Demand-Driven Transparency for Monitoring Intelligent Agents

Mor Vered, Piers Howe, Tim Miller, Liz Sonenberg, and Eduardo Velloso

***Abstract*—In autonomous multiagent or multirobotic systems, the ability to quickly and accurately respond to threats and uncertainties is important for both mission outcomes and survivability. Such systems are never truly autonomous, often operating as part of a human-agent team. Artificial intelligent agents (IAs) have been proposed as tools to help manage such teams; e.g., proposing potential courses of action to human operators. However, they are often underutilized due to a lack of *trust*. Designing *transparent* agents, who can convey at least some information regarding their internal reasoning processes, is considered an effective method of increasing trust. How people *interact* with such transparency information to gain situation awareness while avoiding information overload is currently an unexplored topic. In this article, we go part way to answering this question, by investigating two forms of transparency: *sequential transparency*, which requires people to step through the IA's explanation in a fixed order; and *demand-driven transparency*, which allows people to request information as needed. In an experiment using a multivehicle simulation, our results show that demand-driven interaction improves the operators' trust in the system while maintaining, and at times improving, performance and usability.**

***Index Terms*—Decision support systems, intelligent systems.**

I. Introduction

ARTIFICIAL intelligent agents (IAs) have been commonly used to help manage and supervise large, heterogeneous, robotic systems as a means of alleviating the workload on the human operator [1]–[3]. For this cooperation to succeed in complex scenarios, the human operator must rely on the IAs to perform part of the monitoring and supervision tasks. Reliance on IAs may prove hard for human operators as there may be many instances in which they do not fully *trust* the agent [4].

Manuscript received May 24, 2018; revised March 8, 2019, September 7, 2019, January 12, 2020, and March 22, 2020; accepted April 5, 2020. Date of publication May 18, 2020; date of current version May 18, 2020. This work was supported by a Sponsored Research Collaboration grant from the Commonwealth of Australia Defence Science and Technology Group and the Defence Science Institute, an initiative of the State Government of Victoria. This article was recommended by Associate Editor J. Y. C. Chen. *(Corresponding author: Mor Vered.)*

Mor Vered is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: mor.vered@monash.edu).

Piers Howe is with the Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, VIC 3010, Australia (e-mail: pdhowe@unimelb.edu.au).

Tim Miller, Liz Sonenberg, and Eduardo Velloso are with the School of Computing and Information Systems, University of Melbourne, Melbourne, VIC 3010, Australia (e-mail: tmiller@unimelb.edu.au; l.sonenberg@unimelb.edu.au; eduardo.velloso@unimelb.edu.au).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2020.2988859

Transparency is an effective method for increasing the operators' trust in the IA. It refers to an IA's ability to convey information regarding its internal reasoning process and the possible outcomes of its proposed actions to a human operator [5]. A better understanding of the IA's decision-making process often increases trust in the system [5]–[9].

The optimal level of transparency is unknown. Too little may result in no trust in the system, whereas too much may result in an unacceptably large increase in the workload of the human operator [10]. Mercado *et al.* [6] investigated the effects of different levels of IA transparency on the operator's task performance, trust, and workload, and found that increased transparency improved performance without increasing workload or response times. However, they considered a comparatively simple scenario in which the intelligent agent (IA) could make only one type of error: optimizing for the incorrect mission objective. It is unclear whether these findings would continue to hold in more complex scenarios. Our expectation was that in more complex scenarios, increasing transparency would likely result in an increased workload due to the increase in the amount of knowledge conveyed to the user.

The way in which users *interact* with information is also important. An adaptive system is one that adapts to the user's specific needs and specific context enabling more personalized, flexible interactions. The adaptation may involve altering the design of the interface, opening a dialogue between the IA and the user, or changing the manner in which the system's knowledge is represented [11]–[14]. Although transparency has been shown to improve different aspects of situation awareness, sometimes at the cost of others, to the best of our knowledge, no studies investigate how users interact with transparency information to guide their own situation awareness while minimizing information overload.

In this article, we present a new mode of transparency acquisition that we call *demand-driven transparency* (DDT). This model provides users with basic, coarse-grained control over which transparency information they acquire about specific components of the IA's reasoning process. In this way, the human operator can acquire information on demand. Our expectation was that providing flexibility for the human operator to choose which information to access would result in a more efficient interaction when compared with presenting the information in a fixed sequence. We contrast this approach with a baseline approach we refer to as *sequential transparency* (ST), in which the human operator must review the information regarding the IA's reasoning process in a predefined order.

We designed an experimental scenario comprising 12 different unmanned vehicles (UxVs) with nine different categories of capabilities divided into volatile capabilities (such as offensive capabilities that may decrease over time) and nonvolatile capabilities (such as possessing night vision). An IA was used to derive two potential plans for achieving several objectives. Both plans could contain several possible types of errors. The human operators (participants in our experiments) were tasked with determining which of the two plans, if either, was the most suitable plan. Our aim was to better portray the amount of information needed to make an informed decision in a plausible, real-world scenario. We expected that, in such scenarios, increased transparency would necessarily result in an increased workload for the human operator.

II. Related Work

As technology continues to evolve so does the use of robotic systems to assist with highly critical and influential tasks, such as mission planning within military applications [6], search and rescue (SAR) missions [15], and firefighting operations [16]. Due to the increasing complexity and size of these systems, failures or malfunctions frequently occur [17]. As the nature of these tasks brings them in close interaction with people in vulnerable situations, these failures can have catastrophic implications.

Typically, a human operator is used to manage these systems and to detect possible problems. This, however, is not an easy task. Supervising and monitoring large, robotic systems can be overwhelming and may result in suboptimal use [18]–[21]. Therefore, it is common to add IAs to serve as artificial team members, helping to manage and control these resources [1]–[3]. For this collaboration to work, the human operator must rely on the IAs to perform part of the monitoring and supervision task [22].

Humans do not always interact appropriately with IAs. Parasuraman *et al.* addressed theoretical, empirical, and analytical studies pertaining to human use of automation [23]. They identified both *misuse* and *disuse* of systems as challenges for human-agent collaboration. "Misuse" refers to users excessively relying on automation, which can result in a detection failure or decision biases [23]. To illustrate this issue, Parasuraman and Riley give an example of an accident that occurred near Columbus, OH, USA, in 1994, in which a pilot placed over reliance on the automatic pilot. Due to low visibility, the automatic pilot failed to monitor the aircraft's airspeed resulting in a landing short of the runway. On the other hand, "disuse" of automation refers to users rejecting the capabilities of a system resulting in underutilization [23]. For example, Sorkin [24] gave several examples of accident occurrences caused by operators intentionally bypassing or disabling important warning systems.

Misuse and disuse of IAs are more likely if *trust* is not appropriately calibrated [4]. Excessive trust in an AI may lead to "Misuse" whereby users' excessive reliance may fail to detect automation malfunction or erroneous behavior, whereas a lack of trust may lead to "Disuse" and therefore may result in IA underutilization and poor task performance. Trust is particularly important in situations that are complex and hard to analyze. In such instances, IA decisions may seem counterintuitive or surprising. For IAs to be used effectively in such situations, trust must be fostered. One way to do this is to design software that displays characteristics that are similar to the user [25]. Another way is to specify acceptable human-agent behavior as computer etiquette [26]. It is also possible to specifically design interaction systems to enhance trust and acceptance by altering graphics design, content design, structure design, or social cue design [27].

