Comparing Levels and Types of Situational-Awareness based Agent Transparency in Human-Agent Collaboration

Sylvain Daronnat, Leif Azzopardi, Martin Halvey
University of Strathclyde

Increasing agent transparency is an ongoing challenge for Human-Agent Collaboration (HAC). Chen et al. proposed the three level SAT framework to improve Agent Transparency and users' Situational Awareness (SA) by informing about (1) *what* the agent is doing, (2) *why* the agent is doing it and (3) what the agent will do *next*. Explanations can be *descriptive* (informing the user decision-making process) or *prescriptive* (guiding the user toward a pre-determined choice). To study these differences, we conducted a 3 (SA level) x 2 (explanation types) online between-group user experiment (n=180) where we designed six visual explanations and tested their impact on task performance, reliance, reported trust, cognitive load and situational awareness in a goal-oriented HAC interactive task. We found that SA level 1 explanations led to better task performance, while SA level 2 explanations increased trust. Moreover, descriptive explanations had a more positive impact on participants compared to prescriptive explanations.

Introduction

A wide range of visual explanations have been employed in Human-Agent studies to help users understand an agent's intent or to better digest information coming from the environment of interaction. In past work on Situational Awareness, SA was often studied by measuring how much information is processed or understood by the user in a task, at a given time (Endsley, 2017). A better SA is often linked to better task performance (Graafland et al., 2017) and a more appropriate reliance on an automated system. The "Situation Awareness-based Agent Transparency model" (SAT) (Chen et al., 2014) defines three different SA levels designed to assess users' understanding of a situation or a system's decision. This framework represents a good avenue to inform on the design of effective visual explanations and anticipate their impact on users. In a study centred around the design of Head-up Displays (HUD), Charissis et al. stated that "*a successful human-centered interface should enhance human actions* [...] *senses* [...] *and judgement* [...]. *Furthermore, it should guide the user rather than constrain his/her* [...] *abilities.*" (Charissis & Papanastasiou, 2010). While Charissis' work was focused on Human-Machine Interfaces (HMI) in an automotive environment, this statement can be applied to any task that, similarly to driving, require users to understand changes in the environment and respond to it quickly and appropriately.

A range of visual explanation techniques have been studied to communicate information as quickly and efficiently as possible (Charissis & Papanastasiou, 2010; Shekhar et al., 1991). Alphanumeric (alphabetical and numerical) symbols are among the oldest and most widely used methods to convey information visually. Lohse et al. classified different types of visual representations and found that numeric elements are often "unattractive" and likely to over-load users when utilized to emphasize parts of a specific representation (Lohse et al., 1994). As a more compact way of displaying information, other systems rely on icons or symbols, which impart an unambiguous meaning to a picture (Shekhar et al., 1991). Icons are used when the meaning of the icon is apparent to users, and were found to be interpreted much faster by human operators in fast-changing scenarios, where iconic displays led to response times three times smaller than with alphanumeric displays (Shekhar et al., 1991). In fast-paced tasks, icons can also be used as attention indices where symbols only serve as pointers for users to quickly know on which element(s) to focus their attention on (Storm & Pylyshyn, 1988). Most current visual explanation techniques rely on alphanumeric and/or icons or symbols to display information to users. While these modalities remain the same, their implementation can vary greatly depending on the task.

In a comprehensive report by Chen et al. (2014), Situational Awareness-based agents are touted to have a positive impact on reported trust in agents by improving trust calibration through the display of more information about an agent's inner working in a simplified form (Lee & See, 2004). The study presented here investigate the impact of different types of *visual explanations* (i.e., additional visual information to help explain the agent's decisions) on the human-agent relationship in a collaborative scenario. Each visual explanation was designed according to studies related to SA levels (Chen et al., 2014) and was informed by empirical implementations of visual explanations from past Human-Agent Interaction (HAI) and Human Factor research. More specifically, this work aims to answer the following research questions:

RQ1. Which SA level(s) provide the best overall support for HAC?

