Human Error, Reliability, Resilience, and Performance, Vol. 33, 2022, 8–15
https://doi.org/10.54941/ahfe1001562

Decision Support Systems for Route Planning: Impacts on Performance and Trust

Mary E. Frame[1], Jessica Armstrong[2], and Bradley Schlessman[2]

[1]Parallax Advanced Research, RDT&E Beavercreek, OH, USA
[2]Air Force Research Laboratory, 711th Human Performance Wing Wright Patterson AFB, OH, USA

ABSTRACT

Decision Support Systems (DSS) and other performance augmentation tools are increasingly leveraged by the military to recommend courses of action and augment analyst performance on critical tasks. This is particularly important for path planning operations, where analysts must consider complex tradeoffs and contingencies based on available assets, distance, and target priority. Emulating a more general applied context, we developed a long-range truck dispatch path planning task. Participants provided a quality control check of four simulated DSS, which ranged from perfect (100%) to sub-par (40%) accuracy. Participants reported lower trust of lower accuracy DSS, but their quality control performance was significantly lower when the DSS was below perfect accuracy. This demonstrates that while participants successfully calibrated trust in their DSS, they nevertheless experienced performance decrements, possibly due to anchoring on the DSS's incorrect results. The findings of this study provide the groundwork to understand the relationship between automation-reliance, trust, and performance.

Keywords: Human factors, Decision support systems, Human-machine teaming, Quality control

INTRODUCTION

Increasingly, military operational environments are incorporating automated tools to improve task performance. As these tools are developed, it is critical to evaluate operator trust, algorithm usage, and task performance. Decision Support Systems (DSS) are leveraged to help provide suggested solutions to problems, which are then either accepted or rejected by a human (Bonczek, et al., 2014). DSS are particularly useful in problem spaces where human discretion or problem solving is required for implementing an optimal solution, or in sensitive applied spaces where a human must be the final decision-maker. The DSS filters information provided to the human operator but has no decision-making authority (Wickens, et al., 2010). Tools that provide information on the feasibility of possible assignments are in development for application in multiple military environments with a goal of determining the feasibility of intelligence collection routes to yield mission-critical information. Conducting multiple mathematical operations and simulations quickly, DSS can provide guidance to operators on recommended courses of action, but those operators must make final mission decisions. As these tools are in development, it is critical to understand how DSS are perceived by users based on their understanding of how the algorithm is performing the task and as a function of the reliability of suggestions.

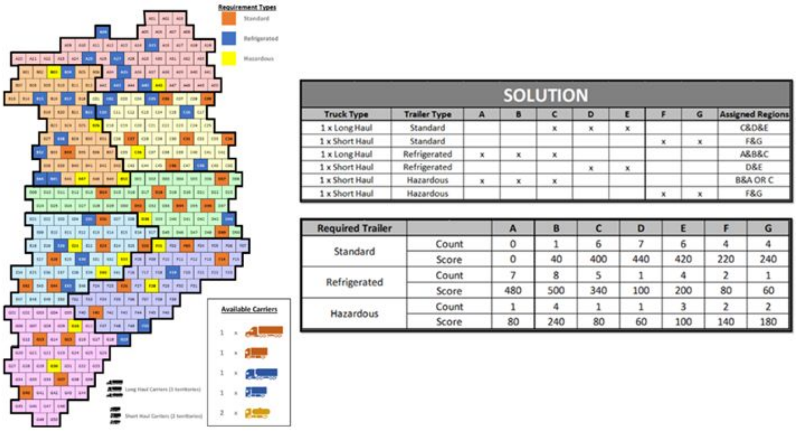
One major issue with any DSS is the issue of transparency into how the tool solves problems that are presented to human teammates. Although many studies have demonstrated the value of algorithm transparency in human-machine teams (Fox, et al., 2017) many current-generation algorithm processes are opaque (Clos, et al., 2017). Transparency refers to a user's ability to understand the processes that an algorithm uses to solve a particular problem. Increased transparency has been found to help reduce user bias and prevents inappropriate over or under-use of automated systems (Hepenstal, et al., 2019). Hepenstal (2020) asserted a model of needs for algorithm transparency that includes technical explanation of algorithms that are interpretable by the user. Then users can understand how the functional relationships are mapped against system goals or constraints, in addition to context for interpreting the explanation. Previous research by Hoff & Bashir (2015) demonstrates that an optimal explanation may not be universal, but may vary between individuals. Another critical factor in establishing trust in AI systems is the perceived reliability of the AI or DSS. Reliability refers to an AI behaving in a manner that is consistent and predictable, allowing a human teammate to have consistent, reasonable expectations. Trust in AI is reduced if a system has poor accuracy (indicating it doesn't follow expected procedures) or if it has inconsistent results (Glickson & Woolley, 2020). Previous studies have demonstrated that decreased accuracy or reliability of a tool or decision support system can lead to inappropriately calibrated operator trust (Lyons, et al., 2017). This can be further complicated if advice is accurate, but unintuitive upon superficial evaluation. Having too much or too little trust in a system can mean not relying appropriately on a particular tool, which can lead to problems in operational environments. Misuse occurs when a tool is relied upon excessively or unchecked, despite it not being accurate or reliable (Guidotti, 2021). Disuse, by contrast, refers to inadequate reliance on automation due to a lack of trust, yielding poorer performance. The confidence route planners have in their toolsets, improved by transparency and explanations of assistive algorithms, and their own expertise, influences their confidence in their decisions and their willingness to use the tool.