Mercado *et al.* [6] investigated the effects of different levels of IA transparency on an operator's task performance, trust, and workload. They conducted an experiment simulating a multi-UxV military scenario in which participants performed the role of an operator whose job it was to work with the IA to find the most appropriate course of action to take in different scenarios. In each scenario, participants needed to evaluate and compare two plans suggested by the IA, knowing that one of these plans would always be correct. Participants were asked to evaluate the appropriateness of a plan using three metrics: speed, coverage, and capabilities. Transparency was provided following the principles of the SAT model introduced in [5]. In these experiments, the only mistake the IA could make was to optimize for the wrong metric. Consequently, this was the only aspect of each plan the human operator needed to check, so the effective workload was minimal. Stowers *et al.* [28] continued this line of research using a slightly more complex scenario. However, even in these experiments, the human operator needed only to determine for which metric each of the plans was optimized. Consequently, all other aspects of the plan could be ignored.

Although these experiment scenarios comprised realistic military settings, the scenarios were simple enough that they did not place a sufficient amount of workload on the human operator. In more complex scenarios, increasing transparency can result in an increased workload for the human operator [29], [30], which, in turn, can negatively impact operator self-confidence [31] and trust in the system [23].

Transparency can be increased by making a system adaptive. An adaptive or user-driven system is a system that adapts to the user's specific needs and specific context. The adaptation may involve a different design of interface, opening a dialogue between the operator and the IA, or a different representation of the systems' knowledge [13], [14]. Adaptive user interfaces are commonly used to facilitate smoother human–computer interaction [11], [12], [32]. When dealing with human operators, we must consider that different users employ different processes when making decisions [33]. As such, we hypothesize that there would be significant advantages for allowing the operator to request information at their own discretion. In this manner, the system is more flexible and better suited to the individual needs of each different user, which we argue should not only reduce workload but also improve trust [34].

This concept has been further developed by the emergence of Explainable AI [35]. An explanation is commonly defined as a reason or justification for an action or belief. This field examines how an AI-based system clarifies its complex behavior

to a human operator. Research has included evaluating how users would respond to explanation systems [36], evaluating how human experts perform a task and form explanations [37] as well as developing and evaluating explanation systems. For instance, Fukuchi *et al.* [38] developed instruction-based behavior explanation as a method of explaining machine learning robotic systems' future behavior and Hayes *et al.* [39] introduced robots that automatically synthesize policy descriptions as explanations in response to general and targeted queries.

With this in mind, we introduce a new mode of transparency that we refer to as DDT. Crucially, this form of transparency is adaptive in that it allows the human operator to request information about specific parts of the IA's reasoning process, without being presented with all the information that is available. To do this, DDT utilizes a new model of IA transparency. By conceptualizing transparency in this way, we make it easier for the human operator to request only that information about the agent's reasoning process that is needed at a certain point in time. We expected that this would both reduce the operator's workload and increase the operator's trust in the system.

III. IA Transparency

In this section, we present our three-level model of transparency, which is based on Endsley's model of situation awareness [40]. We then describe two modes of transparency acquisition that utilize this model: *DDT*, in which a human is given flexibility over which levels they examine; and *ST*, in which the levels are presented in the order in which Endsley's model predicts people will gain situation awareness.

A. Three-Level Transparency Model

Endsley's model [40] is a widely-used model of situation awareness that consists of three consecutive levels of reasoning:

1) Perception (Level 1): Perception of the individual elements in the environment and their properties.
2) Comprehension (Level 2): Integration of the Level 1 information, in particular, inference of the relationships between the elements and the significance of these relationships.
3) Projection (Level 3): Prediction of the future state(s) of the system and therefore, necessarily, of the individual and her effect on the system, at least in the near term, based on Levels 1 and 2 understanding.

Endsley argues that these three levels model how people come to an understanding of their situation, enabling them to make time-critical decisions.

While there have been several criticisms of Endsley's model from the perspective of cognitive science—in particular, that it addresses only the knowledge states attained but does not address the cognitive processes involved [41]–[43]—we assert that this model serves as a useful basis for transparency. While it may well be the case that this model does not accurately reflect the processes used to derive situation awareness, its knowledge states reflect the way that people *think* they achieve situation awareness.

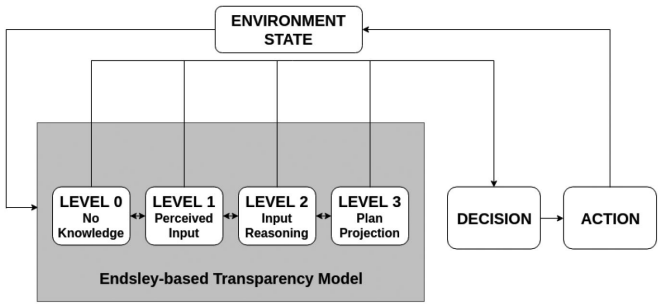


Fig. 1. Endsley-based transparency model (ETM).

Endsley models people's reasoning as they acquire situation awareness. Our model of transparency is based on the hypothesis that presenting an IA's reasoning in this manner is a natural way for an observer to achieve an understanding of the agent's reasoning process, and thus better judge the plans provided by the agent. In this, we are similar to the SAT model introduced by Chen *et al.* [5]. However, as we discuss later, our hypothesis is that not all of this information is useful, and giving people flexibility to review only some of this information can lead to improved results. Therefore, the new model further separates the transparency levels, giving people additional choices in terms of the information they acquire.

Using Endsley's model as a basis, we define a four-level model of IA transparency (see Fig. 1), which we call the ETM. The four levels represent the reasoning process of the IA. The initial level corresponds to no knowledge regarding the IA's internal reasoning except the proposed plans generated by the IA. This enables the human operator to make a decision without being exposed to the underlying reasoning. The subsequent three levels correspond to increasing levels of transparency, based on Endsley's model of situation awareness. As Fig. 1 shows, a decision can be made at any point without the necessity of viewing all levels of transparency. More specifically, the four levels correspond to the following.

1) *No Knowledge* (Level 0): This level presents the IA's decision(s) with no corresponding explanation. In many cases, a decision taken by the IA corresponds to a decision that the human observer would have made, so it may not require any explanation.
2) *Perceived Input* (Level 1): In this level, basic factual information about the input, *as perceived by the IA*, is made transparent to the user.

   Crucially, we do not display any *inferences* made by the IA. This level of transparency is sufficient for the user to detect errors arising from the IA's erroneous natural language processing or missing input; for example, a missing time or weather conditions alerts which was lost due to noise or miscommunication.
3) *Input Reasoning* (Level 2): In this level, we list the immediate inferences the IA makes based on Level 1 information. This level therefore directly corresponds to Endsley's Level 2. Examples would be the possible ramifications of weather conditions on the availability of an unmanned

aerial vehicle, or the possible intentions of potentially hostile vehicles entering an area. Crucially, this level does not address any specific plan.

4) *Plan Projection* (Level 3): In this level, we list the IA predictions regarding future events and the uncertainties of the occurrences of these future events. For example, this level would list the estimated travel time and expected future capabilities of the various assets, taking into account expectations as to how predicted conflicts will likely reduce these capabilities. This level also explains why a particular plan of action is preferable (likely to result in a better outcome) than another plan.

B. Process Model

In this section, we define two simple process models for instantiating this SA-based model: *ST* and *DDT*.

1) Sequential Transparency (ST): We characterize *ST* as the manner of acquiring information about the IA's reasoning process in predefined, ordered steps. The manner or substance of the information conveyed is controlled externally and provided in an identical manner for each decision, irrelevant of the context or the person using it. This approach has the advantages of leveling the field and maintaining a unified level of knowledge among all users while also making sure the operator will be exposed to specific information that may influence the decision-making process. However, because this approach targets the lowest common denominator, we hypothesize that it may be redundant or time consuming for some of the users in terms of the amount of information displayed.

2) Demand-Driven Transparency (DDT): DDT provides user autonomy by granting the user control of the flow of knowledge-state information. DDT allows a human operator/observer to determine not only the order at which to request the information but also the type of information requested. This allows for a much more personalized form of interaction. Note, however, that in scenarios where the user has chosen to visit all of the ETM levels sequentially, there may not be any difference between the performance of both approaches.

To illustrate the advantages and drawbacks of both approaches, consider the following example, based on the scenario that we use in our evaluation (see Section IV). The scenario is of a human operator monitoring a team of unmanned vehicles (UxVs) for surveillance of a set of offshore assets. The operator is assisted by an IA that produces candidate plans to respond to situations. In this particular instance, the IA missed an important alert concerning foggy weather conditions due to a transmitter malfunction.

Missing this alert could have one of two possible implications: It may negatively influence the final decision by the human unknowingly choosing to deploy a UxV that is not equipped to handle those weather conditions, or it may have no negative influence since the most appropriate UxV for the task may coincidentally happen to be equipped to handle foggy weather. If the first of these possibilities occurs the DDT model may have the advantage over the ST model. In the DDT model once concluding that the IA missed an important alert, the operator may directly skip to the end of the assessment and reject the plans suggested by the IA, disregarding all other information and hence saving valuable time, while the operator working with the ST model would first have to sequentially progress through all levels of transparency. However, if the second possibility occurs whereby the missed alert did not negatively influence the final decision, early abandonment of both plans due to a missed alert would be wrong and it would be better to request further information regarding the IA's reasoning process. In this instance, the ST model would have the advantage in that the system forces the operator to review all the information about the IA's reasoning process, thereby possibly reducing the chance of the operator making a rash decision.

IV. Evaluation Method

We wish to evaluate the advantages and disadvantages of *DDT* acquisition when compared to the baseline *ST* technique. We aim to give an insight into the effects of the technique on performance (measured by successful missions), efficiency (measured by trial duration), and participants' perceived trust and usability.

A. Participants

Participants for this experiment included undergraduate and graduate students from the University of Melbourne. We used SONA, an online participation recruitment system and public announcements to recruit 36 participants (17F/19M), between the ages of 18 and 43 with an average age of 26.2 years and standard deviation of 5.9 years. Six of the participants were disqualified due to insufficient understanding of the experiment, as further explained in Section IV-D. No prior knowledge was required except proficiency in the English language. The experiment lasted between 1.5 and 2 h and participants were reimbursed $20 AUD in gift vouchers.

B. Scenario

Our scenario replicates a highly complex, military planning task with the aim of making the scenario as realistic as possible. In the scenario, participants are tasked with the surveillance of a set of offshore and onshore assets using a range of unmanned vehicles (UxVs), following standard protocols and responding to alerts. The scenario is designed to be complex enough that the participants would need to rely on the advice of an intelligent planning agent. The agent provided two different plans, from which the operator must either select one as being the best plan, or indicate that neither is suitable.

The scenario included 12 different UxVs with a range of 9 different capabilities comprised of both volatile and nonvolatile capabilities. Volatile capabilities refer to capabilities whose values may change over time; for example, the amount of firepower a UxV possesses decreases upon use. Nonvolatile capabilities refer to capabilities that remain constant throughout the experiment, such as whether the UxV possesses night vision.

Aside from complying with the *asset capabilities*, each possible course of action must also abide by the *rules of engagement*.

The rules of engagement constitute a set of ordered and clearly defined hard and soft constraints that guide the actions needed to respond to any situation.

Participants were asked to deal with 12 tasks, with each task having one or more properties, such as a confirmed or suspected terrorist attack, a confirmed or suspected natural disaster, a confirmed or suspected civilian in distress, general surveillance, and a passenger pickup. To properly satisfy each scenario, the plan needed to comply with a specific set of rules; for example, in the case of a *confirmed terrorist attack*, the hard constraints would include sending three UxVs, one of whom must be aerial. Additional hard constraints would be that two of the UxVs must possess offensive capabilities and must be able to communicate with each other. These conditions were labelled as *hard* constraints because they are essential to completing the scenario and must be complied with. Any plan that did not satisfy all the hard constraints was necessarily unacceptable.

The soft constraints are not essential to satisfying the scenario, although complying with them will accomplish the scenario in the best possible manner. In the case of a *confirmed terrorist attack*, these soft constraints would additionally indicate that it is best for *all* UxVs to possess offensive capabilities and communicate with each other and that the incident be dealt with as soon as possible. Both soft and hard constraints were prioritized, within their own category. Participants were informed that plans should conform with as many soft constraints as possible, but, in the case of conflicts, higher priority soft constraints should be satisfied first.

The role of the participants in this experiment was to work with an intelligent planning agent (the IA) to determine the *best* course of action, where "best"' means to achieve all hard constraints as the first objective, and as many soft constraints as possible, conforming to the soft constraint prioritization order, as the second objective. The experiment incentive involved the surveillance and protection of a set of three oil-rig assets and one drilling rig off the Australian coastline. The IA used was simulated using a Wizard of Oz technique [44]. Each mission encompassed dealing with five or six different mission objectives at once. The mission objectives had to be correctly interpreted from the input which comprised a commanders' statement outlining the situation and several mission alerts.

The IA interpreted the different mission objectives and provided the operator with two candidate plans. What the agent considered to be the best fitting plan was suggested as Plan A and the alternate plan as Plan B. The participants were instructed that the IA could make two types of errors.

1) Incorrectly interpreting or missing some of the initial input; for example, the IA missing input regarding certain weather conditions and then proceeding to allocate UxVs that were unable to operate under those weather conditions.
2) Failure to achieve some of the soft constraints in the rules of engagement due to errors in reasoning; for example, in the case of a confirmed terrorist attack, the IA might compute two plans that both comply with the hard constraints of sending three UxVs, two of which possess offensive capabilities. However, it may fail to compute that the better option is to send three UxVs, *all* possessing offensive capabilities.

Participants were told that the IA always achieved the hard constraints, providing it did not miss or incorrectly interpret some of the input.

The operator's task was either to accept the (recommended) Plan A, reject Plan A and choose Plan B, or choose neither plan if a better plan was possible, whereas taking into consideration the initial intelligence, different asset capabilities, and the rules of engagement. Through the different levels of transparency, the operator was provided with basic information and some of the internal reasoning processes of the agent. Using this information, the operator could determine if an error of any kind has been made.

The experiment was divided to two conditions: half of the participants used DDT, whereas the other half used ST. We hypothesized that the DDT participants would have an advantage when it comes to workload, with a decreased amount of time spent on the experiment, because people would not be forced to view all of the information. However, we expected that DDT would have a negative influence on performance, resulting in decreased success levels, because the participants may commit to certain plans too early.

C. User Interface

The experiment was run using a computer-based simulator designed for the purpose. Fig. 2 presents the user interface. The top left region displays the general input comprising the commander's statement and different alerts. This is the raw input, unaltered by the IA and available to the operator throughout the experiment. The bottom left and right regions display the agents' suggested Plan A and Plan B in two manners: a textual representation listing the actions of the individual agents; and a visual map representation enabling the operator to discern initial UxVs locations and trajectories, illustrated by straight and dotted lines. The dotted lines were used to indicate a secondary trajectory in the event of UxV reuse (that is, a single UxV achieving two separate objectives in a single task).

To facilitate making the decision, the human operator was able to access different levels of transparency by navigating through the different tabs in the top right region of the user interface (see Fig. 2). Initially the "No Knowledge" tab would be selected. Under this configuration, the information available would be the raw input along with the two possible plans with no additional information. The other levels correspond directly to the ETM levels of transparency: *Input* corresponds to Level 1, *Reasoning* to Level 2, and both *Plan 1* and *Plan 2* correspond to Level 3 for each of the possible plan hypotheses suggested by the IA. Fig. 3 presents an example of the Level 3 information interface. Due to the large amount of information to be displayed we presented Level 3 as a table, with different animals representing different UxVs corresponding to their code names and the medical cross's representing UxVs with medical capabilities. The rows represented the different UxVs, whereas the columns represented the different UxV capabilities with regards to each objective. We used color to illustrate additional information,

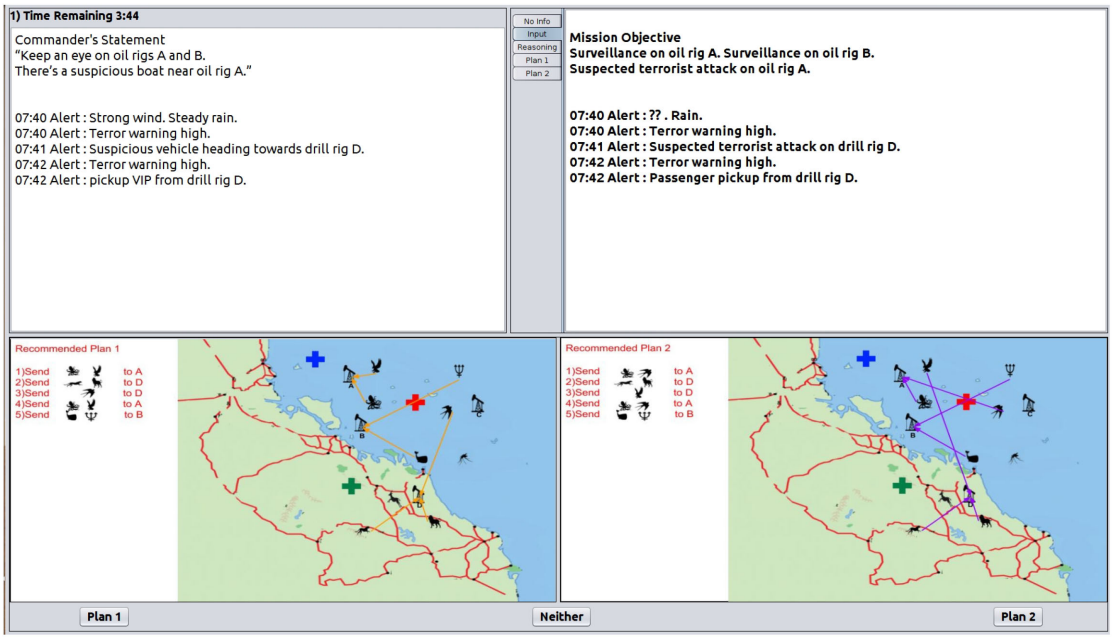
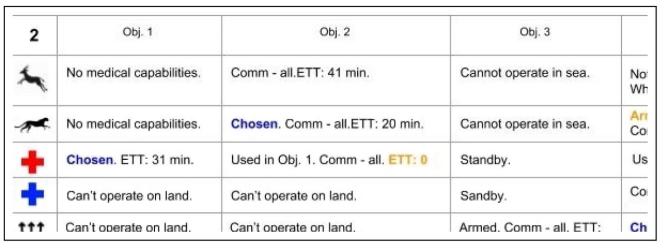


Fig. 2. Experiment user interface. Top left: input. Top right: different transparency levels. Bottom left: Plan A. Bottom right: Plan B.



2	Obj. 1	Obj. 2	Obj. 3	
	No medical capabilities.	Comm - all.ETT: 41 min.	Cannot operate in sea.	No Wh
	No medical capabilities.	**Chosen**. Comm - all.ETT: 20 min.	Cannot operate in sea.	Arr Co
	Chosen. ETT: 31 min.	Used in Obj. 1. Comm - all. ETT: 0	Standby.	Us
	Can't operate on land.	Can't operate on land.	Sandby.	Co
	Can't operate on land.	Can't operate on land.	Armed. Comm - all. ETT:	Ch

Fig. 3. Transparency Level 3 example.

such as information the IA regarded as uncertain in yellow and the UxVs chosen for the mission by the IA in blue.

Both the ST and the DDT participants were initially presented with the "No Knowledge" information and then able to access all of the ETM transparency levels. The differences between the two interfaces arose from the manner in which the information was accessed. The DDT participants could choose to opt out of some levels and could directly influence the order in which these levels were accessed, whereas the SST participants had to access all of the ETM transparency levels in a sequential manner. Participants from both sessions had full control over how much time was spent in each level.

D. Procedure

The experiment procedure started with two sessions of training. The first session familiarized participants with the asset capabilities and rules of engagement. Participants first reviewed hard copies of both the asset capabilities and the rules of engagement, which were available to the participants throughout the experiment. After this, they answered a set of seven questions, about key points from the rules of engagement, to assess their understanding. To be able to continue with the experiment, participants had to successfully answer all of the questions. In case of mistakes they were instructed to read specific clauses in the reading material before attempting to answer the questions again. This session lasted about 30-min after which the participants had a 5-min break.

The second session was designed to familiarize the participants with the user interface and to evaluate their understanding prior to the actual experiment. At this point, the participants were randomly divided into two groups: DDT-based versus ST-based versions of the simulator. Participants viewed four tutorial videos emphasizing each aspect of the user interface, as well as examples of mistakes the IA might make. This was followed by a practice session of three tasks, 4-min per task, in the same format as the actual experiment session. If the participant made a mistake, choosing the wrong plan option, hints were given on the screen as to the correct answer and the task started again. Only after successfully completing all three tasks did the participant proceed to the final evaluation. The evaluation comprised an additional three tasks, 4-min per task. This time participants were not given any feedback after each task. Only those participants that had successfully completed two out of the three tasks were allowed to proceed to the final experiment. Participants who failed more than one of the evaluation tasks were disqualified from further participation.

We now proceed to describe the actual experiment. Each participant was evaluated over 12 tasks, delivered in a random order, with each task having a time limit of four minutes. A failure to respond within the given time limit was considered to be an incorrect response. Table I illustrates the division of the different tasks. In 5 out of the 12 tasks (41.7%) the IA's recommended plan was not the best plan. Two of the errors were attributed to the IA incorrectly recommending Plan A over Plan B, the correct

TABLE I
EXPERIMENT TASK COMPONENTS

Task	Answer	Logic
1	Plan A	The IA recommended the best course of action
2	Neither	The IA missed a weather alert (Level 1)
3	Plan B	Plan B was better optimised with regards to the soft constraint of minimum agent reuse (Level 3)
4	Plan A	The IA recommended the best course of action
5	Neither	Hard constraint violation, the IA did not consider all objectives
6	Plan A	The IA recommended the best course of action
7	Plan A	The IA recommended the best course of action
8	Plan B	Plan B was better optimised with regards to the soft constraint of maximum speed (Level 3)
9	Neither	The IA misunderstood the commander's intent (Level 1)
10	Plan A	The IA recommended the best course of action
11	Plan A	The IA recommended the best course of action
12	Plan A	The IA recommended the best course of action

answer being Plan B. These errors could be first discerned by a human operator given Level 3 transparency information. An additional two errors were a result of missing, or misunderstood initial intelligence in which case the correct answer was *neither*, as both plans did not comply with the rules of engagement. These errors could be first discerned by a human operator given Level 1 transparency information. The last error was a result of an IA error in the hard constraints, in direct opposition to the explicit instructions given to the participants prior to the experiment, in which it was stated that the IA can only err with regard to the soft constraints. We devised this task to evaluate whether the users were so overloaded as to blind them to other aspects of the problem. Since automation misuse may be based on complacency and reflected in an inappropriate checking and monitoring of automated functions, exposing participants to rare automation failures is often examined [45]. Therefore, not being able to detect a violation of the hard constraints could imply *misuse* of the system whereby the user may place excessive reliance on the IA and consequently fail to detect critical IA errors [23].

E. Measures

For both groups, we recorded the following measures.

1) *objective performance*: success rate in selecting the best plan;
2) *hard constraint violation*: success rate in detecting incomplete plans;
3) *completion time*: how quickly participants made their selection (with an upper bound of 4 min).

In DDT, we additionally evaluated the sequence of transparency acquisition as well as which of the different transparency levels were visited. As part of the statistical analysis, we ran a Shapiro–Wilk test on all independent samples to ascertain a normal distribution following which we evaluated significance using a two tail, independent samples, equal variance *t*-test with $\alpha = 0.05$.

At the end of the experiment, the participants were asked to fill out a 10-item summative usability survey [46] and a 12-item trust between people and automation questionnaire [47].

TABLE II
PERFORMANCE PERCENT SUCCESS RESULTS COMPARING RANDOM, DDT, AND ST CONDITION

	Random	ST	DDT
Performance	33.3%	50.6%	55%

TABLE III
COMPARISON OF PERFORMANCE AND EFFICIENCY

	Performance			Response Time		
	M	SD	statistics	M	SD	statistics
DDT	55.00	26.87	$t(28)=0.5$	**55.46**	12.56	$t(28)=1.83$
ST	50.56	21.24	$p=0.62$	64.99	15.80	$p=$**0.04**

Performance measured as percentage correct (higher is better). Efficiency measured as the average response time for each task, expressed as a percentage of the time allocated for that task (lower is better).

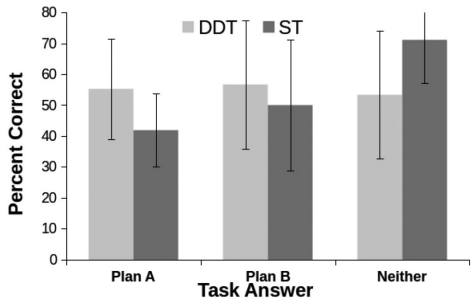


Fig. 4. Percent correct for both the DDT and the ST conditions as a function of whether Plan A, Plan B, or neither plan was the correct answer.

V. RESULTS

In this section, we outline the main results from our evaluation.

A. Performance

Table II compares average mission success percent over all experiments. Both ST and DDT conditions greatly improve on random mission success with 50.6% and 55%, respectively. When comparing between ST and DDT conditions, the DDT participants ranked higher by 5% indicating that the ability to determine which information to examine and at what time while intuitively aiming to increase efficiency may also increase performance.

Table III presents the performance of participants in the DDT and ST sessions, as the average mission success percent over all 12 tasks, higher values indicating better performance. Over the tasks, DDT succeeded on average 5% more than ST but this result is not statistically significant with *p*-value of 0.62.

Throughout the experiment, participants were asked to either agree with the IA and choose *Plan A*, or to disagree with the IA, in which case they could either choose *Plan B* or *Neither*. Fig. 4 displays the distribution over the different possibilities. The *X*-axis denotes the different solution possibilities and the *Y*-axis denotes percent out of all missions whose correct answer corresponded to the option on the *X*-axis. For example, out of all missions whose correct answer was *Plan B*, DDT participants

TABLE IV
STATISTICAL ANALYSIS OF THE AVERAGE PROPORTION OF THE TRIALS FOR WHICH THE PARTICIPANT WAS CORRECT FOR THE DDT VERSUS THE ST CONDITION AS A FUNCTION OF WHETHER PLAN A, PLAN B, OR NEITHER PLAN WAS THE CORRECT ANSWER

	Plan A	Plan B	Neither
t(28)	1.29	0.0.43	1.38
p-value	0.21	0.67	0.09

were successful in 57% and ST participants only answered 50% of those tasks correctly.

Statistical analysis reveals no significant differences among the plans (see Table IV), nonetheless the results were interesting and indicative. Among the plans whose correct answer was *Neither*, the ST participants performed better with 71% success rate versus DDT with only 53% success rate. Among the plans whose correct answer was *Plan A*, the DDT participants outperformed the ST participants with a mean of 55% versus 41%.

B. Hard Constraint Violation

The necessity of visiting every level of transparency in the ST session could potentially result in an increase in the task demand level which may ultimately effect performance [48]. We devised a task to evaluate whether the participants' were still able to discern scenarios which were clearly erroneous. As a prerequisite we clearly stated, within the rules of engagement, that the IA could not err with regards to the hard constraints. We then deliberately inserted a task in which the IA erred with regards to a basic Hard Constraint, only solving four objectives when initially five were given and therefore not completing the mission.

None of the participants in the ST session detected the error. However, 27% of the DDT participants noted the error and gave a correct verbal justification ($t(28)=2.26$, $p=0.03$). Better performance along this measure might be due to the differences in task demands indicating that the amount of information the ST participants were presented with was excessive and detrimental to the performance of the task.

C. Completion Time

Table III presents the efficiency of participants in the DDT and ST sessions. Efficiency was measured by the average time spent on each task, expressed as a percentage of the total time available for that task. As previously mentioned, each experiment was comprised of 12 different tasks, each task with a time limit of 4 min (for a total of at most, 2880 s across all 12 tasks).

The difference in the average response time between the two transparency modes proved to be significant with a p-value of 0.04, with DDT participants only using 55% of the allotted 4 min, on average, with a range of 54.6 sec to 4 min. In contrast, ST participants used an average of 65% of the allotted 4 min., with a range of 2.0–4 min. This reflects a substantial reduction in completion time which may indicate a reduction of workload although we did not measure this directly. In fact, for each task, we counted the number of different ETM levels that were visited by the participants. We found that the ST participants navigated through an average of 9.8 (SD=5.2) different ETM levels per task, whereas the DDT participants navigated through an average of 14.6 (SD=17.5) different ETM levels per task ($t(358)=3.48$, $p=0.001$). So, while the completion time was reduced, significantly more ETM levels were visited by the DDT participants. As discussed later, we found no evidence that visiting more ETM levels increased the workload for the DDT participants. In particular, we found no evidence that DDT participants found the system to be less usable than the ST participants and the DDT participants were in fact better able to spot a violation of a hard constraint suggesting, if anything, a reduction in their workload.

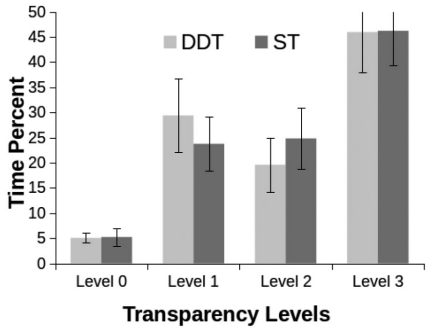


Fig. 5. Time distribution over the different transparency levels. Error bars represent standard deviation.

TABLE V
STATISTICAL ANALYSIS OF THE TIME FRACTION SPENT ON EACH TRANSPARENCY LEVEL FOR THE DDT VERSUS THE ST CONDITION

	Level 0	Level 1	Level 2	Level 3
statistics	$t(28)=1.47$ $p=0.15$	$t(28)=0.19$ $p=0.84$	$t(28)=2.27$ $p=\mathbf{0.03}$	$t(28)=0.64$ $p=0.53$

Fig. 5 summarizes the time distribution over each of the different transparency levels, although we have no clear indication exactly as to which information was focused on, within each level. In both DDT and ST, most time was spent in viewing Level 3, *Plan Projection*. This was understandable as this level had the most information to be conveyed to the user. The least time was spent in the initial level, *No Knowledge*. As this level did not present any transparency information, this was also understandable.

A statistical analysis of the DDT and ST sessions shows that mostly there are no significant differences (see Table V). However, we can see a general indication of DDT participants spending less time on Level 2 transparency, *Input Reasoning*, with only 19% compared with ST with 25%.

D. Insufficient Time

As previously mentioned, participants were allotted 4 min per task. If they were not able to reach a decision within the given time, the task was marked as failed and the next task commenced.

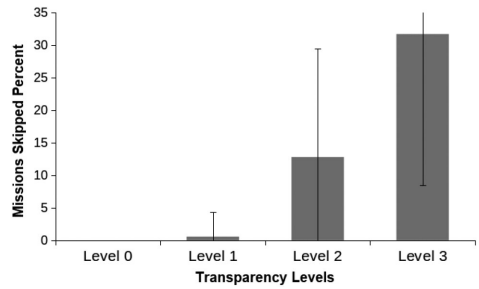


Fig. 6. Distribution percent of the transparency levels not accessed in DDT. Error bars represent standard deviation.

We wanted to evaluate whether the different modes of transparency acquisition affected the tasks in which the participants had insufficient time.

Each session had total of 15 participants with 12 different tasks per experiment for overall 180 tasks. In DDT there was insufficient time for 2 of the tasks, 1.11%, whereas in ST, there was insufficient time for 6 of the tasks, 3.33% ($t(358)=1.43$, $p=0.07$).

E. Unnecessary Transparency Information

The reduction of *completion time* exhibited in the DDT session arises in part from the freedom not to visit every transparency level in the ETM. We evaluated which ETM levels the participants chose to exclude to provide insight into the information deemed important by the user in the presented scenario.

Fig. 6 presents the proportion of tasks in which different ETM levels were not visited. The *Y*-axis denotes the percentage out of all tasks (12 tasks for each of the 15 participants for a total of 180 tasks overall). The *X*-axis denotes the transparency levels not visited.

The initial level, Level 0, is the initial DDT level in which no transparency was conveyed but the two possible plans were presented. This level comprised the initial interface the participants were presented with and therefore could not be skipped. Level 1 was only skipped in one instance, 0.56% of all tasks. Level 2, *Input Reasoning*, was skipped in 23 instances, roughly 13% of all missions.

The last transparency level corresponded to the last level in the ETM which presented the *plan projections* and uncertainties. This level was skipped in 57 tasks, roughly 32% of the missions. These cases include instances in which the participants chose not to examine the additional transparency information for either of the projected plans or both.

One point to consider is that when in DDT, participants who do not trust the system might be led not to acquire any additional information. There was one case study in which the participant viewed only Levels 0 and 1 transparency and acquired no further information. When asked afterwards, the participant voiced a lack of trust in the IA. This mode of action led to poor results with the participant succeeding only in 1 out of the 12 tasks.

TABLE VI
PERCEPTIONS OF USABILITY (HIGHER IS BETTER) AND TRUST (HIGHER IS BETTER)

	Usability			Trust		
	M	SD	statistics	M	SD	statistics
DDT	**2.23**	1.28	$t(28)=0.93$	**4.04**	1.75	$t(28)=0.22$
ST	2.13	1.25	$p=0.25$	3.66	1.39	$p=0.01$

F. Transparency Levels Acquisition Ordering

In DDT, participants were given the freedom to choose which of the levels of transparency to examine and at what times. We wanted to see if that freedom affected the order with which they acquired the different levels of transparency information. For the most part, participants maintained a sequential order of visiting the different transparency levels. In only 36 out of the 180 tasks did the participants choose to visit the levels in a different order, corresponding to only 20% of all tasks. This result further supports the use of the ETM. In the ETM, the order of the transparency levels corresponds to Endsley's model of situation awareness and therefore represents an intuitive, gradual ordering of transparency acquisition.

G. Perceived Usability

We evaluated the effects of DDT on the perceived usability of the system. Participants were presented with the ten-item summative usability survey [46] at the end of the experiment. Table VI presents the results. There were no significant differences between the two modes of transparency acquisition showing that DDT improved efficiency while not unnecessarily encumbering the system.

H. Perceived Trust

To establish a measure of perceived trust in the IA, the participants were presented with the 12-item trust between people and automation questionnaire [47] at the end of the experiment. Participants answered 12 perceived trust questions, allowable answers ranged between 1 and 7 with higher values indicating a higher level of trust. This enabled us to evaluate how DDT would impact trust in the system. Table VI presents the results of the trust questionnaire. The analysis revealed significant differences between the DDT and ST sessions with DDT participants rating their perceived trust in the IA much higher than ST participants.

As mentioned in Section IV-D, the IA errs in 41.7% of the tasks (5 out of 12), thereby being correct 58.3% of the time. The perceived trust associated with the DDT approach was 4.04. Since the trust associated with the DDT approach cannot be directly compared to the IA error rate (given their different units of measurement), it cannot be known with certainty whether the increased perceived trust is commensurate with the IA's success rate. However, the DDT participants higher levels of trust and better performance suggest benefits of the increased trust that deserve further study. Specifically, participants in the DDT condition were both more accurate and faster than those in the ST condition, although only the latter result reached

statistical significance. Further exploration of this finding by evaluating different levels of trustworthiness on the side of the IA remains for future work.

I. Discussion

The results show that giving the operator control over which information they accessed led them to complete the tasks quicker, with a slight (not significant) improvement in performance. The fact that in the ST session, more tasks had been timed out further complements these findings, indicating that ST increases the time necessary to complete each task and illustrating that this may have a detrimental effect on performance.

With regard to the participants perceived trust in the AI, DDT participants rated their trust in the system much higher indicating that trust in the IA's ability to suggest or make decisions increased when participants were able to control the manner in which they acquired the transparency information. We believe that the basis for this finding is in the reduction of *completion time*, resulting from the ability to easily obtain information regarding certain questions instead of being presented with all of the information regardless.

DDT participants rated their trust in the IA higher than the ST participants did. Additionally, DDT participants were more likely to choose Plan A than ST participants who had a higher tendency to reject both plans. These results are consistent with previous studies that suggest that the more the participants trust the IA, the more likely they are to accept the IA's recommendations [4]. Drilling down, we see that participants in the DDT group considered lower level information more important than higher level information in this scenario, choosing not to utilize higher level information more often, even though they typically followed the levels in the sequential order.

We interpret these results as indicating that complex decision-making tasks may be better achieved with *interactive explanation*, which would allow people to interactively pose questions and receive explanations, rather than via predefined transparency models using structured, sequential information presentation. In complex scenarios as studied here, users preferred examining specific information, having to filter through to the information that they wanted. The preference exhibited by users for viewing lower ETM levels is suggestive of users' aversion of overly complex and detailed information. Aware that the domain of explanation is itself the subject of many approaches [49], we acknowledge that more investigation is needed to identify the factors influencing the observed users' preferences, but we observe that our conclusion coincides with that of Miller [35] who stated that people prefer to engage in *contrastive explanations* whereby an event is explained with relation to some other event that may or may not have occurred. Allowing the user to interact with the system does not reduce the availability of complex or detailed information, but does allow the user to readily seek what is relevant for the question *currently* in their mind, for example: *Why is the recommendation to send UXV X rather than Y?* We aim to investigate this and to what degree our results generalize to other situations in future work. We hypothesize that interactive explanation that allows more fine-grained access to information and reasons would be more effective and perceived as more trustworthy.

a) Limitations: The results presented in this article are based on a single experimental case study conducted on 30 trained participants recruited from a pool of undergraduate and graduate students at the University of Melbourne. While our study has sufficient power to show that complex decision-making tasks are better achieved with *interactive explanations* in the particular situations that we considered, to show that our results are more broadly applicable, future work will need to replicate our findings in other situations. Another consideration is that in the presented experiments, the interactive explanation available to participants was coarse grained, with participants only able to interactively choose which transparency levels they wanted to view and in which order. Participants were not able to ask specific questions or request partial information. As can be seen in Fig. 3, Level 3 information, the *Plan Projection*, was quite overwhelming. Better performance might have been achieved had the participants been able to elicit only certain information through some kind of dialogue, enabling them to pose specific questions concerning the unclear aspects of the IA transparency.

b) Future work: In the future, we would further like to directly evaluate the effects of the different transparency acquisition models on user task demand level and also measure the effects of expertise on the side of the participants. In particular, we would like to investigate whether making explanations interactive increases or decreases task demand level, and the degree to which this depends on expertise. In the current study, no expertise was assumed and participants were trained and evaluated to make sure that a certain basic level of proficiency was achieved. We hypothesize that with additional training, participants may be able to better utilize the DDT level and achieve a significantly higher level of task efficiency [50]. Furthermore, it is well known that experts approach problems in their area differently to novices whereby, as a consequence of their extended experience, they can make quicker and more intuitive decisions relying on acquired patterns and familiar planning [51]. Hence, it would be appropriate to explore the relationship between expertise and the need and use of explanations. We would propose to monitor gaze so as to have a better idea of the information utilized by participants.

We hypothesize that expert participants will display a higher level of understanding of the internal reasoning process of the IA and therefore focus on areas in which the IA may potentially make mistakes such as missing alerts while being possibly more prone to IA misuse [23]. We predict that such differences would be more evident with the DDT participants as they have more scope to determine what information they focus on.

VI. Conclusion

In an attempt at increasing human trust in an IA and reducing the interaction time associated with making IAs more transparent, we have presented DDT, a mode of transparency acquisition that allows the human operator to request information as needed and in no particular order as an attempt at alleviating the task demand level.

We contrasted our approach with a baseline approach we refer to as ST in which the human operator must follow all levels of transparency in a particular order. We evaluated both approaches on a highly complex military scenario in which increased transparency would necessarily result in increased task demand level. We also examined different types of errors the IA may make, as well as allowed the possibility of both plans being erroneous.

We have shown that the DDT significantly reduced the response time by roughly 10% when compared with the baseline ST approach, as well as significantly increased the participants trust in the IA, while maintaining the same level of performance. As noted by Mercado *et al.* [6], one might expect the opposite: increasing transparency is likely to provide more information to the operator, which, in turn, might be expected to slow down decision making. Indeed Stowers *et al.* [28] found that although increasing transparency improved performance, it also reduced usability. DDT is noteworthy precisely because it avoids this downside: we are able to obtain some of the benefits of increasing transparency without increasing response times. We have further gained insight into which transparency levels people were mostly interested in, per our experiment settings, and in which order they acquire transparency, when given the freedom to do so. Our findings led us to hypothesize that an interactive explanation would be a better model of transparency than a structured information presentation.

REFERENCES

[1] A. Rosenfeld, N. Agmon, O. Maksimov, and S. Kraus, "Intelligent agent supporting human–multi-robot team collaboration," *Artif. Intell.*, vol. 252, pp. 211–231, 2017.

[2] J. M. Bradshaw *et al.*, "Coordination in human-agent-robot teamwork," in *Proc. Int. Symp. Collaborative Technol. Syst.*, 2008, pp. 467–476.

[3] B. Hardin and M. A. Goodrich, "On using mixed-initiative control: A perspective for managing large-scale robotic teams," in *Proc. Int. Conf. Human-Robot Interact.*, 2009, pp. 165–172.

[4] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[5] J. Y. Chen, M. J. Barnes, J. L. Wright, K. Stowers, and S. G. Lakhmani, "Situation awareness-based agent transparency for human-autonomy teaming effectiveness," in *Proc. SPIE*, 2017, Paper 101941V.

[6] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, and K. Procci, "Intelligent agent transparency in human–agent teaming for multi-UXV management," *Human Factors*, vol. 58, no. 3, pp. 401–415, 2016.

[7] R. F. Kizilcec, "How much information? Effects of transparency on trust in an algorithmic interface," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2016, pp. 2390–2395.

[8] Y. Brand, M. Ebersoldt, D. Barber, J. Y. Chen, and A. Schulte, "Design and experimental validation of transparent behavior for a workload-adaptive cognitive agent," in *Proc. Int. Conf. Intell. Human Syst. Integr.*, 2018, pp. 173–179.

[9] J. B. Lyons, K. S. Koltai, N. T. Ho, W. B. Johnson, D. E. Smith, and R. J. Shively, "Engineering trust in complex automated systems," *Ergonom. Des.*, vol. 24, no. 1, pp. 13–17, 2016.

[10] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? Ways explanations impact end users' mental models," in *Proc. IEEE Symp. Vis. Lang. Human-Centric Comput.*, 2013, pp. 3–10.

[11] J. Mitchell and B. Shneiderman, "Dynamic versus static menus: An exploratory comparison," *ACM SIGCHI Bull.*, vol. 20, no. 4, pp. 33–37, 1989.

[12] A. Cockburn, C. Gutwin, and S. Greenberg, "A predictive model of menu performance," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2007, pp. 627–636.

[13] V. Alvarez-Cortes, V. H. Zárate, J. A. R. Uresti, and B. E. Zayas, "Current challenges and applications for adaptive user interfaces," in *Human-Computer Interaction*. Rijeka, Croatia: InTech, 2009.

[14] J. A. Riascos, L. P. Nedel, and D. C. Barone, "An adaptive user interface based on psychological test and task-relevance," in *Proc. Latin Amer. Workshop Comput. Neurosci.*, 2017, pp. 143–155.

[15] H. Balta *et al.*, "Integrated data management for a fleet of search-and-rescue robots," *J. Field Robot.*, vol. 34, no. 3, pp. 539–582, 2017.

[16] J. Saez-Pons, L. Alboul, J. Penders, and L. Nomdedeu, "Multi-robot team formation control in the guardians project," *Ind. Robot Int. J.*, vol. 37, no. 4, pp. 372–383, 2010.

[17] J. Casper and R. R. Murphy, "Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 33, no. 3, pp. 367–385, Jun. 2003.

[18] J. Chen and P. Terrence, "Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment," *Ergonomics*, vol. 52, no. 8, pp. 907–920, 2009.

[19] P. Squire and R. Parasuraman, "Effects of automation and task load on task switching during human supervision of multiple semi-autonomous robots in a dynamic environment," *Ergonomics*, vol. 53, no. 8, pp. 951–961, 2010.

[20] M. L. Cummings and P. J. Mitchell, "Predicting controller capacity in supervisory control of multiple UAVs," *IEEE Trans. Syst. Man, Cybern. A, Syst. Humans*, vol. 38, no. 2, pp. 451–460, Mar. 2008.

[21] M. Lewis, "Human interaction with multiple remote robots," *Rev. Human Factors Ergonom.*, vol. 9, no. 1, pp. 131–174, 2013.

[22] B. M. Muir, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

[23] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[24] R. D. Sorkin, "Why are people turning off our alarms?" *J. Acoust. Soc. Amer.*, vol. 84, no. 3, pp. 1107–1108, 1988.

[25] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction," *J. Exp. Psychol., Appl.*, vol. 7, no. 3, pp. 171–181, 2001.

[26] C. A. Miller, "Definitions and dimensions of etiquette," In C. Miller (Ed.), Etiquette for Human-Computer Work: Technical Report FS-02-02 (pp. 1–7). Menlo Park, CA: American Association for Artificial Intelligence, 2002, pp. 1–7.

[27] Y. D. Wang and H. H. Emurian, "An overview of online trust: Concepts, elements, and implications," *Comput. Human Behav.*, vol. 21, no. 1, pp. 105–125, 2005.

[28] K. Stowers, N. Kasdaglis, O. Newton, S. Lakhmani, R. Wohleber, and J. Chen, "Intelligent agent transparency: The design and evaluation of an interface to facilitate human and intelligent agent collaboration," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, 2016, vol. 60, no. 1, pp. 1706–1710.

[29] T. Helldin, U. Ohlander, G. Falkman, and M. Riveiro, "Transparency of automated combat classification," in *Proc. Int. Conf. Eng. Psychol. Cogn. Ergonom.*, 2014, pp. 22–33.

[30] T. Helldin, "Transparency for future semi-automated systems: Effects of transparency on operator performance, workload and trust," Ph.D. dissertation, Sch. Sci. Techn., Örebro Universitet, Örebro, Sweden, 2014.

[31] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Human-Comput. Stud.*, vol. 40, no. 1, pp. 153–184, 1994.

[32] L. Rothrock, R. Koubek, F. Fuchs, M. Haas, and G. Salvendy, "Review and reappraisal of adaptive interfaces: Toward biologically inspired paradigms," *Theor. Issues Ergonom. Sci.*, vol. 3, no. 1, pp. 47–84, 2002.

[33] D. N. Kleinmuntz and D. A. Schkade, "Information displays and decision processes," *Psychol. Sci.*, vol. 4, no. 4, pp. 221–227, 1993.

[34] D. Ariely, "Controlling the information flow: Effects on consumers' decision making and preferences," *J. Consum. Res.*, vol. 27, no. 2, pp. 233–248, 2000.

[35] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019. [Online]. Available: https://arxiv.org/abs/1706.07269

[36] S. Penney, J. Dodge, C. Hilderbrand, A. Anderson, and M. Burnett, "Toward foraging for understanding of starcraft agents: An empirical study," in *Proc. 23rd Int. Conf. Intell. User Interfaces*, Mar. 2018, pp. 225–237.

[37] J. Dodge, S. Penney, C. Hilderbrand, A. Anderson, and M. Burnett, "How the experts do it: Assessing and explaining agent behaviors in real-time strategy games," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2018, Paper 562.

[38] Y. Fukuchi, M. Osawa, H. Yamakawa, and M. Imai, "Autonomous self-explanation of behavior for interactive reinforcement learning agents," in *Proc. 5th Int. Conf. Human Agent Interact.*, 2017, pp. 97–101.

[39] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2017, pp. 303–312.

[40] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 32–64, 1995.

[41] J. M. Flach, "Situation awareness: Proceed with caution," *Human Factors*, vol. 37, no. 1, pp. 149–157, 1995.

[42] S. Tremblay, *A Cognitive Approach to Situation Awareness: Theory and Application*. Evanston, IL, USA: Routledge, 2017.

[43] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 1: Alternative perspectives," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 70–73, Jul./Aug. 2006.

[44] L. D. Riek, "Wizard of Oz studies in HRI: A systematic review and new reporting guidelines," *J. Human-Robot Interact.*, vol. 1, no. 1, pp. 119–136, 2012.

[45] J. E. Bahner, A.-D. Hüper, and D. Manzey, "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience," *Int. J. Human-Comput. Stud.*, vol. 66, no. 9, pp. 688–699, 2008.

[46] J. Brooke *et al.*, "SUS—A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.

[47] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cogn. Ergonom.*, vol. 4, no. 1, pp. 53–71, 2000.

[48] C. D. Wickens, "Multiple resources and mental workload," *Human Factors*, vol. 50, no. 3, pp. 449–455, 2008.

[49] G. Gigerenzer, "How to explain behavior?" *Topics Cogn. Sci.*, 2019, pp. 1–19. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12480

[50] R. H. Grabner, A. C. Neubauer, and E. Stern, "Superior performance and neural efficiency: The impact of intelligence and expertise," *Brain Res. Bull.*, vol. 69, no. 4, pp. 422–439, 2006.

[51] K. A. Ericsson, R. R. Hoffman, A. Kozbelt, and A. M. Williams, *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge, U.K.: Cambridge Univ. Press, 2018.