RQ2. Which type of visual explanation (prescriptive or descriptive) offers the most benefits for HAC?

Method

Study Design. We conducted a remote online study where participants completed a 2D goal-oriented task with the help of agents. The same setup was employed for past human-agent studies (Daronnat et al., 2021). In this experiment, participants oversee the defence of cities from incoming missiles by firing projectiles at them. Some missiles are heading toward cities (True Positives or "TP") while others are not (True Negatives or "TN"). Participants can choose to leave the aiming to an aiming agent or to manually control the crosshair and gain priority over the aiming agent on the controls. During the task, the screen is partially occluded by moving "clouds" which can

This is a peer-reviewed, version of record of the following paper: Daronnat, S., Azzopardi, L., & Halvey, M. (2022). Comparing levels and types of situational-awareness based agent transparency in human-agent collaboration. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 66(1), 1169-1173. https://doi.org/10.1177/1071181322661498

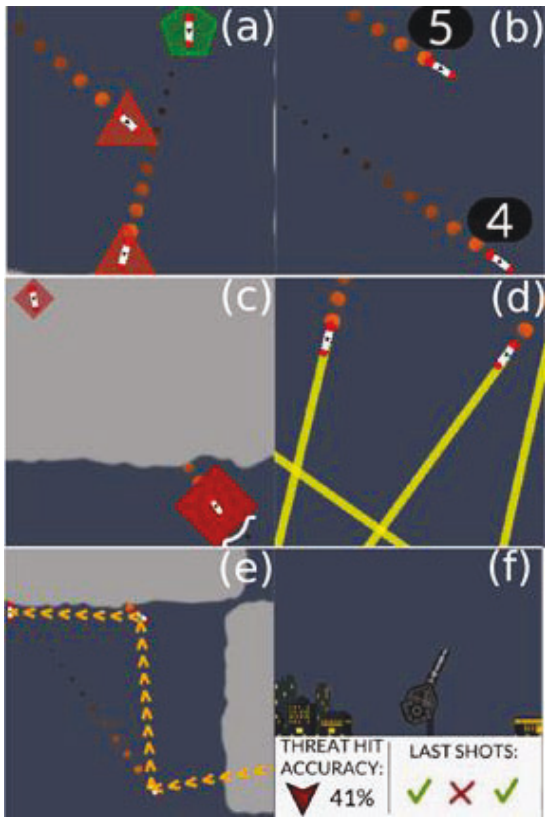

Figure 1: All types of visual explanations designed for this study. (a) and (b) support SA level 1, (c) and (d) SA level 2 and (e) and (f) SA level 3.

hide incoming missiles. In this study, participants are helped by *aiming agents* (moving the crosshair for the user) and/or *visual explanations* (displaying information to the user). We used a between-groups design where every participant interacted with the same *aiming agent* and one type of *explanation* (out of a total of six different visual explanations).

Procedure. The entirety of the study was conducted online using the Prolific© platform. Participants received £5.5 for undertaking the 40- minutes-long experiment. A total of 180 participants were recruited, divided in six groups of 30 participants. Each participant (1) completed a demographic survey, (2) played a tutorial, (3) played one session without any agent, (4) one session with an aiming agent, (5) one session with a visual explanation and (6) with an aiming and visual explanation. (items 4, 5, 6 were rotated using a Latin Square design). The type of visual explanation was selected depending on the group participants were recruited for (out of six possible groups, each corresponding to different visual explanations). Each session was comprised of two difficulty levels ("easy" and "hard") that lasted 120 seconds each. Survey instruments to measure reported trust, cognitive workload and situational awareness were presented after completion of levels.

Independent Variables

Task Difficulty. Every session was comprised of two the "Easy" level, 3 missiles spawned every 5 seconds at a speed of either 30 or 60 pixels per second. 30% of the missile spawned were "False Positives" or "FP" (not heading toward cities). In the "Hard" difficulty level, 3 missiles spawned every 4 seconds with a speed of either 60 or 80 pixels per second. 30% of the missile spawned were "False Positives" (not heading toward cities).

Aiming Agent. The aiming agent was set to have a performance level of 80%, meaning that 20% of the agent's decision would be incorrect (either targeting False Positive missiles or not targeting True Positive missiles at all).

Visual Explanation. The explanation displayed was selected among the six visual explanations presented in Figure 1. Each visualization is described before participants begin the task. Below is a description of the visual explanation developed for this study, as well as their type (descriptive or prescriptive):

Threat Shapes (SA Level 1). *Prescriptive.* indicates which target(s) the agent recognizes as threats with a red triangle and a green polygon for non-threats (Figure 1.a).

Priority Number (SA Level 1). *Descriptive*. displays the results of the agent's prioritization process via numbers, indicating which missiles the user should target **(**Figure 1.b)

Agent's Prioritization (SA Level 2). *Prescriptive***.** Targets deemed as threats are highlighted with red squares of different sizes and opacity (the bigger and opaquer, the more important according to the agent). This indicates the priority in which participants are recommended to deal with them (see Figure 1.c)

Missiles' Path (SA Level 2). *Descriptive***.** This visual explanation focuses on SA level 2 and understanding "why" the agent aims at certain targets based on their current paths. This visual explanation consists in displaying the paths of missiles and where they are pointing toward (Figure 1.d).

Agent's Plan Display (SA Level 3). *Prescriptive***.** This visual explanation displays paths between targets in the order where the agent is aiming at them. This gives an explanation regarding why the agent is heading in a particular direction (see Figure 1.e)

Performance Graph (SA Level 3). *Descriptive***.** This visual explanation support SA level 3 and gives more general information about the current level of performance as well as the current trend (whether the team is getting better - with a green checkmark or worse - with a red cross) (see Figure 1.f).

Dependent Variables

Reported Trust. A single item instrument where participants rated the statement "I can trust the agent" on a 7-point Likert scale (Jian et al., 2000) was used after each level.

Reported Cognitive Workload. The NASA TLX 6-items survey workload (Hart & Staveland, 1988) was used to measure cognitive workload after each session.

Reported Situational Awareness. The short "3D" SART questionnaire (Endsley et al., 1998) was used after each session.

Reliance. Participants' reliance on the agent was studied by monitoring how many times participants corrected the agent and for how long (User Control Time). A greater control time indicates lower reliance on the agent.

Task Performance was assessed using "Threat" Precision, Threat Recall and Threat F1 scores. Threat Recall is computed by dividing the amount of Threatening Missiles hit (TP) by the amount of total Threatening Missiles Spawned (FN+TP) and gives an overall estimate of how well participants performed at the task. Threat Precision is computed by dividing the amount of Threatening Missile hit (TP) by the amount of Threatening Missile hit (TP) and Non-Threatening Missile Hit (FP) and gives an estimate of how efficient participants were. F1 is the harmonic mean of both. A higher score indicates better performance. The analysis will focus of "Relative" metrics which represent a relative gain or loss in one session compared to another and are computed by subtracting a score in one session to another score in a reference session.

Demographics. Ethics approval for this study was obtained from the University of Strathclyde CIS Departmental Ethics Committee (App. No.1395). We recruited 180 (93M, 87F) participants aged from 18 to 24 (n=104) and 25 to 34 years old (n=76). In terms of level of education, most participants reported having a bachelor's degree (n = 77) while the rest reported having a college degree (n = 33), high school diploma (n = 32), master's degree (n = 20) or other (n = 38).

Results

We focused our analysis on the comparison of *relative* between groups differences compared to a baseline condition where participants were not helped by an aiming or visual explanation. These relative scores help us understand the relative impact of adding a visual explanation, an aiming agent or both.

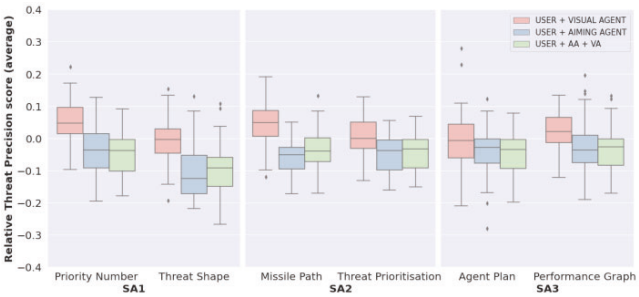

Figure 3: Relative threat precision score. A higher score indicates more relevant target(s) being hit.

Relative Task Performance. We studied performance with the Threat Recall, Precision and F1 score. A higher score indicates better performance. Overall, participants scored higher across all metrics when an aiming agent was present, no matter whether a visual explanation was provided or not. Without aiming agent, participants performed better with some visual explanations in SA 1 and 2 (see Figure 3 and Table 1).

A Welch ANOVA yielded significant results for Relative Threat Recall scores *(F=11.78, p<0.0001, np^2=0.14)*. Further pairwise comparisons using Gameshowell tests indicated that participants supported by a visual explanation in the Agent Plan (SA3) group performed significantly worse than participants supported by a visual explanation in the Priority Number (SA1) *(T=5.56, p=0.001, CLES=0.76)*, Missile Path (SA2) *(T=5.21, p=0.001, CLES=0.75)*, Threat Prioritization (SA2) *(T=7.16, p=0.001, CLES=0.82)* and Performance Graph (SA3) groups *(T=-93, p=0.0021, CLES=0.31)*. In addition, participants with a visual explanation in the Performance Graph (SA3) group scored significantly lower in terms of Relative Threat Recall than participants with a visual explanation in the Threat Prioritization (SA2) group *(T=4.3, p=0.001, CLES=0.71)*.

Similarly, an ANOVA yielded significant results for Relative Threat Precision scores *(F=8.41, p<0.0001, np^2=0.11)*. Further pairwise comparisons using Tukey tests indicated that participants with a visual explanation in the Priority Number (SA1) group performed significantly better than in the Threat Shape (SA1) *(T=4.65, p=0.001, CLES=0.73),* Agent Plan (SA3) *(T=4.59, p=0.001, CLES=0.72*) and Threat prioritization (SA2) groups *(T=3.65, p=0.004, CLES=0.68)*. In addition, participants with a visual explanation scored significantly higher Relative Threat Precision scores in the Missile Path (SA2) group compared to the Threat Shape (SA1) group *(T=-4.28, p=0.001, CLES=0.29)* and in the Missile Path (SA2) group compared to the Agent Plan

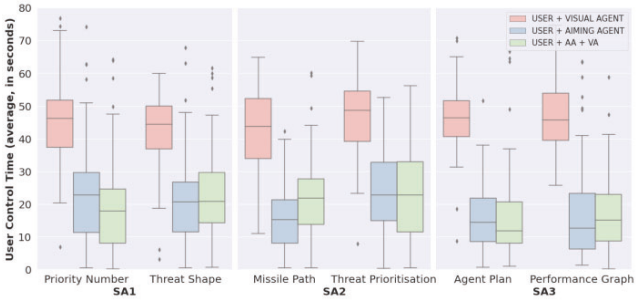

Figure 2: Average User Control Time across all sessions. A higher score indicates less reliance on an agent

(SA3) group *(T=4.22, p=0.001, CLES=0.71)*.

Reliance. User control times were used to study reliance. A higher amount of time represents less reliance on the aiming agent (see Figure 2 and Table 1). Participants controlled the crosshair less when aided by an aiming agent, with or without a visual explanation. However, participants relied on the aiming agent more in sessions with a visual explanation for the Priority Number (SA1) and Agent Plan (SA2) groups.

A Kruskal Wallis test yielded significant results for user control time *(H=21.99, p=0.0005)*. Further pairwise comparisons using paired T-TESTS indicate that participants supported by a visual explanation and an aiming agent relied on the aiming agent significantly more in the Agent Plan (SA3) group than in the Threat Shape (SA1) *(U=2246, p=0.0004, CLES=0.69)*, Missile Path (SA2) *(U=2113, p=0.0015, CLES=0.67)* and Threat prioritization (SA2) groups *(U=2030, p=0.0039, CLES=0.66)*. In addition, participants relied on the aiming agent significantly more in the Performance Graph (SA3) group than in the Threat Shape (SA1) group *(U=2120, p=0.0026, CLES=0.66)*.

Trust. Reported trust was studied using ratings to the statement "I can trust the agent" on a 7-items Likert scale. From consulting the results (see Table 1 and Figure 4) we can notice that participants' trust levels changed the most when a visual explanation was provided in sessions where an aiming agent was present.

Table 1: Scores for sessions with users and visual explanations only. Top: Performance and reliance metrics. Bottom: self-reported measures. "D." and "P." indicate Descriptive and Prescriptive Explanations. For Performance metrics, higher scores indicate higher performance. For User Control Time, higher scores indicate less reliance on the agent. For trust and SART, higher scores indicate a greater trust and a better SA, while higher TLX scores indicate a more cognitively taxing task.

	Threat Shape (SA1) (P.)	Priority Number (SA1) (D.)	Threat Prio. (SA2) (P.)	Missile Path (SA2) (D.)	Agent Plan (SA3) (P.)	Perf. Graph (SA3) (D.)
Threat Recall	0.72 ± 0.03	0.72 ± 0.02	**0.74 ± 0.02**	0.67 ± 0.03	0.59 ± 0.03	0.68 ± 0.03
Threat Precision	**0.92 ± 0.01**	0.90 ± 0.01	0.84 ± 0.01	0.90 ± 0.01	0.84 ± 0.01	0.84 ± 0.01
Threat F1	**0.79 ± 0.02**	0.78 ± 0.02	0.77 ± 0.02	0.75 ± 0.02	0.66 ± 0.02	0.73 ± 0.02
Relative Threat Recall	-0.01 ± 0.03	0.04 ± 0.01	**0.07 ± 0.01**	0.03 ± 0.01	-0.09 ± 0.02	-0.00 ± 0.01
Relative Threat Precision	-0.01 ± 0.01	**0.05 ± 0.01**	0.01 ± 0.01	0.05 ± 0.01	-0.01 ± 0.01	0.02 ± 0.01
Relative Threat F1	-0.04 ± 0.02	**0.05 ± 0.01**	0.04 ± 0.01	0.04 ± 0.01	-0.08 ± 0.02	0.01 ± 0.01
User Control Time	43.22 ± 1.61	45.42 ± 1.68	**47.48 ± 1.41**	42.67 ± 1.62	46.12 ± 1.45	47.02 ± 1.49
Reported Trust	3.25 ± 0.21	3.72 ± 0.23	4.27 ± 0.22	**4.75 ± 0.19**	4.05 ± 0.21	4.30 ± 0.24
Overall Raw TLX	54.66 ± 2.36	55.93 ± 2.78	57.78 ± 1.86	**58.04 ± 2.22**	57.72 ± 1.40	53.68 ± 2.02
Overall 3D SART	77.78 ± 5.18	80.68 ± 4.19	75.70 ± 4.37	75.48 ± 4.21	79.88 ± 5.21	**82.02 ± 4.40**

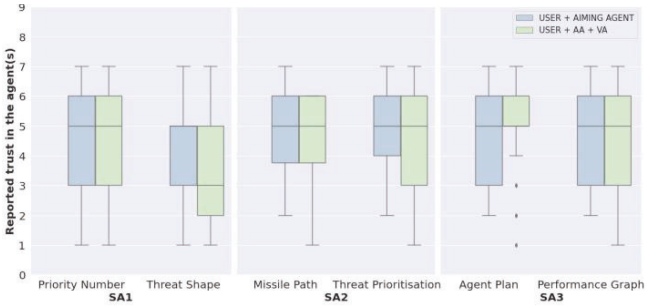

Figure 4: Reported trust. A higher score indicates a higher reported trust in the agent.

