided with each other [collisions]
  * Safe overfly diverted incorrectly [divertFCC]

- Danger Zone Management:
  – Successes (DZ success):

* Safe incursions allowed [dzSafeIncursions]
* Diverting dangerous incursions [divertDZDI]
* Cancelling unnecessary diversions (of what was going to be a safe incur- sion) [divertCancelDZSI]
– Errors (DZ error):
* Crashes in the danger zone [dzCrashes]
* Dangerous incursions [dzDangerousIncursion]
* Diverting safe incursions [divertDZSI]
* Cancelling necessary diversions (inadvertently causing a dangerous incur- sion) [divertCancelDZDI]