For the present study, we investigated how individuals interact with potential DSS for path planning. Critical to the adoption of these intelligence collection tools is the ability for humans to appropriately gauge their levels of trust to the performance of the automated systems and execute quality control. In this study, participants chose truck shipping routes and were provided with input from a DSS-informed bot on each scenario that recommended where shipping trucks should be dispatched. Participants engaged in a quality control task that allowed them to revise each bot's proposed solution and then rated their trust in each DSS. We hypothesized that: 1) Participants would report higher trust in more detailed algorithm explanation, compared to a simple written explanation, 2) Participants would have lower trust in a

DSS that provided less accurate recommendations, and 3) Lower bot accuracy would yield lower human performance, as would higher trust in lower reliability algorithms.

METHOD

Data was collected from a total of 247 participants through Amazon's Mechanical Turk experiment platform in compliance with the Air Force Research LaboratoryâŁ™s Institutional Review Board. Participants had to pass a screener consisting of data quality and task comprehension questions to be credited for participation and inclusion in the analysis. The mean age of the sample was 39 and the majority (76%) had completed at least a 4-year degree. 36.8% of the sample reported that they had a degree in a STEM field and 33.6% indicated they had a high degree of familiarity with maps/routing and 21.4% were highly familiar with algorithms. The task consisted of allocating a set of shipping trucks to regions of delivery destinations. Participants were told the monetary value of delivering to each location within a region and were told to select a limited number of regions for each truck to travel to. The goal was to maximize the total revenue of all routes. For each scenario, they were responsible for long haul vehicles, which could travel through up to 3 contiguous regions, and short haul vehicles, which could travel through up to 2 contiguous regions. There were three types of trucks indicated with different colors on the map: standard shipping trucks, refrigerated shipping trucks, and hazardous materials shipping trucks. On each trial, participants were provided with a map of destinations and a table of revenues, and a recommended route set from a simulated decision support bot (Figure 1). Participants received instructions and were informed that they were to evaluate the recommendations from separate bots for each scenario and perform a quality control task to determine which regions each truck should be allocated to. At the start of the experiment, each participant was presented with one of four descriptions of the DSS algorithm: (1) no explanation apart from the explanation of experimental procedures, (2) a text explanation of the optimization, (3) a flowchart of the algorithm calculations, and (4) example solution tables with process annotations. For the main task, participants were presented with a series of 4 maps with a summary of the total number of points for each truck type in each region and a proposed "optimal solution" proposed by the decision support bot. The bot accuracy varied as either 100%, 80%, 60%, or 40% of true optimal points by region allocation. The unreliable estimates were designed to appear plausible, but under good quality control checking, a novice participant could detect the suboptimality and make modifications. After viewing each scenario, participants were asked to choose the best solution for the scenario and queried on their trust in each bot. Participants then answered a series of Likert-rated questions pertaining to their trust in the bot based on the trust measurement questions by Lyons et al. (2017). After the primary task, participants viewed all three algorithm descriptions, provided feedback on their description preferences, and then answered a series of demographic questions.



SOLUTION

Truck Type	Trailer Type	A	B	C	D	E	F	G	Assigned Regions
1 x Long Haul	Standard			x	x	x			C&D&E
1 x Short Haul	Standard						x	x	F&G
1 x Long Haul	Refrigerated	x	x	x					A&B&C
1 x Short Haul	Refrigerated				x	x			D&E
1 x Short Haul	Hazardous	x	x	x					B&A OR C
1 x Short Haul	Hazardous						x	x	F&G

Required Trailer		A	B	C	D	E	F	G
Standard	Count	0	1	6	7	6	4	4
	Score	0	40	400	440	420	220	240
Refrigerated	Count	7	8	5	1	4	2	1
	Score	480	500	340	100	200	80	60
Hazardous	Count	1	4	1	1	3	2	2
	Score	80	240	80	60	100	140	180

Figure 1: Map of delivery locations by truck type (left), with summary table (bottom right) and DSS allocation recommendation (top right). Participants were also provided with a detailed list of locations and their point values, not pictured here due to space.

RESULTS

Algorithm Explanation Preferences. After completing all scenarios, participants were presented with all 3 potential explanations for the algorithm and ranked their preferences. A plurality of the sample (40.1%) ranked the written explanation as their top choice, 32.0% preferred the flowchart and 27.9% preferred the annotated tables. Participants provided long-form written responses on their preferred explanation, which we used to determine the top preference and calculated congruence between reported ranking and articulated preference. From the long-form responses, a plurality preferred the annotated tables (39.7%), followed by the written explanation (32.4%), and finally the flowchart (27.1%). The written and annotated tables explanations were robustly preferred, regardless of initially presented explanation. We further examined understanding of each algorithm representation on a scale from -3 to +3. All explanations were rated relatively high, with the written rated highest (M = 1.77), followed by the annotated tables (M = 1.64), and the flowchart (M = 1.21). Taken together, these results indicate that the flowchart was least preferred and rated as the most difficult to understand. An algorithm explanation that contains a simple step-by-step explanation seems to be better understood and more preferred by users.

Trust Results. We hypothesized that trust would be lower when the bot was less accurate, since participants could determine the bot's incorrectness using the data available. Trust was measured by averaging the trust subscales, aggregated across scenarios. Trust was significantly higher when the bot's accuracy was 100% compared to when it was only 40%, $F(3{,}261) = 3.44$, $p = .02$, $\chi^2 = 0.038$ (Figure 2, left). Regardless of bot accuracy, trust decreased significantly over time, $F(3{,}261) = 4.47$, $p = .003$, $\chi^2 = 0.013$,. The Tukey's HSD post hoc test indicated that trust in the bot was significantly higher during the first scenario compared to later scenarios, indicating that trust declined over time, but there was still a significant impact of bot inaccuracy

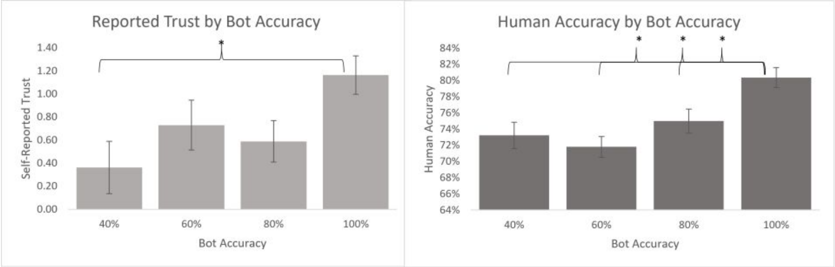


Figure 2: Trust (left) and human performance (right) was highest when bot accuracy was 100% compared to when bot performance was lower than 100%.

on decreasing trust. Participants rank ordered the bots, based on which they trusted most to least. Rank order was based on bot accuracy, $F(3{,}984) = 6.89$, $p < .001$, $\chi^2 = 0.021$. Tukey's HSD indicated that the accuracy of the most preferred bot was significantly higher than the bots ranked as worst and second-worst by participants, indicating that participants ascertained which bots were the most accurate during the task and scaled their ranking appropriately to bot performance.

Accuracy Results. To compare results across scenarios it is important that the scenarios were of relatively equal difficulty. Human performance was consistent across the 4 scenarios (ranging from 72.5% to 76.9%), with no significant differences in performance between them. This indicates that the four scenarios were of relatively equivalent difficulty. One of the primary goals of this experiment was to determine if human performance would vary as a function of a DSS bot's accuracy. We hypothesized that when bot accuracy was lower, human performance would also be lower due to anchoring on the bot' incorrect values. There was a significant difference in performance as a function of the bot accuracy $F(3{,}261) = 7.77$, $p < .001$, $\chi^2 = 0.082$ (Figure 2, right). A Tukey's HSD post hoc test found that participant accuracy was significantly higher when the bot was 100% accurate compared to the other three conditions.

We conducted a series of regressions between bot accuracy and human performance, as well as bot accuracy and trust in the algorithm for each given scenario. This allows us to determine if there was a relationship between bot performance and human performance and to further determine if trust was influenced by bot accuracy. By examining each scenario individually, we can see how this relationship persists over the course of the experiment over time. Table 1 summarizes these regressions. For the initial two scenarios, bot accuracy was highly predictive of human accuracy, but became less so over time, as indicated by non-significant regressions for Scenario 3 and Scenario 4. This indicates that it is possible that bot accuracy has a lower impact on human accuracy over time. As stated previously, trust in the bot declined generally over time, so it is possible that individuals relied on the bot's recommendation less over time and their performance became less correlated. Generally, bot accuracy was highly correlated with user trust in the bot

Table 1. Regressions results of bot accuracy, trust, and human performance.

Scenario	Regression	R^2	F	p
1	Bot Accuracy v Human Accuracy	.034	8.74	.003
2	Bot Accuracy v Human Accuracy	.039	10.02	< .001
3	Bot Accuracy v Human Accuracy	.01	1.78	> .05
4	Bot Accuracy v Human Accuracy	.003	0.86	> .05
1	Bot Accuracy v Trust in Bot	.023	5.79	.02
2	Bot Accuracy v Trust in Bot	.065	16.94	< .001
3	Bot Accuracy v Trust in Bot	.02	4.79	.03
4	Bot Accuracy v Trust in Bot	.015	3.85	.05

for all four scenarios. Higher bot accuracy led to greater reported trust in the bot.

Individual Differences. We examined the relationship between relevant self-reported participant characteristics, namely reported familiarity with algorithms and with route planning, on human task accuracy and on self-reported trust in the algorithms. There was a significant negative correlation between familiarity with algorithms and task accuracy r = -.337, p < .001. This indicates that those who were more familiar with algorithms in general had poorer task performance. This could be due to misappropriated trust, as higher reported familiarity with algorithms was significantly positively correlated with algorithm trust, r = .247, p < .001. If there was too much trust in the algorithm, this could lead to worse performance by relying on inaccurate bots. Furthermore, we found a significant negative correlation between reported route planning familiarity and performance, r = -.230, p < .001, and significantly positively correlated with bot trust r = .230, p < .001. This indicates that self-reported task familiarity overall led to higher reported trust but lower overall task accuracy. There was a high degree of correlation between reported route planning familiarity and algorithm familiarity, r = .743, p < .001.

DISCUSSION

In this study participants conducted a quality control task to optimally route a series of shipping trucks to delivery regions. During the task, they were provided recommendations from Decision Support System (DSS) bots that varied in their accuracy. Human performance was significantly lower when the DSS performed at 80% accuracy or worse, compared to when the DSS provided perfect recommendations. This indicates potential over-reliance on, or anchoring to, the initial values proposed by the bot. This performance detriment occurred even when participants accurately reported lower trust in worse-performing bots. This indicates that even when the inaccuracy of the bot is detected by the participant, and they report distrust, this does not mitigate the detriment to performance imposed by the initial incorrect region assignments. The task in this case was quality control, rather than generating a plan from scratch, and our results indicate that this may be difficult when recommendations are poorer. The effect of the bot on human accuracy indicates that a DSS should be developed to be as accurate as possible, and that a DSS that is not 100% accurate or reasonably close to 100% accurate may be detrimental to performance and need to be improved upon before being included in decision making to ensure that human decision-makers do not anchor onto incorrect values, as this is difficult for people to overcome and solve on their own.

The results of our inquiries into different algorithm explanations indicate that simpler, process-focused explanations are generally preferable for understanding the underlying calculations used by a DSS. Although richer visualized explanations (i.e., flowcharts) are engaging and preferable to a sizable minority of participants, they are not as well understood by the general sample. Additionally, Guidotti (2021) notes that the amount of time a user has to understand an explanation impacts how well that explanation will be understood. Under non-time-constrained conditions, a more exhaustive explanation can be provided. However, under time pressure, a more concise, easy to read explanation would be preferred. In our study, the task was self-paced, allowing participants to spend as much time as they desired to study the explanations provided, contributing to a preference for a step-by-step explanation. It is important to note that the sample in our study was drawn from a general novice population using MTurk, rather than expert operators in an applied setting, such as a professional long-distance truck dispatcher. Professional operators may have greater intrinsic motivation to perform well on the task, regardless of perceived DSS accuracy. Previous research has found that a combination of verbalization and visualization is beneficial for effectively communicating the structure and decision-making processes of machine learning models to users (Sevastianova, et al., 2018). Importantly, visuals should not be too complex or overwhelming, as this leads to information being ignored. This may have been the case with our flowchart, contributing to its ranking as the most poorly understood and least preferred of the three process explanations. Taken together with the reported explanation preferences, this indicates that a process focused explanation paired with a rich visual display, either of the tables or of visual task elements themselves, would be optimal for communicating the DSS's methodology for determining delivery truck allocation.

These results have implications for Decision Support Systems more broadly. Taken together, it is recommended that one has a careful understanding of the user population and the time constraints involved in the use of a DSS, to select the appropriate explanation or visualizations for different tasks. Additionally, it is critical to ensure excellent performance and adherence to the proper algorithmic method to ensure high accuracy. When a DSS is insufficiently accurate, even if a user can detect this inadequacy, they may be unable to overcome anchoring to these initial values and suffer poor performance themselves in attempting to correct it. There are many potential opportunities to extend our present paradigm to examine how DSS are understood and trust is established for allocation and path planning tasks, particularly in applied domains such as military operations. Our future studies will seek to further understand the relationship between automation-reliance, trust, and performance to determine when it is appropriate to allow automation to make recommendations to users in operational environments in varying difficulty and complexity.

REFERENCES

Bonczek, R. H., Holsapple, C. W., & Whinston, A. B.: Foundations of Decision Support Systems. Human Systems Management, 3, 324–328 (2014)

Clos, J., Wiratunga, N., & Massie, S.: Towards explainable text classification by jointly learning lexicon and modifier terms. IJCAI-17 Workshop on Explainable AI (XAI), 19 (2017)

Fox, M., Long, D., & Magazzeni, D.: Explainable planning. arXiv e-prints (2017)

Glikson, E., & Woolley, A. W.: Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14 (2), 627–660 (2020)

Guidotti, R.: Evaluating local explanation methods on ground truth. Artificial Intelligence, 291, 103428 (2021)

Hepenstal, S., & McNeish, D.: Explainable artificial intelligence: What do you need to know? International Conference on Human-Computer Interaction, 266–275 (2020)

Hepenstal, S., Kodagoda, N., Zhang, L., Paudyal, P., & Wong, B.: Algorithmic transparency of conversational agents. (2019)

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

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., Smith, D., Johnson, W., & Shively, R.: Shaping trust through transparent design: Theoretical and experimental guidelines. Advances in human factors in robots and unmanned systems (pp. 127–136). Springer (2017)

Sevastjanova, R., Beck, F., Ell, B., Turkay, C., Henkin, R., Butt, M., Keim, D. A., & El-Assady, M. Going beyond visualization: Verbalization as complementary medium to explain machine learning models. Workshop on Visualization for AI Explainability at IEEE VIS (2018)

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B.: Stages and levels of automation: An integrated meta-analysis. Proceedings of the human factors and ergonomics society annual meeting, 54 (4), 389–393. (2010)