A Kruskal Wallis tests yielded significant result for Relative Trust scores *(H=12.31, p=0.03)*. Further pairwise comparisons indicate that participants supported by an aiming agent and a visual explanation in the Threat Shape (SA1) group trusted the aiming agent significantly less than in the Agent Plan (SA3) group *(U=1263, p=0.0039, CLES=0.35)*.

Relative Cognitive Workload. Raw NASA TLX scores were used to study reported cognitive workload. A higher score indicates a more cognitively taxing experience. From the results (see Table 1), we can notice that adding an aiming agent, no matter whether a visual explanation is present or not, reduces reported Raw TLX scores in all groups.

Overall, no statistically significant results were found when performing comparisons of Raw TLX scores between sessions without an aiming agent and with visual explanations *(F=0.43, p=0.82, np^2=0.01)* or sessions with an aiming agent and visual explanations *(F=1.66$, p=0.14, np^2=0.04)*.

Situational Awareness. The 3-items "3D" SART survey was used to measure situational awareness. A higher score indicates a better SA. From consulting the results (see Table 1), we can notice that reported situational awareness varied widely between sessions and SA groups. For most groups, the addition of an aiming agent improved reported situational awareness.

No statistically significant results were found when performing comparisons of Overall 3D SART scores between sessions without an aiming agent and with a visual explanation *(H=6.51, p=0.26)* or sessions with an aiming agent and visual explanation *(H=9.64, p=0.08)*.

Discussion

In this work, we created six visual explanations designed to increase transparency of an agent's actions by providing more information about the environment of interaction and reasoning of the agent during a HAC task. We designed each visual explanation to support a specific SA level (1, 2 or 3) and type (descriptive or prescriptive). We used task-specific metrics to study the evolution of task performance and reliance, as well as survey instruments to study reported trust, cognitive load and situational awareness. Our results indicate that some types of visual explanations have more impact on task performance and reliance, and that prescriptive explanations provide a better overall support than descriptive ones.

With our first research question (RQ1), we sought to investigate which SA level(s) were supporting users best. We found that visual explanations designed to support SA at level 1 and 2 had the best impact on task performance and reliance, but that differences in terms of reported metrics were less noticeable.

For task performance, the most interesting results were found when analysing Relative Threat Precision scores, which represent improvements or deterioration in Threat Precision scores compared to a baseline session (without visual explanation). Overall, participants supported by a visual explanation in the Priority Number (SA1) group achieved significantly higher Relative Threat Precision scores than in most other groups, which was surprising, as other groups that presented information that were faster to act upon (for instance, Threat Prioritizations (SA2) or Agent Plan (SA3)) resulted in either lower performance increase (Threat Prioritization) or even slight performance degradation (Agent Plan). Similar trends were observed for reliance, where SA1 and SA2 visual explanations increased reliance on the aiming agent, compared to SA3 visual explanations. When no aiming agent was present, the Priority Number (SA1) and Missile Path (SA2) visual explanations led to lower user control times compared to most other groups, which coincided with significant increases in Relative Threat Precision scores. The opposite happened in the Priority Number (SA1) group, but also resulted in better Relative Threat Precision scores. Overall, these changes show that visual explanations at lower SA levels (1 and 2) increase task performance while promoting more appropriate reliance on the aiming agent.

In terms of reported metrics, we did not find important differences between groups. For Trust, we found that

participants, in sessions without aiming agents, reported higher levels of trust in the visual explanations of the SA2 and SA3 groups compared to the SA1 group. When supported by an aiming agent and visual explanation, participants in the Threat Shape (SA1) group reported significantly higher trust in the aiming agent than participants in the Agent Plan (SA3) group. These results are not consistent with changes in reliance, signifying that perception of the agent's trustworthiness evolved independently of participants actual reliance on the aiming agent, which is at odds with previous HAI work that linked transparency as a positive factor for reported trust (Mercado et al., 2016). For Cognitive Workload and Situational Awareness, we did not find any significant differences between any of the visual explanations. These results are likely the reflection of the innate complexity of the task, which wasn't widely affected by the type of visual explanation and led participants to report comparable Raw TLX or SART scores across sessions.

With our second research question (RQ2), we studied the role played by the *type* of visual explanation (descriptive or prescriptive). We found that participants performed better with descriptive visual explanations compared to prescriptive ones. Overall, descriptive visual explanations such as Priority Number (SA1) and Missile Path (SA2) led to the most increase in task performance compared to other types of visual explanations. From our results, it seems that descriptive visualizations led participants to gain a better understanding of the task, which resulted in less false positive errors (higher Relative Threat Precision scores). Other visualizations that focused on processing more data for participants (Threat Shape at SA1 and Threat Prioritizations at SA2) gave more information regarding the agent's reasoning which induced better performance in terms of missiles hit (higher Relative Threat Recall scores) but made it harder for participants to distinguish between true and false positives (lower Threat Precision scores). From our findings, the best visual explanations provided participants with information regarding *what* targets to hit first (Priority Number) or *why* based on targets' predicted trajectories (Missile Path). This indicate that these explanations, which helped participants to make their own decisions, led them to get a better understanding of the task and, as a result, perform better at it. Our results are in line with past HAI work that found increased transparency to lead to better task performance (Mercado et al., 2016).

Conclusion

In this study, we have designed six visualizations based on previous SA work and tested their influence on human-agent collaboration in an interactive scenario. With our findings, we found that better performance can be achieved, by providing explanations about *what* the agent is doing, while trust in an agent can be increased by providing explanations focused on *why* the agent is acting in such a way. No clear improvements were observed with higher order (SA3) explanations. Moreover, participants reacted more positively to descriptive explanations rather than prescriptive explanations even though this increased cognitive load.

References

Charissis, V., & Papanastasiou, S. (2010). Human-machine collaboration through vehicle head up display interface. *Cognition, Technology and Work*, *12*(1), 41–50. https://doi.org/10.1007/s10111-008-0117-0

Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. J. (2014). Situation Awareness–Based Agent Transparency. *US Army Research Laboratory*, *April*, 1–29.

Daronnat, S., Azzopardi, L., & Halvey, M. (2021). Investigating the Impact of Visual Environmental Uncertainty on Human-Agent Teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 1185–1189. https://doi.org/10.1177/1071181321651209

Endsley, M. R. (2017). Toward a theory of situation awareness in dynamic systems. *Human Error in Aviation*, *37*(1), 217–249. https://doi.org/10.4324/9781315092898-13

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). Comparative analysis of SAGAT and SART for evaluations of situation awareness. *Proceedings of the Human Factors and Ergonomics Society*, *1*, 82–86.

Graafland, M., Bemelman, W. A., & Schijven, M. P. (2017). Game-based training improves the surgeon's situational awareness in the operation room: a randomized controlled trial. *Surgical Endoscopy*, *31*(10), 4093–4101. https://doi.org/10.1007/s00464-017-5456-6

Hart, S. G., & Staveland, L. E. (1988). *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research* (Vol. 43, Issue 5). https://doi.org/10.1016/S0166-4115(08)62386-9

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lohse, G. L., Biolsi, K., Walker, N., & Rueter, H. H. (1994). A Classification of Visual Representations. *Communications of the ACM*, *37*(12), 36–49. https://doi.org/10.1145/198366.198376

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 401–415. https://doi.org/10.1177/0018720815621206

Shekhar, S., Coyle, M. S., Shargal, M., Kozak, J. J., & Hancock, P. A. (1991). Design and validation of headup displays for navigation in IVHS. *SAE Technical Papers*. https://doi.org/10.4271/912795

Storm, R. W., & Pylyshyn, Z. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197.