A Probabilistic Model for Personality Trait Focused Explainability

Mohammed N. Alharbi
Department of Computer & Electrical Engineering and Computer Science
Florida Atlantic University
Boca Raton FL USA
malharbi2016@fau.edu

Shihong Huang
Department of Computer & Electrical Engineering and Computer Science
Florida Atlantic University
Boca Raton FL USA
shihong@fau.edu

David Garlan
Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh PA USA
garlan@cs.cmu.edu

***Abstract*— Explainability refers to the degree to which a software system's actions or solutions can be understood by humans. Giving humans the right amount of explanation at the right time is an important factor in maximizing the effective collaboration between an adaptive system and humans during interaction. However, explanations come with costs, such as the required time of explanation and humans' response time. Hence it is not always clear whether explanations will improve overall system utility and, if so, how the system should effectively provide explanation to humans, particularly given that different humans may benefit from different amounts and frequency of explanation. To provide a partial basis for making such decisions, this paper defines a formal framework that incorporates human personality traits as one of the important elements in guiding automated decision-making about the proper amount of explanation that should be given to the human to improve the overall system utility. Specifically, we use probabilistic model analysis to determine how to utilize explanations in an effective way. To illustrate our approach, Grid – a virtual human and system interaction game -- is developed to represent scenarios for human-systems collaboration and to demonstrate how a human's personality traits can be used as a factor to consider for systems in providing appropriate explanations.**

Keywords*— *explainability, human system co-adaptation, human computer interaction (HCI), personality traits, self-adaptive systems, human-in-the-loop, model checking, probabilistic model.

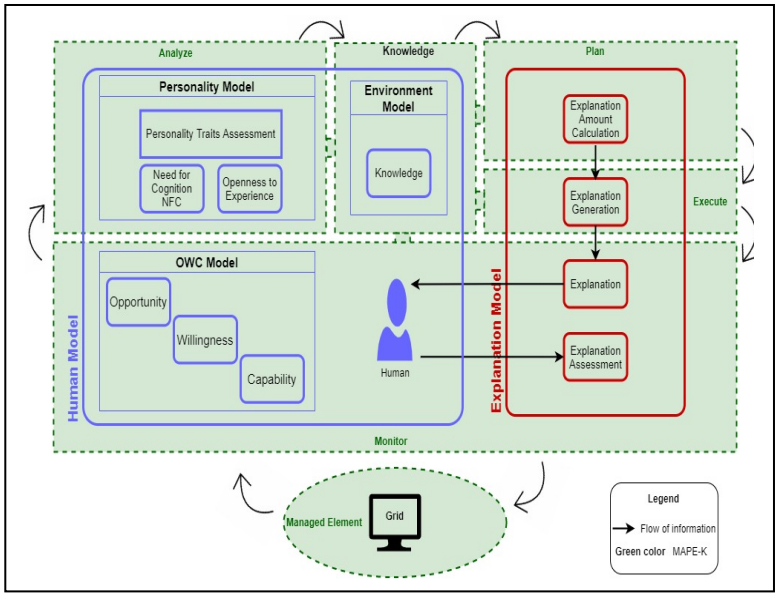


Fig. 1 A probabilistic model for personality trait focused explainability framework: this framework incorporates two basic personality traits (Openness and Need for Cognition) as important elements in a human model that can be used to guide a system in deciding the appropriate amount of explanation that should be given to the human

I. Introduction

As systems become more autonomous and intelligent through the incorporation of AI techniques and self-adaptive approaches, it becomes increasingly important for those systems to be able to "explain" themselves to their human users and collaborators [1][2]. In particular, there are four main purposes of explainability: (1) explain to justify: use explanations to justify some results to the human, particularly when decisions are made suddenly; (2) explain to control: explanations can help not only to justify, but also to prevent systems from going wrong; (3) explain to improve: improving the systems continuously through human involvement; (4) explain to discover: discovering and gathering new facts that help us to learn and to gain knowledge. In the context of this paper, explainability refers to the degree to which a software system's actions or solutions can be understood by humans, and explainability is used to *improve* a system's overall utility.

While explanation is an increasingly desirable – even, essential – capability of a system, it is not at all obvious when and how explanation should be given, particularly since explanation comes with a cost on human attention and delays in system-human interaction and the fact that different humans may need different kinds of explanation. To partially address this problem this paper defines a formal framework, as illustrated in Figure 1, for reasoning about the proper amount of explanation that a system should provide to the human *based on their personality traits*. Specifically, leveraging research in the psychology of human personality, this framework incorporates two basic personality traits (*Openness* and *Need for Cognition*) as important elements in a human model that can be used to guide a system in deciding the appropriate amount of explanation that should be given to the human in order to improve overall system utility. The effects of given explanations (which are determined based on personality traits of the human) affect human-system co-adaptation, represented through the Opportunity-Willingness-Capability (OWC) model, a commonly used model for adaptive systems' reasoning about human-in-the-loop behavior [3]. We incorporate our approach into the MAPE-K architecture [4] to formally model and analyze human involvement at different stages of system management and adaptation. To illustrate our approach, Grid – a virtual human and system interaction game – is developed to represent

scenarios for human-systems collaboration and to demonstrate how a human's personality traits can be used as a factor to consider for systems in providing appropriate explanations.

The organization of the paper is as follows: Section II describes the research problem and goals, Section III represents background information and related work, Section IV shows methodology, Section V shows the Stochastic Multi-player Games (SMG) model while Section VI shows results and analysis, Section VII represents discussion and future work and the last section focuses on the conclusion.

II. PROBLEM STATEMENT AND RESEARCH GOALS

A. Problem Statement

A co-adaptation system is symbiotic human-in-the-loop system where human-system cooperation is required in achieving shared goals, and system and human actions mutually impact each other's behavior in accomplishing coordinated tasks [5]. In this context, providing effective explanations to humans is an important factor in maximizing the co-adaptation outcomes between the system and the human [6]. Maximizing co-adaptation outcomes implies that the relationship between system and humans has become a partnership, or collaborative relationship, in which humans and systems act semi-autonomously – in contrast to traditional systems that wait for the human's inputs and commands to take action [6].

Given that different humans may benefit from different amounts and frequency of explanation, in this paper we argue that adapting the explanation to the particular human through knowledge of their personality traits can help the system in determining what are appropriate explanations and, therefore, maximize the benefits of co-adaptation. In particular, given that there are tradeoffs in determining what kind of explanations to give, it is important to be able to tailor the explanations to the user [7]. Providing longer and more-frequent explanations may increase the effectiveness of collaboration between the system and the human; however, this comes at the cost of taking more time for humans to understand the explanations and respond accordingly. Thus, key questions that must be answered by a system are: What should the contents of an explanation be, and how frequently should they be given? Further, how can we formalize and mechanize the decision process that a system uses in determining the answers to these questions?

B. Research Goals

In this paper we attempt to answer these questions by defining a formal framework for reasoning about how a self-adaptive system should provide explanations based on its knowledge of a person's personality traits. This framework uses probabilistic analysis to decide how explanations should be given, based on a formal human model that includes psychologically relevant aspects of personality. Specifically, we focus on answering the following research question: How to use knowledge about an individual's personality traits to improve the overall system utility?

The main contributions of this paper are:

- A formal framework that incorporates human personality traits and guides adaptive human-in-the-loop systems to decide how much explanations should be given in order to improve system utility.
- An evaluation system based on a collaborative game, to simulate the effects of decision making under various scenarios.

III. BACKGROUND AND RELATED WORK

This section introduces some background on personality traits, the OWC (Opportunity-Willingness-Capability) model, model checking of stochastic multi-player games (SMG), and some of state-of-the-art studies that focus on explainability and human-system co-adaptation. Section IV will then illustrate how this background and related work are related to what we do in this research.

A. Personality Traits

Psychological studies have demonstrated that human personality traits play a strong role in determining human behavior [8]. Personalities can be characterized in terms of traits that are relatively stable characteristics of a human that influence our behavior across many situations. An individual's personality is the combination of traits and patterns that influence his/her behavior, thought, motivation, and emotion. It drives individuals to consistently think, feel, and behave in specific ways.

There are, of course, many differences between individuals; however, personality traits are one of the more important measurable characteristics that can be used to distinguish one person from another. In the psychological literature the *Big Five* (also called the *Five Factor*) model of personality is one of the most widely accepted personality taxonomies. In the Big Five model, the five dimensions of personality include extraversion, neuroticism, openness to experience, agreeableness, and conscientiousness [9].

Openness to experience is one of the personality traits that is used to describe individual personality in the Five Factor Model. Open people tend to be intellectually curious, creative and imaginative. Open people have a high openness to embrace new things, fresh ideas, and novel experiences [10].

In addition to the Five Factor Model, the psychological literature also identifies Need for Cognition is an important distinguishing characteristics of human personality trait [9][11].

Need for Cognition (NFC) is defined as the "individual's tendency to engage in and enjoy effortful cognitive tasks." People with higher NFC levels typically prefer more detail, while those with low levels of NFC want to quickly understand the big picture and avoid engaging through more detail. Based on the NFC 10-item testing instrument [9][11], a score above 80 is generally considered to be High NFC (or high personality trait), and below 50 is Low NFC.

As we elaborate later, we adopt these two basic personality traits (*Openness to Experience* and *Need for Cognition*) as important elements in a human model that can be used to guide a system in deciding the proper amount of explanation that should be given to the human to improve overall system utility.

B. OWC (Opportunity-Willingness-Capability) Model

Prior research in adaptive systems has investigated various models of humans that can be used at run time to effectively characterize humans when deciding how best to incorporate them into a co-adaptive system. One of the more prominent

models is the OWC (Opportunity-Willingness-Capability) model [3].

OWC categorizes human attributes into: (1) *Opportunity*: indicates whether a human is available to participate in a cooperative task with the system (such as whether the human is physically present). (2) *Willingness*: identifies the human's inclination to perform the task (affected by cognitive load, human attention, stress level, and motivation). (3) *Capability*: defines the human's abilities and skills that are necessary to execute the task successfully (affected by level of experience or training, knowledge of the task, and cognitive or physical skills) [3].

This model has been used effectively in a number of papers to determine, for example, whether to involve the user in a task or to carry it out automatically [12][5], whether to proactively gain the user's attention [13], and when to provide an explanation [14]. As we detail later in this paper we use OWC to capture the co-adaptation attributes of the human (see Section IV. B).

C. Model Checking Stochastic Multiplayer Games (SMG) and PRISM

Probabilistic model checking is used as a technique to analyze the systems that exhibit stochastic behavior. Stochastic Multi-player Games (SMG) is a form of probabilistic modelling that allows us to reason quantitatively about reward-based properties and probability such as time, usage, and resources in a multi-agent system [15][16][17]. Our approach is to use SMG models to reason about the appropriate amount of explanation that should be given to the humans based on their personality traits where we model the system and humans as (cooperating) players in a game.

PRISM is "a probabilistic model checker, a tool for formal modelling and analysis of systems that exhibit random or probabilistic behavior" [18]. PRISM-games is an extension of PRISM that is used to analyze probabilistic systems where players can incorporate competitive or collaborative behavior, modelled as stochastic multiplayer games SMG [19]. Analyzing systems using PRISM has been carried out in variety of application domains, including: security protocols, communication and multimedia protocols, randomized distributed algorithms, biological systems and many others. PRISM can analyze a wide range of quantitative properties of stochastic models automatically (e.g., "what is the probability of a failure causing the system to shut down within 4 hours?"). PRISM further supports the specification and analysis of properties based on *costs* and *rewards*. These allow it to reason, not only about the probability that a model behaves in a certain way, but about a wide range of quantitative measures related to the behavior of the model (e.g., "expected number of lost messages", "expected time", or "expected power consumption").

In this paper we use PRISM to dynamically determine appropriate levels of explanation to maximize expected utility (expressed as a reward).

D. Human-in-the-Loop Self-Adaptation and Explainability

Human-system integration or human-system co-adaptation is advancing the fields of human-system interaction. Integration here means that the relationship between system and humans has become a partnership or symbiotic relationship in which humans (i.e., users) and systems act with autonomy instead of the system waiting for the user's inputs and commands to take an action. Self-adaptation refers to a process in which an interactive system co-adapts its behavior to a human based on its internal model of the human, dynamic information acquired about the human, the context of use and its surrounding environment [4][5][6].

Several related works have studied explainability focused on a human-system co-adaptation perspective. In [20] the authors propose a method that generates verbal explanations of multi-objective probabilistic planning. This method explains why a particular behavior is chosen on the basis of the optimization objectives. Their explainability method relies on describing the values of the objective of a generated behavior and, therefore, explaining tradeoffs that were made to reconcile competing objectives.

In [21], the authors define a formal framework to reason about explainability of co-adaptive system behaviors and the situations under which they are warranted. Specifically, they characterized explainability in terms of explainability cost, effect, and content. They propose a dynamic adaptation approach that uses a probabilistic reasoning technique, similar to ours, in order to determine when the explanations should be used for the purpose of improving system utility.

In another related work [14], the authors use a similar framework of [21] to reason about explainability of adaptive system behaviors and the conditions under which they are warranted. They characterize explainability in terms of the effects on a human operator's ability to engage in co-adaptive actions effectively. They present a decision-making mechanism to plan in self-adaptation that provides a probabilistic reasoning tool to determine when explanations should be used in an adaptation.

While this prior work shares with our research the goal of reasoning about explanation in the context of human-system co-adaptation, and also use probabilistic reasoning to account for inherent uncertainties in our human models, none of these studies take into consideration specific personality traits of humans – the main focus of our work.

IV. Methodology

In this section, we illustrate how we use explanation as a tactic (or action) that systems can use to improve the efficiently and effectiveness of human-system co-adaptation based on human personality traits. We describe also how we utilize a probabilistic planner [19] to determine the optimal amount of explanation according to those personality traits.

A. Selection of Personality Traits

An important question is which personality traits to consider with respect to explanation? As noted earlier, the psychological literature has classified a variety of important distinguishing characteristics for human personality. However, not all traits are relevant to explainability. In this work we have adopted two personality traits: *Need for Cognition* (NFC) and *Openness to experience*, since there is a direct relationship between NFC and explainability and between *Openness* and capability in OWC [9][10] (see Section IV. B).

We use the "Openness to experience" trait as one factor that affects the human's capability to continue and complete a task, since open people tend to be intellectually curious and have a high level of capability to do creative tasks [10]. We consider the Openness level as an important human factor

since an individual's Openness level reflects their capability to engage in cognitive tasks.

In our work we assume that the human's personality traits are known (for example, by using the NFC 10-item testing instrument in [9][11]) and do not change over the time horizon of a particular set of interactions with the system. While the traits are assumed to be known, there does, however, remain some uncertainty about the impact of the amount of explanation that should be provided to the human, which we incorporate into our reasoning framework. We will further assume for concreteness that both selected personality traits are relevant, and that their weights are equally important (although the relative importance can be adjusted in the model).

B. *Incorporating the OWC (Opportunity-Willingness-Capability) Model*

We use the OWC model (described in Section III.B) to capture the co-adaptation attributes of the human. In this paper, the following indicators show the connection between our model and the OWC model and how the OWC is incorporated in the context of the collaborative Grid game: (1) time and location represent the set of variables of the *Opportunity* category. Is the player located at the correct location? Has the timer expired? (2) Human satisfaction represents the *Willingness* category. Is the human satisfied with the given explanation? That category is applied through the playerFeedback (pF) tactic. (3) Human performance represents the *Capability* category. The Capability category identifies the ability of the human to complete Grid task. Giving an explanation increases the capability of the human to successfully carry out that particular task [12].

C. *Utlizing Model Checking Stochastic Multiplayer Games (SMG)*

The probabilistic model checker (PRISM-games) is utilized to formally model our approach. PRISM-games is particularly suitable for our study because it helps us to reason quantitatively under unpredictability and uncertainty about "how much" explanations should be given. The uncertainty (or stochasticity) that is relevant in this context is about the proper amount of explanations and the impact of different amounts of explanations that should be given to the human.

We model the system (the Grid game described in Section IV.D below) as a turn-based SMG, which means exactly one player in each state of the modeled system can choose an action, where the outcome of that state will be probabilistic. Players in a SMG may cooperate to achieve a common goal, or compete to accomplish different goals. In our examples, we model two players[1], the *human* and the *system*, and we assume that they share a common goal.

We use rPATL, a probabilistic temporal logic, to express properties of stochastic multi-player games quantitatively. rPATL helps us to reason about the collective ability of a group of players to achieve a goal relating to the probability of an occurring event [22].

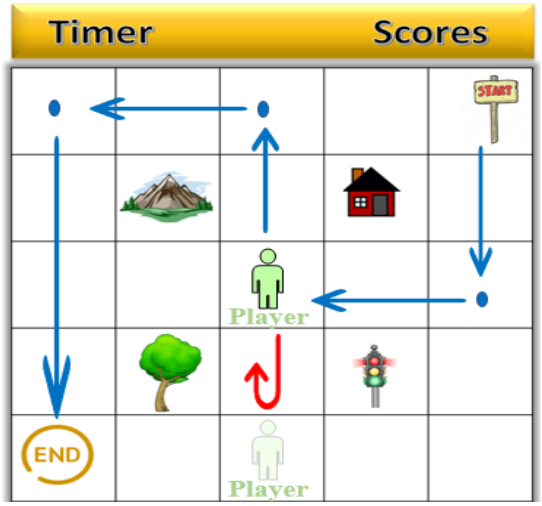


Fig. 2 The Grid Game we defined that embodies a representative scenario for human-system ao-adaptation

D. *Grid Game*

To illustrate our approach, we defined the Grid game– a virtual game -- as shown in Figure 2, as a game that embodies a representative scenario for human-system co-adaptation.

In the Grid game the system S instructs a player P verbally to move on a 5×5 grid from the top right corner (start) to the bottom left corner (end). The game is designed to rely on explanations, at various levels of detail, to instruct the user on what tasks to perform and how to perform them.

Game objectives:

- Follow the system instruction through a certain path within a certain maximum amount of time (60 seconds).
- Minimize the time *t* to complete the task.
- Traverse an optimal number of blocks to complete the end-to-end task, avoiding obstacles.

Game rules:

- The player can move either horizontally or vertically.
- Game score (100 points): points are deducted for traversing extra blocks or moving into or through obstacle squares (e.g., in Figure 2 there are four obstacles: the house, a traffic light, a mountain, and a tree).

The Grid game can involve the use of five tactics for interacting with the player, as shown in Table 1. The system provides two levels of explanation to command the human to move from one point to another. The choice of level of explanation is based on the run-time calculation and explanation generation based on the probabilistic model.

In this case "less explanation (LExp)" provides an abbreviated command (e.g., "Go 2 blocks left"), while "more

[1] Note that a multiplayer here (i.e., two players) does not mean that the Grid game is a multiuser game. The concept "multiplayers" in PRISM refers to multiple agents, such as system, human, or environment. In our model, the *system* and the *human* are the only two players and they are working cooperatively (taking turns) to achieve the best possible outcome.

TABLE I: GRID GAME TACTICS FOR INTERACTING WITH THE PLAYER

Model	Categories	Tactics	Role	Example
System	Less Explanation	lessExplain (lExp)	Commands the human to carry out an action.	"Go 2 blocks left" "Move south 4 blocks"
	More Explanation	moreExplain (mExp)	The system further explains information when the human is confused and loses track.	"You will go between a house and traffic light" "You go straight, and you see a car on your left side"
Human	Clarification Request	Check (Chk)	The human requests the system to confirm information that they not entirely sure about.	"North?" "Should I continue above the tree?"
	Feedback	playerFeedback (pF)	Human feedback is collected about his satisfaction for each given explanation	Helpful, Not helpful, Neutral
	Acknowledgement	confirm (conf)	The human confirms information and follows the instructions.	"Yeah", "Thanks" "Okay"

explanation (mExp)" provides an abbreviated command (e.g., "You will go between a house and traffic light") contains additional details. The human may request clarification about a given explanation if they are not entirely sure about it (Chk) (e.g., "Should I continue above the tree?"). Or the user can confirm the information and follow the instructions (conf). The human also gives feedback (pF) about the given explanation as to whether it was (a) helpful, (b) not helpful, or (c) neutral. (This supports explanation assessment in the framework - Figure 1).

1) Utility Attributes

The four utility attributes of the game are: RequiredTime (t), Blocks (B), LengthOfExplanations (xL), and ExplainEfficiency (xE). B and t are used for calculating the game score, and xL and xE are used as explainability attributes. Game score(s) depend on the time elapsed for completing the game (t), associated with the optimal number of the blocks (B) that the player is supposed to end the task with:

- RequiredTime (t): the total elapsed time for completing the game.
- Blocks (B): the number of the blocks traversed to complete the task.
- LengthOfExplanations (xL): the amount of delay (or time) required to explain.
- ExplainEfficiency (xE): a measurement that determines how happy the player is with the given explanations. xE is associated with the playerFeedback (pF) tactic which can be one of the following values: Helpful, Not helpful, or Neutral.

2) Tactics Cost/Benefit and Utility Dimensions

Table 2 lists the tactics in the Grid game, and their impacts on utility dimensions. Different tactics cause an increase in Time (three seconds for lExp, Chk, and conf; six seconds, for mExp). The upward ↑ or downward arrow ↓ reflects utility increments and decrements, respectively. For example, the lExp tactic increases both t and xL by three seconds, which is associated with a smaller amount of costs. Human feedback is collected about the user's satisfaction for the given explanation (lExp) which can be:

a) Helpful (H) reflects utility increments (↑),

b) Not helpful (NH) reflects utility decrements (↓),

c) Neutral (N) reflects neither utility increments nor decrements (-).

3) Utility Functions

To compare different explainability tactics (i.e., lengths of explanation), we use probabilistic temporal logic with rewards, rPATL, which enables us to analyze the utilities of the system that explainability can influence. rPATL (described in Section IV.C) is used to reason about the ability of a group of players (system and human) to collectively achieve a specific goal [18].

In the formal model we define formulas that represent the accrued utility (The Scores function $\cup s$ and the ExplainEfficiency Function $\cup xE$) as the maximum real immediate utility that the human can achieve along the whole task.

TABLE II: COST/BENEFIT ' IMPACTS ON UTILITY DIMENSIONS

Tactics	Time	Δ ExplainEfficiency (xE) [a]		
	Δ *RequiredTime (t)* Δ *LengthOfExplanations (xL)*			
	↑	↑	↓	-
lessExplain (lExp)	+3	H	NH	N
moreExplain (mExp)	+6	+1	-1	0
Check (Chk)	+3			
confirm (conf)				
playerFeedback (pF)				

[a] H: Helpful, NH: Not helpful, N: Neutral

The Scores function $\cup s$, as shown in function (1), maps high scores to high utility derived by dividing the number of blocks B by the maximum level of RequiredTime t^{max} (t^{max}= 60), where B must be greater than or equal to $optimalB$ (the optimal number of blocks that the player is supposed to complete the task with):

$$\cup s(B) = \left(1 - \frac{B}{t^{max}}\right) x100 \text{ where } B \geqslant optimalB \qquad (1)$$

The ExplainEfficiency Function $\cup xE$, as shown in function (2), maps higher levels of ExplainEfficiency (xE) derived by dividing the accumulated player Feedback ($\sum pF$)

by the total number of feedbacks (pF^{max}), where $pF \epsilon [1,0,-1]$ represent Helpful, Neutral, Not helpful, respectively:

$$\cup xE(pF) \approx \left(\frac{\sum \boldsymbol{pF}}{pF^{\boldsymbol{max}}}\right) x100 \qquad (2)$$

Both personality trait variables (Openness and NFC) are initialized with some constants (as inputs) that represent the human traits. Personality traits are directly mapped to the level of explanation (the amount), and are used to calculate the probability of getting explanations in that amount. Function (3) shows combined personality traits, which will be 0 in case both traits are 0, or 1 in case the human has the highest personality traits levels (i.e., $0 \rightarrow 1$). Openness^{max} and NFC^{max} are 100, which represent the highest personality trait levels. The values of personality traits are determined based on the NFC 10-item testing instrument in [9][11] that produces scores between 0-100.

For example, if a human has 75 Openness and 90 NFC. The combined human traits are 0.82 which means he has high personality traits (by using function (3)). That means the system will explain less 18% of the time (i.e., lExp) and explain more 82% of the time (i.e., mExp) during the task. As another example, suppose a human with low personality traits has 43 openness and 49 NFC. The combined human traits are 0.46 (using function (3)). That means the system will explain less 54% of the time (i.e., lExp) and explain more 46% of the time (i.e., mExp) while playing the Grid game.

$$Human\ Traits = \frac{Openness+NFC}{Openness^{max}+NFC^{max}} \qquad (3)$$

TABLE III: AN EXAMPLE DIALOGUE OF A SCENARIO BETWEEN THE SYSTEM (S) AND A HUMAN (H)

Scenario	*Tactics*	*Time*	*pF*
S: Can you go 2 blocks down?	(lExp)	3s	H
H: Yeah	(conf)	3s	-
S: Then go 2 blocks left.	(lExp)	3s	NH
H: Could you repeat that?	(Chk)	3s	-
S: Go west. You will go between a house and traffic light.	(mExp)	6s	H
H: Okay	(conf)	3s	-
S: Go after that 2 blocks up.	(lExp)	3s	N
H: The human is on the wrong track			-
S: No, not south. You go north	(mExp)	6s	H
H: Okay	(conf)	3s	-
S: Go 2 blocks left	(lExp)	3s	N
……			
S: Go south 4 blocks.	(lExp)	3s	H
H: Okay, thanks a lot.	(conf)	3s	-

4) Example Scenario

Figure 2 and Table III show an example dialogue of a scenario between the system (S) and a human (H).

The human spent 42 seconds (t) and used 15 blocks (B) to finish the task. However, the number of blocks B that the player is supposed to end the task with is 12 ($optimalB$). The system took 27 seconds for explanations (xL). At the end of the task, the score of the player is 75 (by using the function (1), where B=15 and tmax= 60), and the ExplainEfficiency (xE) is 43 (by using the function (2), where $\sum pF$ is three and pF^{max} is seven (which means seven feedbacks are collected)).

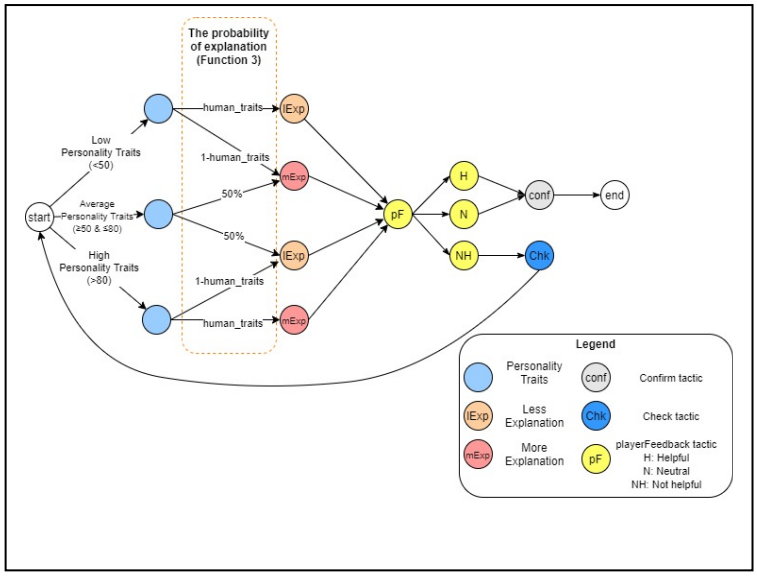


Fig. 3 The strategy we use to model the SMG: the proper amount of explanation is determined based on the three personality levels of the human (represented in light blue). The two explanation amounts (less or more) are determined by using function 3. Human feedback is collected that will can be helpful, neutral, or not helpful (represented in yellow). The human confirms information that means he moved successfully to the next point (conf), or checks/requests the system to clarify information that they not entirely sure about (chk).

V. THE STOCHASTIC MULTI-PLAYER GAMES (SMG) MODEL

We model the Stochastic Multi-player Games (SMG) model as two players, where the players try to collaboratively maximize accumulated reward(s): (1) Player SYS specifies the actions that are controlled by the system (i.e., it represents the Grid game). (2) Player HUMAN specifies the actions belonging to the human (i.e., it represents the game player). The models represent the behavior of a set of agents (or "players") that take turns making moves, where the choice of move is specified probabilistically or non-deterministically. A game solver for such a system (such as PRISM-games [19]) determines an optimal strategy for the players by resolving the non-deterministic transitions in such a way that the expected reward for each player is maximized (assuming rational play by each). Figure 3 shows the strategy we use to model the SMG. The proper amount of explanation is determined based on the three personality levels of the human (i.e., *Low*, *Average*, or *High* personality levels (represented in light blue)). The two explanation amounts (less or more) are determined by using Function 3 (describes in the previous section). Human feedback is collected that is of the form *Helpful*, *Neutral*, or *Not Helpful* (represented in yellow). The human confirms information that means he moved successfully to the next point (conf), or checks/requests the system to clarify information that they not entirely sure about (Chk).

The Stochastic Multi-player Games (SMG) model consists of the following four parts:

A. Player Definition

Player definition includes the declaration of the two players in the SMG and different modules that each player has control of. The two players in our game are shown in Listing 1. Player SYS (lines 1-2) specifies the actions that are controlled by the system (i.e., it represents the Grid game). Player HUMAN (lines 3-4) specifies the actions belonging to the human (i.e., it represents the game player). Our Grid game is played in turns by the two players SYS and HUMAN. Turn (line 5) is a global variable used as a controller to take turns between different players, ensuring that only one player can take an action at each state of the model execution. Tactics are executed sequentially in our model.

```
1.  player SYS
        Game, [lExpLow],[ lExpAvg],[ lExpHigh],
        [mExpLow],[mExpAvg],[mExpHigh]
2.  endplayer
3.  player HUMAN
        Play, [conf], [Chk]
4.  endplayer
5.  global turn:[SYS..HUMAN] init SYS;
6.  const SYS=1; const HUMAN=2;
```

Listing. 1 Player definition includes the declaration of the two players in the SMG and different modules that each player has control of

B. Game Model

Player SYS has control of the Game model, illustrated in Listing 2. Opportunity elements are used as execution conditions of different tactics such as: the human is at the correct location ((x= 1)&(y=1)) and is not involved in a crash, and the time has not expired (t<60). The Game module is parameterized by the variables (lines 1-2), which indicate the state of tactic execution, where false means this tactic is not in use (i.e., lExp_state, and mExp_state).

During the system's turn, the system executes these tactics sequentially: lExp (lines 5-13), and mExp (lines 15-23). For the sake of clarity, we will describe only the lExpLow tactic to illustrate how tactic execution is modeled. The other explainability tactics follow the same structure. The system instructs the human with low personality traits through executing the command labeled as lExpLow (line 5). This tactic executes only if:

- It is the turn of the SYS.
- The human traits are low (<0.50).
- The player position is on a certain block (x1,y1).
- The end time of the task has not been reached yet (t<60).

If the guard is satisfied, the system will explain more by flagging mExp_state tactic true with probability human_traits (line 6). Otherwise, the system will explain less by flagging lExp_state tactic true with probability 1-human_traits (line 7) and the system will:

- Commands the player to move to the position (x2,y2).
- Increases the time 3 seconds (xL'=xL+3)&(t'=t+3).
- Flags the lExp tactic as true (lExp_state'= true).
- Updates the value of the variable turn, changing control to the human player (turn'=HUMAN).

Similarly, the system instructs the human with average personality traits through executing the command labeled as lExpAvg (line 8-10), or the system instructs the human with high personality traits through executing the command labeled as lExpHigh (line 11-13).

The human wins by executing the command labeled as win (line 25). That means the human (turn=SYS) has arrived at the bottom left corner ((x1= 1)&(y1=1)) within the time limit (t<60). However, the human loses the game by executing the command labeled as lose (line 26) when the end time of the task has been reached (t=60).

```
1.  global lExp_state: bool init false;
2.  global mExp_state: bool init false;
3.  …
4.  module Game
5.  [lExpLow] (turn=SYS)&(human_traits<.5)&(x1= 5)
        &(y1=5)&(t<60)
6.      ->human_traits:(mExp_state'= true)
7.      + 1-human_traits:(x2'=5) & (y2'=3)&(xL'=xL+3)
        & (t'=t+3)&( lExp_state'= true)&(turn'=HUMAN);
8.  [lExpAvg] (turn=SYS)&(x1= 5)&(y1=5)&(t<60)
9.      ->0.5:(x2'=5)&(y2'=3)&(xL'=xL+3)&(t'=t+3)
        &(lExp_state'= true)&(turn'=HUMAN)
10.     +0.5:(mExp_state'= true);
11. [lExpHigh] (turn=SYS)&(human_traits>.8)
        &(x1= 5)&(y1=5)&(t<60)
12.     ->human_traits:(mExp_state'= true)
13.     + 1-human_traits:(x2'=5) & (y2'=3)&(xL'=xL+3)
        & (t'=t+3)&( lExp_state'= true)&(turn'=HUMAN);
14. …
15. [mExpHigh] (turn=SYS)&(human_traits>.8)&
        (conf_state= false)&(Chk_state= true)&(t<60)
16.     ->human_traits:(mExp_state'=true)&(xL'=xL+6)
        &(t'=t+6)&(Chk_state'= false)&(turn'=HUMAN)
17.     + 1-human_traits: (lExp_state'= true);
18. [mExpAvg] (turn=SYS)&(conf_state= false)
        &(Chk_state= true)&(t<60)
19.     ->0.5:(mExp_state'= true)&(xL'=xL+6)&(t'=t+6)
        &(Chk_state'= false)&(turn'=HUMAN)
20.     +0.5:( lExp_state'= true);
21. [mExpLow] (turn=SYS)&(human_traits<.5)
        &(conf_state= false)&(Chk_state= true)&(t<60)
22.     ->human_traits:( lExp_state'= true)
23.     + 1-human_traits:(mExp_state'=true)&(xL'=xL+6)
        &(t'=t+6)&(Chk_state'= false)&(turn'=HUMAN);
24. …
25. [win] (turn=SYS)&(x1= 1)&(y1=1)&(t<60)
        -> (win'=true)&(turn'=0);
26. [lose] ((turn=SYS)|(turn=HUMAN))&(t=60)
        -> (win'=false)&(loser'= true)&(turn'=0);
27. endmodule
28. …
```

Listing. 2 Game Model

C. Play Model

Player HUMAN has control of the Play model, illustrated in Listing 3. The encodings of the HUMAN module are similar to those of the SYS module. The Play module is parameterized by variables (lines 1-2), which indicate the state of tactic execution, where false means this tactic is not in use (e.g., Chk_state, and conf_state). Personality Traits are initialized with values that represent the human's personality (lines 3-5).

During the human's turn, the human can execute one of these tactics: conf (line 8), and Chk (line 10). We explain only the conf tactic to illustrate how tactic execution is modeled.

The human confirms (conf) and follows the system instructions (i.e., the human moves successfully from the 1st point to the second) by executing the command labeled as conf. This tactic executes only if:

- It is the turn of the HUMAN.
- The system instructs the player to move to the position (x2,y2).
- The end time of the task has not been reached yet (t<60).

```
1.  global Chk_state: bool init false;
2.  global conf_state: bool init false;
3.  const int INIT_OPN; const int INIT_NFC;
4.  global human_Open: [1..100] init INIT_OPN;
5.  global human_NFC:[1..100] init INIT_NFC;
6.  …
7.  module Play
8.  [conf] (turn=HUMAN)&(x2= 5)&(y2=3)&(t<60)
        ->(x1'=5) & (y1'=3)& (t'=t+3)&(B'=B+2)
        &(pF'=pF+1)&(pfMAX'=pfMAX+1)&(conf_state'= true)
        &(lExp_state'= false)&(turn'=SYS);
9.  …
10. [Chk] (turn=HUMAN)&(conf_state= false)&(x1= 5)
        & (y1=3)&(t<60)
        ->(Chk_state'= true)&(t'=t+3)&(pF'=pF1)
        &(pfMAX'=pfMAX+1)&(turn'=SYS);
11. [wrong] (turn=HUMAN)&(x2= 3)&(y2=5)&(loser=false)
        &(t<60)-> (x1'=3) & (y1'=1) &(t'=t+3)
        & (B'=B+2)&(pF'=pF-1)&(pfMAX'=pfMAX+1)
        &(conf_state'= false) &(turn'=SYS);
12. [crash] (turn=HUMAN)& ((x1=obj1x & y1=obj1y)
        | (x1=obj2x & y1=obj2y)| (x1=obj3x & y1=obj3y)
        | (x1=obj4x & y1=obj4y))->(turn'=SYS);
13. endmodule
14. …
```

Listing. 3 Play Model

If the guard is satisfied, the player:

- Moves to the position (x1,y1).
- Increases the time three seconds (t'=t+3).
- Increases the number of Blocks by two (B'=B+2).
- Gets the player feedback (pF= 1 means the explanation was helpful).
- Increases the player feedback counter pfMAX by 1.
- Flagging the conf tactic true (conf_state'= true).

Moreover, the tactic wrong (line 11) will be executed when the human moves in the wrong direction, and the tactic crash (line 12) will be executed when the human moves to one of the obstacle squares (the house, a traffic light, a mountain, or a tree in Figure 2).

D. Utility Profile and Reward Structure

Utility functions are described in Section IV.D and illustrated in Listing 4. Formulas and reward structures are used to encode the utility functions that allow us to quantify the utilities of different task states.

The Scores function, $\cup s$, as in lines (1-2), represents the encoded Function (1) as described in Section IV.D. ExplainEfficiency function $\cup xE$, as in lines (3-4) , represents the encoded Function (2) described in Section IV.D. Line 5 shows the encoded combined traits function (3).

```
1.  rewards "Scores"
        [win] true:(1-(B/tMax))*100;
        [lose] true:0;
        [crash] true: -5;
2.  endrewards
3.  rewards "ExplainEfficiency"
        [win] true:(pF/pfMAX)*100;
        [lose] true:(pF/pfMAX)*100;
4.  endrewards
5.  formula human_traits =
        (human_Open+ human_NFC)/(Max_Open+Max_NFC);
```

Listing 4. Utility profile and reward structure: formulas and reward structures are used to encode the utility functions that allow us to quantify the utilities of different task states. Formulas calculate system utility of the different states.

VI. RESULTS AND ANALYSIS

In this section, we illustrate how our modeling framework can produce optimal decisions with respect to how adaptive systems should explain to the human based on their personality traits. Specifically, we use SMG models of explainability to determine the expected outcome utilities of using different explainability tactics (i.e., explanation amounts) based on the personality traits of the human. Our modeling is done as a simulation (or set of "experiments" in PRISM terms). We use rPATL to ask PRISM a variety of questions such as "what is the maximum/minimum probability a human with high/low personality traits can guarantee to win with high/low utilities?" [22].

Table IV and Figure 4 show the analysis results of 44 rounds run on PRISM. All possible combinations of personality traits are taken into consideration, where high traits are (>80) (represented by orange color), average traits are ($\geqslant$50 and $\leqslant$80) (represented by blue-gray color), and low traits are (<50) (represented by gray color). Plot (a) shows the 44 simulations of different personality traits and the given amounts of explanations (LengthOfExplanations (xL)) to complete the task. The average of different personality traits and the amounts of explanations (xL) is shown in Plot (b). 39% of the iterations (17 rounds) of humans with high personality traits (>80) needed more explanations to finish the task with an average of 21 seconds. 32% of the iterations (14 rounds) of humans with low personality traits (<50) needed less amount of explanations with an average of 20 seconds. The remaining 30% of the iterations (13 rounds) belongs to a human with average personality traits ($\geqslant$50 and $\leqslant$80), where they use average amounts of explanation with an average of 19 seconds to complete the task. Table 5 shows the average of different utilities based on the three personality trait levels.

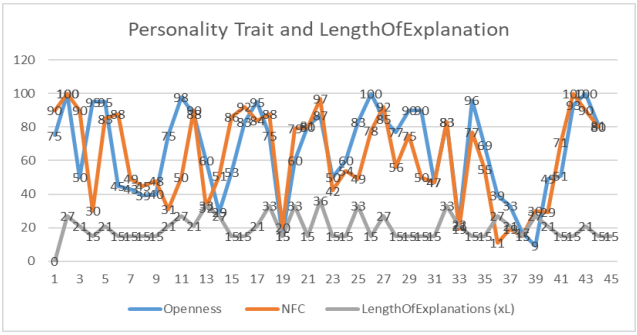
We can conclude from the results that a human with high personality traits needs more detailed information (i.e., explanations), while a human with low personality traits needs less detailed explanation. These conclusions are all consistent with psychology studies (discussed in Section III. A) that human with higher personality trait levels typically prefer more explanations, while those with low levels of personality trait want to quickly understand the big picture and avoid engaging through more explanations [9][11].

TABLE IV: RESULTS OF 44 ROUNDS RUN ON PRISM

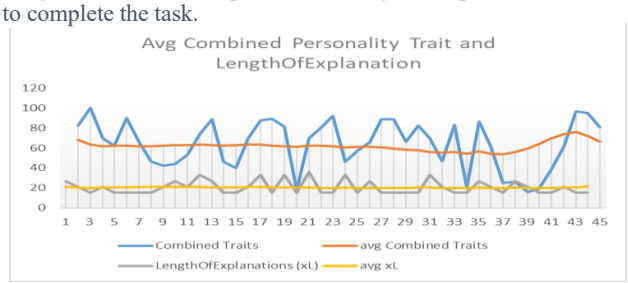
#	Human Traits		Combined Traits	Utilities		
	Openness	NFC		LengthOfExplanations (xL)	ExplainEfficiency (xE)	Scores
1	75	90	82.5	27	28.5	93.4
2	100	100	100	21	50	96.7
3	50	90	70	15	80	100
4	95	30	62.5	21	50	96.7
5	95	85	90	15	80	100
6	45	88	66.5	15	80	100
31	47	47	47	33	12.5	90
32	83	83	83	21	50	96.7
33	22	19	20.5	15	80	100
34	96	77	86.5	15	80	100
35	69	55	62	27	28.5	93.4
36	39	11	25	21	50	96.7
37	33	19	26	15	80	100
38	17	15	16	27	28.5	93.4
39	9	30	19.5	21	50	96.7
40	49	29	39	15	80	100
41	51	71	61	15	80	100
42	93	100	96.5	21	50	96.7
43	100	90	95	15	80	100
44	81	80	80.5	15	80	100
Minimum	9	11	16	15	-12.5	90
Maximum	100	100	100	36	80	100
Average	66.02	61.98	64	20.39	57.07	97.23

TABLE V: AVERAGE UTILITIES OF THE EXPERIMENTS BASED ON THE THREE PERSONALITY TRAITS LEVEL

Personality Traits Level	**Average Utilities**		
	xL	**xE**	**Scores**
High Traits	21.18	53.50	91.388
Average Traits	19.15	62.27	97.708
Low Traits	20.57	56.57	96.921



Plot (a): The 44 simulations of different personality traits and the given amounts of explanations (LengthOfExplanations (xL)) to complete the task.



Plot (b): The average of different personality traits and the amounts of explanations (xL)

Fig. 4 Results of 44 rounds run on PRISM show that human with higher personality trait levels typically prefer more explanations, while those with low levels of personality trait prefer less explanations

VII. DISCUSSION AND FUTURE WORK

In this research we presented an approach based on probabilistic model checking of SMGs to determine how much explanation should be given to the human based on their personality traits. Providing the right amount of explanation to the right human is an important factor to maximize co-adaptation between the system and the human during their interaction.

There is a number of limitations of this research that future research can address based on the foundations that we have described in this paper. These explanation decisions are ideal scenarios without having actual proof of that in reality. To address this, the most important next step is to conduct an empirical study to validate these models on actual real-world systems with humans in the loop.

As we explained earlier, there are many reasons to use explainability and improving a system's overall utility is one of the main reasons (see Section I). Explainability can help not only to *improve* the systems continuously through human involvement, but also to *justify* some information given to the human, particularly when decisions are made suddenly. Gaining more information improves the capability of the human to perform a task. Our results in this paper suggest one of the next steps of research is to go beyond the length of explanations, and examine in more detail questions such as how explanations should be presented: graphically, textually, verbally? A further extension of this research is to have more detailed models that allow the system to determine in a more nuanced way the ideal contents of the explanations that should be considered.

VIII. Conclusion

In this research we presented a formal framework that incorporates human personality traits as one of the important elements in guiding automated decision-making about the proper amount of explanation that should be given to the human to improve overall system utility. To accomplish our goal of this paper, we use probabilistic model analysis to determine how to utilize explanations in an effective way based on the difference of human's personality traits. Grid – a virtual human and system interaction game – was developed to illustrate our approach, to represent scenarios for human-system co-adaptation, and to demonstrate through simulation how a human's personality traits can be used as a factor to consider for systems in providing appropriate explanations.

Acknowledgment

This research was supported in part by the NSA under Award No. H9823018D0008 and Award No. N00014172899 from the Office of Naval Research. Any views, opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSA or the Office of Naval Research.

References

[1] G. Vilone and L. Longo, "Explainable Artificial Intelligence: a Systematic Review." arXiv preprint arXiv:2006.00093, 2020.

[2] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6. pp. 52138–52160, 2018.

[3] D. Eskins and W. H. Sanders, "The multiple-asymmetric-utility system model: A framework for modeling cyber-human systems," Proc. 2011 8th Int. Conf. Quant. Eval. Syst. QEST 2011, pp. 233–242, 2011.

[4] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," Computer (Long. Beach. Calif)., 2003.

[5] E. Lloyd, S. Huang, and E. Tognoli, "Improving Human-in-the-Loop Adaptive Systems Using Brain-Computer Interaction," Proceedings - 2017 IEEE/ACM 12th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, SEAMS 2017. pp. 163–174, 2017.

[6] M. Alharbi and S. Huang, "A Survey of Incorporating Affective Computing for Human-System Co-adaptation," in Proceedings of the 2020 The 2nd World Symposium on Software Engineering, 2020, pp. 72–79.

[7] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," FAT* 2019 - Proc. 2019 Conf. Fairness, Accountability, Transpar., pp. 279–288, 2019.

[8] C. G. H. Jung, "Psychological Factors Determining Human Behaviour," Collect. Work. C.G. Jung, Vol. 8 Struct. Dyn. Psyche, pp. 114–126, 2015.

[9] C. J. Sadowski and H. E. Cogburn, "Need for cognition in the big-five factor structure," Journal of Psychology: Interdisciplinary and Applied, vol. 131, no. 3. pp. 307–312, 1997.

[10] R. R. McCrae, "Openness to Experience as a Basic Dimension of Personality," Imagin. Cogn. Pers., 1993.

[11] R. E. Petty, J. T. Cacioppo, R. E. Petty, J. A. Feinstein, and W. B. G. Jarvis, "Dispositional Differences in Cognitive Motivation : The Life and Times of Individuals Varying in Need for Cognition Dispositional Differences in Cognitive Motivation : The Life and Times of Individuals Varying in Need for Cognition," Psychol. Bull., vol. 119, no. August, pp. 197–253, 2015.

[12] J. