The *Why Agent*

Enhancing user trust in automation through explanation dialog

Rob Cole
Raytheon Company
Intelligence and Information Systems
State College, PA, U.S.A.

Michael J. Hirsch
Raytheon Company
Intelligence and Information Systems
Orlando, FL, U.S.A.

Jim Jacobs
Raytheon Company
Network Centric Systems
Ft. Wayne, IN, U.S.A.

Robert L. Sedlmeyer
Indiana University – Purdue University
Department of Computer Science
Ft. Wayne, IN, U.S.A

***Abstract*— Lack of trust in autonomy is a recurrent issue that is becoming more and more acute as manpower reduction pressures increase. We address the socio-technical form of this trust problem through a novel decision explanation approach. Our approach employs a semantic representation to capture decision-relevant concepts as well as other mission-relevant knowledge along with a reasoning approach that allows users to pose queries and get system responses that expose decision rationale to users. This representation enables a natural, dialog-based approach to decision explanation. It is our hypothesis that the transparency achieved through this dialog process will increase user trust in autonomous decisions. We tested our hypothesis in an experimental scenario set in the maritime autonomy domain. Participant responses on psychometric trust constructs were found to be significantly higher in the experimental group for the majority of constructs, supporting our hypothesis. Our results suggest the efficacy of incorporating a decision explanation facility in systems for which a socio-technical trust problem exists or might be expected to develop.**

Keywords-Semantic modeling; Maritime Autonomy; Trust in Autonomy; Decision Explanation.

I. INTRODUCTION

Large organizations such as the Department of Defense rely heavily on automation as a means of ensuring high-quality product, as well as cost control through manpower reduction. However, lack of user trust has repeatedly stood in the way of widespread deployment. We have observed two fundamental forms of the problem: the technical and the socio-technical form. The technical form is characterized by user reservations regarding the ability of a system to perform its mission due to known or suspected technical defects. For example, an automated detection process might have a very high false positive rate, conditioning operators to simply ignore its output. Trust in such a situation can only be achieved by addressing the issue of excessive false detections, a technical problem suggesting a purely technical solution. As another example, consider a situation in which automation is introduced into a purely manual process characterized by decision making in high-pressure situations. In such a situation, operators might reject automation in favor of the trusted, manual process for purely non-technical reasons. In other words, in the absence of any specific evidence of limitations of the automation, the automation could nonetheless be rejected for reasons stemming from the social milieu in which the system operates. This is the socio-technical form of the problem.

One might address the socio-technical problem through education: train the operators with sufficient knowledge of system specifications and design detail to erase doubts they may have regarding the automation. Such an approach is costly since every operator would have to be trained to a high degree. Operators would essentially have to be system specialists. Instead, we propose an approach intended for non-specialist operators, stemming from the insight that the socio-technical trust problem results from a lack of insight into system decision rationale. If an operator can be made to understand the *why* of system behavior, that operator can be expected to trust the system in the future to a greater degree, if the rationale given to the operator makes sense in the current mission context.

Explanation mechanisms in expert systems have focused on the use of explicit representations of design logic and problem solving strategies [1]. The early history of explanation in expert systems saw the emergence of three types of approaches, as described in Chandrasekaran, Tanner, and Josephson [2]. Type I systems explain how data matches local goals. Type 2 systems explain how knowledge can be justified [3]. Type 3 systems explain how control strategy can be justified [4]. A more detailed description of these types is given by Saunders and Dobbs [5, p. 1102]:

> Type 1 explanations are concerned with explaining why certain decisions were or were not made during the execution (runtime) of the system. These explanations use information about the relationships that exist between pieces of data and the knowledge (sets of rules for example) available for making specific decisions or choices based on this data. For example, Rule X fired because Data Y was found to be true.
>
> Type 2 explanations are concerned with explaining the knowledge base elements themselves. In order to do this, explanations of this type must look at knowledge about

This research was supported by Raytheon Corporate IR&D.

knowledge. For example, knowledge may exist about a rule that identifies this rule (this piece of knowledge) as being applicable ninety percent of the time. A type 2 explanation could use this information (this knowledge about knowledge) to justify the use of this rule. Other knowledge used in providing this type of explanation consists of knowledge that is used to develop the ES but which does not affect the operation of the system. This type of knowledge is referred to as deep knowledge.

Type 3 explanations are concerned with explaining the runtime control strategy used to solve a particular problem. For example, explaining why one particular rule (or set of rules) was fired before some other rule is an explanation about the control strategy of the system. Explaining why a certain question (or type of question) was asked of the user in lieu of some other logical or related choice is another example. Therefore, type 3 explanations are concerned with explaining how and why the system uses its knowledge the way it does, a task that also requires the use of deep knowledge in many cases.

Design considerations for explanations with dialog are discussed in a number of papers by Moore and colleagues ([6], [7], [8] and [9]). These papers describe the explainable expert systems (EES) project which incorporates a representation for problem-solving principles, a representation for domain knowledge and a method to link between them. In Moore and Swartout [6], hypertext is used to avoid the referential problems inherent in natural language analysis. To support dialog with hypertext, a planning approach to explanation was developed that allowed the system to understand what part of the explanation a user is pointing at when making further queries. Moore and Paris [8] and Carenini and Moore [9] discuss architectures for text planners that allow for explanations that take into account the context created by prior utterances. In Moore [10], an approach to handling badly-formulated follow-up questions (such as a novice might produce after receiving an incomprehensible explanation from an expert) is presented that enables the production of clarifying explanations. Tanner and Keuneke [11] discuss an explanation approach based on a large number of agents with well-defined roles is described. A particular agent produces an explanation of its conclusion by ordering a set of text strings in a sequence that depends on the decision's runtime context. Based on an explanation from one agent, users can request elaboration from other agents.

Weiner [12] focuses on the structure of explanations with the goal of making explanations easy to understand by avoiding complexity. Features identified as important for this goal include syntactic form and how the focus of attention is located and shifted. Eriksson [13] examines answers generated through transformation of a proof tree, with pruning of paths, such as non-informative ones. Millet and Gilloux [14] describe the approach in Wallis and Shortliffe [15] as employing a user model in order to provide users with explanations tailored to their level of understanding. The natural language aspect of explanation is the focus of Papamichail and French [16], which uses a library of text plans to structure the explanations.

In Carenini and Moore [17], a comprehensive approach toward the generation of evaluative arguments (called GEA) is presented. GEA focuses on the generation of text-based arguments expressed in natural language. The initial step of GEA's processing consists of a text planner selecting content from a domain model by applying a communicative strategy to achieve a communication goal (e.g. make a user feel more positively toward an entity). The selected content is packaged into sentences through the use of a computational grammar. The underlying knowledge base consists of a domain model with entities and their relationships and an additive multi-attribute value function (a decision-theoretic model of the user's preferences).

In Gruber and Gautier [18] and Gautier and Gruber [19] an approach to explaining the behavior of engineering models is presented. Rather than causal influences that are hard-coded [20], this approach is based on the inference of causal influences, inferences which are made at run time. Using a previously developed causal ordering procedure, an influence graph is built from which causal influences are determined. At any point in the influence graph, an explanation can be built based on the adjacent nodes and users can traverse the graph, obtaining explanations at any node.

Approaches to producing explanations in MDPs are proposed in Elizalde et al. [21] and Khan, Poupart and Black [22]. Two strategies exist for producing explanations in BNs. One involves transforming the network into a qualitative representation [23]. The other approach focuses on the graphical representation of the network. A software tool called Elvira is presented which allows for the simultaneous display of probabilities of different evidence cases along with a monitor and editor of cases, allowing the user to enter evidence and select the information they want to see [24].

An explanation application for JAVA debugging is presented in Ko and Myers [25]. This work describes a tool called *Whyline* which supports programmer investigation of program behavior. Users can pose "*why did*" and "*why didn't*" questions about program code and execution. Explanations are derived using a static and dynamic slicing, precise call graphs, reachability analysis and algorithms for determining potential sources of values.

Explanations in case-based reasoning systems are examined as well. Sørmo, Cassens, and Aamodt [26] present a framework for explanation and consider specific goals that explanations can satisfy which include transparency, justification, relevance, conceptualization and learning. Kofod-Petersen and Cassens [27] consider the importance of context and show how context and explanations can be combined to deal with the different types of explanation needed for meaningful user interaction.

Explanation of decisions made via decision trees is considered in Langlotz, Shortliffe, and Fagan [28]. An explanation technique is selected and applied to the most significant variables, creating a symbolic expression that is converted to English text. The resulting explanation contains no mathematical formulas, probability or utility values.

Lieberman and Kumar [29] discuss the problem of mismatch between the specialized knowledge of experts

providing help and the naiveté of users seeking help is considered. Here, the problem consists of providing explanations of the expert decisions in terms the users can understand. The *SuggestDesk* system is described which advises online help personnel. Using a knowledgebase, analogies are found between technical problem-solution pairs and everyday life events that can be used to explain them.

Bader et al. [30] use explanation facilities in recommender systems to convince users of the relevance of recommended items and to enable fast decision making. In previous work, Bader found that recommendations lack user acceptance if the rationale was not presented. This work follows the approach of Carenini and Moore [17].

In Pu and Chen [31], a "*Why?*" form of explanation was evaluated against what the researchers termed an Organized View (OV) form of explanation in the context of explanations of product recommendations. The OV approach attempts to group decision alternatives and provide group-level summary explanations, e.g. "these are cheaper than the recommendation but heavier." A trust model was used to conduct a user evaluation in which trust-related constructs were assessed through a Likert scale instrument. The OV approach was found to be associated with higher levels of user trust than the alternative approach.

The important of the use of context in explaining the recommendations of a recommendation system was investigated in Baltrunas et al. [32]. In this study of point-of-interest recommendation, customized explanation messages are provided for a set of 54 possible contextual conditions (e.g. "this place is good to visit with family"). Even where more than one contextual condition holds and is factored into the system's decision, only one can be utilized for the explanation (the most influential one in the predictive model is used). Only a single explanatory statement is provided to the user.

Explanation capabilities have also been shown to aid in increasing user satisfaction with and establishing trust in complex systems [34, 35, 36]. The key insight revealed by this research is the need for *transparency* in system decision-making. As noted by Glass et al., "users identified explanations of system behavior, providing transparency into its reasoning and execution, as a key way of understanding answers and thus establishing trust. [37]" Dijkstra [38] studied the persuasiveness of decision aids, for novices and experts. In one experiment, lawyers examined the results of nine legal cases supported by one out of two expert systems. Both systems had incomplete knowledge models. Because of the incomplete models, the expert systems routinely gave opposite advice on each legal case. This resulted in the lawyers being easily mis-led. Therefore, adequate explanation facilities and a good user-interface must provide the user with the transparency needed to make the decision of trusting the system. Rieh and Danielson [39] Outline four different explanation types of decision aids. Line-of-reasoning explanations provide the logical justification of the decision; justification explanations provide extensive reference material to support the decision; control explanations provide the problem-solving strategy to arrive at the decision; and terminological explanations provide definition information on the decision. In each case, the amount of transparency in the decision-making process is a factor in the trust of the user.

Our approach to providing transparency, the *Why Agent*, is a decision explanation approach incorporating dialog between the user and the system. Rather than attempting to provide monolithic explanations to individual questions, our dialog-based approach allows the user to pose a series of questions, the responses to which may prompt additional questions. Imitative of natural discourse, our dialog approach allows a user to understand the behavior of the system by asking questions about its goals, actions or observables and receiving responses couched in similar terms. We implemented our approach and conducted an evaluation in a maritime autonomy scenario. The evaluation consisted of an experiment in which two versions of an interface were shown to participants who then answered questions related to trust. Results of the experiment show response scores statistically consistent with our expectations for the majority of psychometric constructs tested, supporting our overall hypothesis that transparency fosters trust. The rest of this paper is organized as follows. Section II describes the problem domain and the technical approach. Experiments and results are presented in Section III. In Section IV, we provide some concluding remarks and future research directions.

II. TECHNICAL APPROACH

A. Domain Overview

Our approach to demonstrating the Why Agent functionality and evaluating its effectiveness consisted of a simulation-based environment centered on a maritime scenario defined in consultation with maritime autonomy SMEs. The notional autonomous system in our scenario was the X3 autonomous unmanned surface vehicle (AUSV) by Harbor Wing Technologies [1]. Raytheon presently has a business relationship with this vendor in which we provide ISR packages for their AUSVs.

The X3 was of necessity a notional AUSV for our demonstration because the actual prototype was not operational at the time of the Why Agent project. For this reason, a live, on-system demonstration was not considered. Instead, our demonstration environment was entirely simulation-based. An existing route planning engine developed under Raytheon research was modified to serve as the AUSV planner. Additional code was developed to support the simulation environment and Why Agent functionality, as described below.

B. Software Architecture

Our software architecture consists of four components interacting in a service-oriented architecture, as shown in Figure 1.

The Planner component performed route planning functions based on a plan of intended movement. A plan of intended movement is input in the form of a series of waypoints. These waypoints, along with environmental factors, such as weather forecast data, are used in the planning algorithm to determine

[1] http://www.harborwingtech.com

an actual over-ocean route. The planner was a pre-existing component developed on R&D that the Why Agent leveraged for the demonstration. Modifications made to the planner to support the Why Agent project include changes to expose route change rationale to the controller and inform the controller of weather report information.

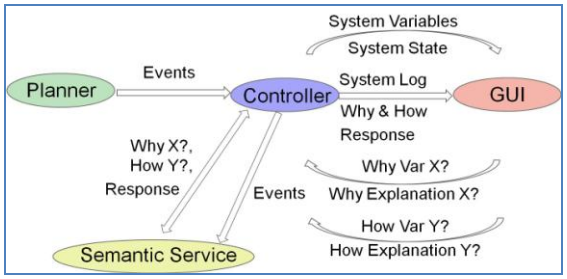

Figure 1: SW architecture for Why Agent.

The Controller represents the embodiment of the majority of the simulated AUSV decision logic and simulation control logic. Because we did not employ an actual AUSV for the Why Agent project, much of the decision logic of an actual AUSV had to be simulated for our demonstration, logic implemented in the Controller. The input to the Controller consisted of a test control file that defined the event timeline for the simulation. In addition to orchestrating simulation events defined in the control file, the Controller mediated queries and responses between the user interface and the semantic service.

The graphical user interface was implemented as a web application. Two versions of the GUI were developed, one with and one without the Why Agent explanation facility. The Why Agent version is shown in Figure 2. It has four screen regions: a map, a status panel, a log data panel and an explanation panel. The map, implemented with Google Map technology, shows the current location and route of the AUSV. The status panel shows various AUSV status values, such as location, speed, current mode, etc. The log panel shows a time-stamped series of event descriptions. Various items in the log panel are user-selectable and have context-sensitive menus to support the user interface functionality of the Why Agent facility. When a user makes a selection, the response from the semantic service is shown in the bottom (explanation) panel. Additionally, responses in the explanation panel are also selectable for further queries. In this manner, the user can engage in a dialog with the system.

The semantic service contains the knowledgebase underlying the decision rationale exposed by the Why Agent. The knowledge consists of event and domain ontology models represented in web ontology language (OWL) format. The semantic service provides responses to queries from the controller through queries against its underlying models.

An example of a domain model is shown in Figure 3. Relationships in this figure encode potential queries linking concepts and events that can be displayed in the user interface. For example, the activity *ConductPatrol* relates to the function *MissionExecution* through the relationship *servesPurpose*. This relationship is statically associated with the query why? at the user level. Thus, the existence of this link connected with the node *ConductPatrol* implies a why? option being made available to the user in the context-sensitive menu for the *ConductPatrol* item. When the user selects the *ConductPatrol* item and the associated *why?* option, a query is generated that contains IDs associated with the *ConductPatrol* node and the *servesPurpose* link. The linked node, in this case *MissionExecution*,is then returned to the user as the result of a query against the associated OWL model.

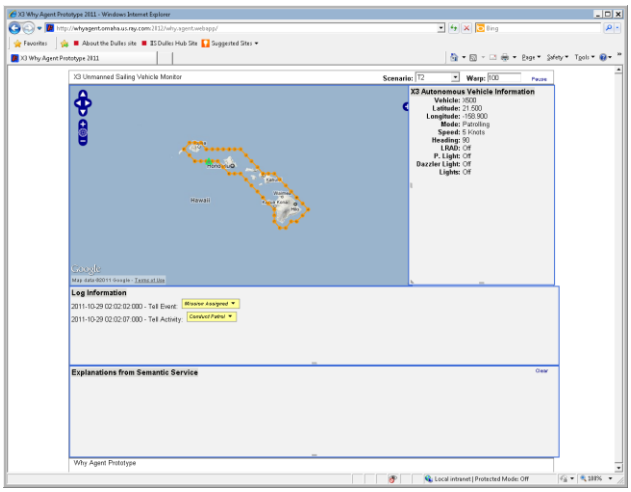

Figure 2: General GUI for Why Agent interface.

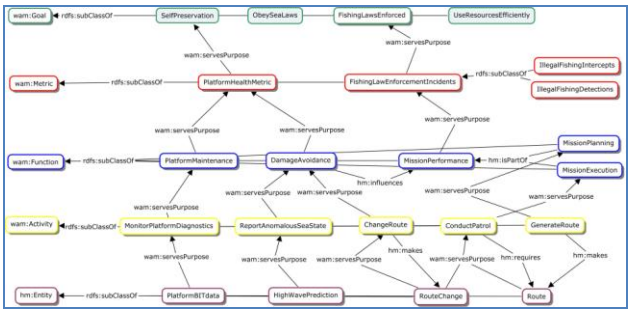

Figure 3: Example domain model.

III. EXPERIMENTATION

Our evaluation approach consisted of an experiment in which the Why Agent was the treatment. Two versions of a prototype operator interface were developed. One version incorporated the Why Agent functionality and the second did not. The two versions were otherwise identical. Screenshots of the two interface versions are presented in Figures 4 and 5.

A. *Demonstration Scenario*

The demonstration scenario consisted of autonomous fishing law enforcement in the Northwestern Hawaiian Islands Marine National Monument. The CONOP for this mission is as follows:

- The AUSV operator selects waypoints corresponding to a patrol area.
- The AUSV route planner finds a route through the waypoints and a patrol is conducted.

- RADAR is used to detect potential illegal fishing vessels (targets)
- Targets are investigated visually after AUSV closes to an adequate proximity.
- Automated analysis of the visual data is used to confirm the target is engaged in illegal fishing.
- Targets engaged in illegal activity are visually identified for subsequent manned enforcement action.

Non-lethal self-defensive actions can be taken by the AUSV in the presence of hostile targets.

To support this demonstration, a software-based simulation environment was developed. The demonstration consisted of capturing video of user interactions with the baseline and Why Agent versions of the operator interface while a scripted series of events unfolded over a pre-determined timeline.

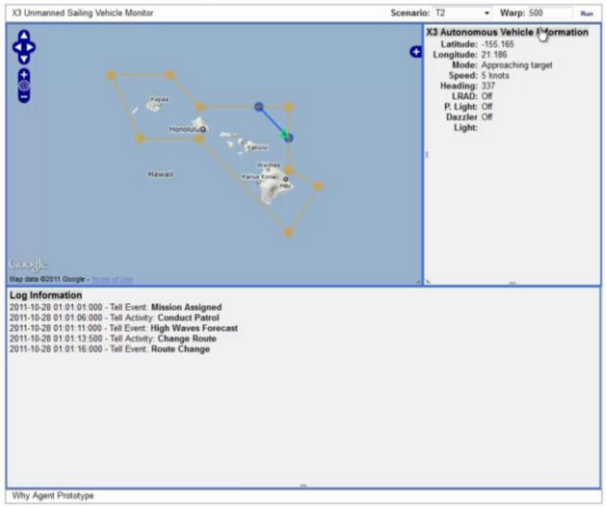

Figure 4: Operator interface without the Why Agent functionality.

B. *Experimental Design*

Our experiment consisted of a single-factor, randomized design. The factor is interface type and has two levels: baseline (control) and Why Agent (experimental). Thus, we have two treatment levels, corresponding to the two factor types. The experimental subjects were Raytheon employees, recruited across multiple Raytheon locations, during the project.

Our general hypothesis is that **the Why Agent fosters a more appropriate level of trust in users than the baseline system**. By utilizing the information provided by the Why Agent, users will be more able to calibrate their trust [33]. To test this hypothesis, we needed to operationalize the concept of "more appropriate level of trust" and thereby derive one or more testable hypotheses. We accomplished this through the following operationalization.

Trust in a particular system, being an unobservable mental aspect of a user, necessitates the use of psychometric readings of constructs related to the overall concept of trust. Given the broad nature of this concept, multiple constructs should be defined. Using our domain insight and engineering judgment, we selected the following set of five psychometric constructs: 1. General Competence, 2) Self-Defense, 3) Navigation, 4) Environmental Conservation and 5) Mission. Each construct is intended to capture the users' belief regarding the system's ability to effectively perform in regard to that construct, i.e. the user's level of trust for that construct. For example, the construct *Mission* attempts to encompass user attitudes toward the ability of the system to successfully execute its mission. The Environmental Conservation construct was included as an example of a construct under which we would not expect to see a difference in psychometric responses.

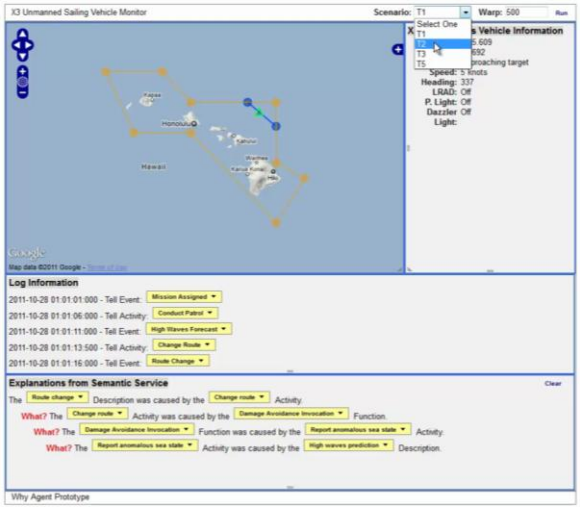

Figure 5: Operator interface with the Why Agent functionality.

For each construct, we have a set of possible trust levels and a set of psychometric participant response scores. Define these as follows (for this study, k=5):

- Set of k constructs C = $\{c_j : 1 \leq j \leq k\}$
- Set of trust levels L = {*low*, *high*}
- Psychometric participant response scores for each construct:

  Control: $R^C = \{r_j^C : 1 \leq j \leq k\}$

  Experimental: $R^E = \{r_j^E : 1 \leq j \leq k\}$

Here, we take the simplest possible approach, a binary trust level set. We simply assume that the trust level for a particular construct should either be low or high, with nothing in between. Clearly, many other trust models are possible. To operationalize the notion of "more appropriate level of trust", we need to define, for each construct, a ground truth assignment of trust level. Thus, we need to define the following mapping T:

- Mapping of construct to trust level: $T(j) \in$ L
  - $T(j) = low$: People *should not* trust the system regarding construct j
  - $T(j) = high$: People *should* trust the system regarding construct j.

Additionally, we need to map the elements of the trust set to psychometric scale values. In other words, we need to normalize the scale as follows:

- Mapping of trust level to psychometric scale values

$$S: S(low) = 1; S(high) = 5.$$

At this point, we can define the concept of "appropriate level of trust" in terms of the psychometric scale through a composition of the above mappings *S* and *T*. In other words, for each construct, the appropriate level of trust is the psychometric value associated with the trust level assigned to that construct:

- Appropriate Level of Trust *with respect to design intent* $A = \{a_j : 1 \leq j \leq k\}$

For each construct c_j, the appropriate level of trust a_j for that construct is given by

$$a_j = S(T(j)),\ 1 \leq j \leq k \quad (1)$$

A key aspect of the above definition is the qualifier *with respect to design intent*. We assume the system functions without defects. *With respect to design intent* simply means "it should be trusted to accomplish X if it is designed to accomplish X." We make this assumption for simplification purposes, fully acknowledging that no real system is defect-free. In the presence of defects, the notion of *appropriate level of trust* becomes more complex.

Having defined *appropriate level of trust*, we are finally in a position to define the key concept, *more appropriate level of trust.* The intuition underlying this notion is the observation that if one's trust level is not appropriate to begin with, any intervention that moves the trust level toward the appropriate score by a greater amount than some other intervention can be said to provide a "more" appropriate level of trust. The Why Agent specifically exposes information associated with the purpose of AUSV actions. Such additional information serves to build trust [33]. If the psychometric score for the experimental group is closer to the appropriate trust level than the score for the control group, then we can say that the experimental treatment provided a more appropriate level of trust for that construct. Formally, we define this concept as follows:

- More appropriate level of trust: Given observed response scores r_j^C and r_j^E for construct *j*, the experimental response r_j^E reflects a more appropriate level of trust when the following holds

$$r_j^E - r_j^C < 0 \text{ if } a_j = 1 \quad (2)$$
$$r_j^E - r_j^C > 0 \text{ if } a_j = 5 \quad (3)$$

We expect the Why Agent to affect observed trust levels only for those constructs for which relevant decision criteria are exposed during the scenario. In these cases, we expect Equations (2)-(3) to hold. In all other cases, we do not. For example, since the AUSV is not designed to protect marine life, we assert that the appropriate level of trust for the Environmental Conservation construct is "low." However, we do not expect to observe response levels consistent with Equations (2) – (3) unless dialog exposing decision rationale relevant to this concept is included in the scenario.

Based on this reasoning, we expect the effect of decision explanation to be one of pushing response scores up or down, toward the appropriate trust level but only in cases where explanation dialog related to the construct under test is exposed. In other cases, we expect no difference in the response scores, as indicated in Table 1. We note that the null hypotheses are derived as the complementary sets to the equations in Table 1. E.g., the 'low, with relevant dialog' null hypothesis equation would be $r_j^E - r_j^C \geq 0$.

A total of 44 control and 50 experimental subjects were recruited for the Why Agent study. The experiment was designed to be completed in one hour. Following a short orientation, a pre-study questionnaire was presented to the participants. The pre-study questionnaire contained questions regarding participant demographics and technology attitudes. The purpose of the pre-study questionnaire was to determine whether any significant differences existed between the experimental and control groups. Following the pre-study questionnaire, participants were given a short training regarding the autonomous system and their role in the study. Participants were asked to play the role of a Coast Guard commander considering use of the autonomous system for a drug smuggling interdiction mission. Following the training, participants were shown the scenario video which consisted of several minutes of user interaction with either the baseline or Why Agent interface. Following the video, participants completed the main study questionnaire. The system training was provided in a series of powerpoint slides. Screenshots taken from the study video were provided to the participants in hardcopy form, along with hardcopies of the training material. This was done to minimize any dependence on memory for participants when completing the study questionnaire.

Table 1: Expected responses as a result of decision explanation.

		Experimental Condition	
		With relevant dialog	Without relevant dialog
Construct trust level	Low	Experimental response less than control response $r_j^E - r_j^C < 0$	Experimental response indistinguishable from control response $r_j^E - r_j^C = 0$
	High	Experimental response greater than control response $r_j^E - r_j^C > 0$	Experimental response indistinguishable from control response $r_j^E - r_j^C = 0$

C. Experimental Results

To investigate whether significant differences exist between the control and experimental groups in terms of responses to the technology attitudes questions, ANOVA was performed. The results are shown in Table 2. Cronbach reliability coefficients, construct variances and mean total response scores are shown for the control and experimental groups in Tables 3 and 4.

To investigate whether significant differences exist between the control and experimental groups in terms of responses to the study questions, ANOVA was performed. For this study,

we focused our analysis on individual constructs. Thus, we do not present any statistics on, for example, correlations among responses related to multiple constructs for either the control or experimental group. The results are shown in Table 6.

Table 2: ANOVA computations analyzing differences between control and experimental groups, for technology attitude questions.

	Q7		Q8		Q9		Q10	
Source	F	Prob>F	F	Prob>F	F	Prob>F	F	Prob>F
Group	0.0331	0.8566	0.702	0.4072	0.4337	0.5141	0.0271	0.87
Gender	0.8431	0.3642	1.5875	0.2152	0.1515	0.6992	0.1713	0.6813
Age	0.8458	0.6845	0.3467	0.9986	0.9452	0.5614	0.8033	0.7359
LaborGrade	1.5714	0.1461	0.988	0.4736	0.49	0.8981	0.3926	0.9509
MilitaryExp	0.1589	0.6924	0.0121	0.9129	0.0252	0.8747	0.04	0.8424
OperatorExp	1.0264	0.3173	0.1311	0.7193	0.1389	0.7113	0.3116	0.5799
MaritimeExp	0.7264	0.4901	0.7418	0.4828	0.9425	0.3983	0.1557	0.8563

Table 3: Cronbach reliability coefficients, construct variances, and means for control group.

Control Results						
	Variances					
Construct	Q1	Q2	Q3	Total	Cronbach Alpha	Mean
1	0.492	0.306	0.348	2.20	0.72	11.11
2	0.710	0.517	NA	1.79	0.63	6.43
3	0.720	0.319	NA	1.05	0.02	7.30
4	0.911	0.670	NA	2.02	0.43	6.73
5	0.953	0.586	NA	2.23	0.62	7.34

T-test results for each construct are shown in Table 5. Two p-values are shown for each construct; p1 represents the p-value resulting from use of the pooled variance while p2 represents the p-value resulting from use of separate variances.

The ANOVA results shown in Table 2 indicate that the experimental and control groups did not significantly differ across any attribute in terms of their responses to the technology attitudes questions. In other words, we do not see any evidence of a technology attitude bias in the study participants.

Table 4: Cronbach reliability coefficients, construct variances, and means for control group.

Experimental Results						
	Variances					
Construct	Q1	Q2	Q3	Total	Cronbach Alpha	Mean
1	0.286	0.262	0.449	1.94	0.73	12.06
2	0.689	0.694	NA	2.18	0.73	7.22
3	0.480	0.367	NA	1.17	0.56	7.64
4	0.571	0.621	NA	1.92	0.76	7.14
5	0.898	0.629	NA	2.05	0.51	7.46

Table 5: T-test computations for each construct.

Construct Hypothesis Tests				
	p-values			
Construct	p1	p2	Null Hypothesis	Result
1	0.001	0.001	Experimental score is not greater than Control score	Reject Null Hypothesis
2	0.004	0.004	Experimental score is not greater than Control score	Reject Null Hypothesis
3	0.058	0.059	Experimental score is not greater than Control score	Accept Null Hypothesis
4	0.158	0.159	Experimental score is equal to Control score	Accept Null Hypothesis
5	0.348	0.347	Experimental score is not greater than Control score	Accept Null Hypothesis

For constructs one and two, the experimental response was greater than the control response ($p = 0.001$ and 0.004, respectively), consistent with our expectations. For construct four, environmental conservation, we see no significant difference between the experimental and control responses ($p = 0.16$), which is also consistent with our expectations as this construct had no associated decision explanation content exposed to the experimental group. The experimental response for construct 3 was not significantly higher than the control response, which is inconsistent with our expectations, although the difference is only marginally outside the significance threshold ($p = 0.059$).

Table 6: ANOVA computations analyzing differences between control and experimental groups, for study questions.

	C1		C2		C3		C4		C5	
Source	F	Prob>F	F	Prob>F	F	Prob>F	F	Prob>F	F	Prob>F
Group	7.7396	0.0083	6.4742	0.015	2.7356	0.1062	1.443	0.2369	0.0993	0.7543
Gender	0.1312	0.7191	0.1283	0.7221	4.5119	0.0401	0.1231	0.7276	6.11E-04	0.9804
Age	0.5682	0.9481	0.6485	0.8944	0.5087	0.9738	0.8843	0.6368	0.7259	0.8225
LaborGrade	0.7763	0.6609	0.8258	0.6156	1.1735	0.3362	1.2032	0.3171	1.1047	0.3834
MilitaryExp	0.3176	0.5763	0.2463	0.6225	0.6621	0.4208	3.9111	0.0551	1.4172	0.2411
OperatorExp	1.4785	0.2313	0.0079	0.9295	0.6746	0.4164	1.0084	0.3215	4.0839	0.0502
MaritimeExp	0.0919	0.9124	0.7755	0.4675	0.424	0.6574	0.0783	0.9248	1.6753	0.2005

While the test results indicate moderate support for the efficacy of the Why Agent approach, they are decidedly mixed, so it is not possible to draw any definitive conclusions. As discussed below, we recognize that a number of significant limitations also hinder the application of our results. A pilot study would have helped to create a stronger experimental design and recruit a more representative sample population, but this was not possible due to budget and schedule constraints. Nevertheless, the study has provided initial evidence for how and to what extent the Why Agent approach might influence trust behavior in autonomous systems, and given impetus for continued investigations.

Construct Reliability: Referring to Table 4, we see that reliability coefficients for some constructs are not above the commonly-accepted value of 0.7. Had schedule permitted, a pilot study could have uncovered this issue, providing an opportunity to revise the questionnaire.

Experiment Limitations: Clearly a variety of limitations apply to our experiment. One is that participants did not interact directly with the system interface; instead entire groups of participants were shown a video of someone else interacting with the system. Also, the participants were not drawn from the population of interest. Consequently, our results may not apply to that target group. Additionally, subjects were asked to play a role with much less information than a real person in that role would have. Also, as noted by a reviewer, the experimental design does not allow us to determine whether decision correctness is related to trust when clearly it should be; an intervention that raises trust regardless of correctness is not desirable. Finally, execution of the experiment could have been improved. In particular, our maritime autonomy SME noted: The Mode should have reflected the simulation events; The LRAD light should have illuminated during the approach phase with an audio warning; The subjects should have been trained on the nonlethal defense functions.

Semantic Modeling: A potentially significant drawback to our approach is the manually-intensive nature of the semantic modeling effort needed to populate our knowledgebase. Identifying ways to automate this process is a key area of potential future work related to this effort.

IV. Concluding Remarks

We draw the following specific conclusions based on the quantitative results reported above. First, the experimental and control groups do not significantly differ across any attribute in terms of their responses to the technology attitudes questions. The experimental and control groups do not significantly differ across any non-Group attribute in terms of their responses to the study questions with the exception of gender differences for construct. Construct reliability is low in some cases, indicating the need for a prior pilot study to tune the psychometric instrument. We accept the null hypothesis for construct 4 and reject it for constructs 1 and 2, as predicted under our assumptions. We cannot reject the hypothesis associated with construct 3, although this is a very marginal case. The results of construct 5 are contradictory to our expectations. Overall, we conclude that the Why Agent approach does increase user trust levels through decision transparency.

References

[1] B. Chandrasekaran and W. Swartout, "Explanations in knowledge systems: the role of explicit representation of design knowledge," *IEEE Expert* vol. 6, no. 3, pp. 47-19, 1991.

[2] B. Chandrasekaran et al., "Explaining control strategies in problem solving," *IEEE Expert* vol. 4, no.1, pp. 9-15, 1989.

[3] William R. Swartout. "XPLAIN: a system for creating and explaining expert consulting programs," *Artificial Intelligence*, vol. 21, no. 3, pp. 285-325, 1983.

[4] William J. Clancey, "The epistemology of a rule-based expert system — a framework for explanation," *Artificial Intelligence*, vol 20., no. 3, pp. 215-251, 1983.

[5] V. M. Saunders and V. S. Dobbs, "Explanation generation in expert systems," in *Proceedings of the IEEE 1990 National Aerospace and Electronics Conference*, vol. 3, pp. 1101-1106, 1990.

[6] J. Moore and W. Swartout, "Pointing: A Way Toward Explanation Dialog," *AAAI Proceedings*, pp.457-464, 1990.

[7] Swartout et al., 1991. "Explanations in knowledge systems: design for explainable expert systems," *IEEE Expert*, vol. 6, no. 3, pp. 58-64, 1991.

[8] Johanna D. Moore and Cécile L. Paris, "Planning text for advisory dialogues," in *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, 1989.

[9] Giuseppe Carenini and Johanna D. Moore, "Generating explanations in context," in *Proceedings of the 1st international conference on Intelligent user interfaces*, 1993.

[10] J. D. Moore, "Responding to 'HUH?': answering vaguely articulated follow-up questions," in *Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind*, 1989.

[11] M.C. Tanner and A.M. Keuneke, "Explanations in knowledge systems: the roles of the task structure and domain functional models," *IEEE Expert*, vol. 6, no. 3, 1991.

[12] J. L. Weiner, "BLAH, a system which explains its reasoning," *Artificial Intelligence*, vol. 15, no. 1-2, pp. 19-48, 1980.

[13] Agneta Eriksson, "Neat explanation of Proof Trees," in *Proceedings of the 9th international joint conference on Artificial intelligence*, vol. 1, 1985.

[14] C. Millet and M. Gilloux, "A study of the knowledge required for explanation in expert systems," in *Proceedings of Artificial Intelligence Applications*, 1989.

[15] J.W. Wallis and E.H. Shortliffe, "Customized explanations using causal knowledge," in *Rule-based Expert Systems*, Addison-Wesley, 1984.

[16] K. N. Papamichail and S. French, "Explaining and justifying the advice of a decision support system: a natural language generation approach," *Expert Systems with Applications*, vol. 24, no. 1, pp. 35-48, 2003.

[17] Carenini and Moore, "Generating and evaluating evaluative arguments," *Artificial Intelligence*, vol. 170, no. 11, pp. 925-952, 2006.

[18] T. R. Gruber and P. O. Gautier, "Machine-generated explanations of engineering models: A compositional modeling approach," *IJCAI*, 1993.

[19] Patrice O. Gautier and Thomas R. Gruber, "Generating Explanations of Device Behavior Using Compositional Modeling and Causal Ordering," *AAAI*, 1993.

[20] B. White and J. Frederiksen, "Causal model progressions as a foundation for Intelligent learning," *Artificial Intelligence*, vol. 42, no. 1, pp. 99-155, 1990.

[21] F. Elizalde et al., "An MDP approach for explanation. Generation," In *Workshop on Explanation-Aware Computing with AAAI*, 2007.

[22] O. Z. Khan et al., "Explaining recommendations generated by MDPs," In *Workshop on Explanation Aware Computing*, 2008.

[23] S. Renooij and L. Van-DerGaa, "Decision making in qualitative influence diagrams," In *Proceedings of the Eleventh International FLAIRS Conference*, pp. 410–414, 1998.

[24] C. Lacave et al. "Graphical explanations in bayesian networks," In *Lecture Notes in Computer Science*, vol. 1933, pp. 122–129. Springer-Veralg, 2000.

[25] Andrew Ko and Brad Myers, "Extracting and answering why and why not questions about Java program output," *ACM Transactions on Software Engineering and Methodology*, vol. 20, no. 2, 2010.

[26] F. Sørmo et al., "Explanation in case-based reasoning – perspectives and goals," *Artificial Intelligence Review*, vol 24, no. 2005, pp. 109–143, 2005.

[27] A. Kofod-Petersen and J. Cassens, "Explanations and context in ambient intelligent systems, in *Proceedings of the 6th international and interdisciplinary conference on Modeling and using context*, 2007.

[28] C. P. Langlotz et al., "A methodology for generating computer-based explanations of decision-theoretic advice," *Med Decis Making*, vol. 8, no. 4, pp. 290-303, 1988.

[29] H. Lieberman and A. Kumar, "Providing expert advice by analogy for on-line help," in *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pp. 26-32, 2005.

[30] Baderet et al., "Explanations in Proactive Recommender Systems in Automotive Scenarios," *Workshop on Decision Making and Recommendation Acceptance Issues in Recommender Systems Conference*, 2011.

[31] P. Pu and L. Chen, "Trust building with explanation interfaces," *in: 11th International conference on Intelligent User Interfaces*, pp. 93-100, 2006.

[32] Baltrunas et al., "Context-Aware Places of Interest Recommendations and Explanations," in *1st Workshop on Decision Making and Recommendation Acceptance Issues in Recommender Systems*, (DEMRA 2011), 2001.

[33] J. D. Lee K. A. See, Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, vol 46, no 1, pp. 50-80, 2004.

[34] D. L. McGuinness et al., "Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study," in *Workshop on the Models of Trust for the Web*, 2006.

[35] I. Zaihrayeu, P. Pinheiro da Silva, and D. L. McGuinness, "IWTrust: Improving User Trust in Answersfrom the Web," in *Proceedings of the 3rd International Conference on Trust Management*, pp. 384-392, 2005.

[36] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proceedings of the 27th international conference on Human factors in computing systems*, pp. 2119-2128, 2009.

[37] A. Glass, D. L. McGuinness, and M. Wolverton, "Toward establishing trust in adaptive agents," in *Proceedings of the 13th international conference on Intelligent user interfaces*, pp. 227-236, 2008.

[38] J. J. Dijkstra, "On the use of computerised decision aids: an investigation into the expert system as persuasive communicator," Ph.D. dissertation, 1998.

[39] S. Y. Rieh and D. R. Danielson, "Credibility: a multidisciplinary framework." In Annual Review of Information Science and Technology, B. Cronin (Ed.), Vol. 41, pp. 307-364, 2007.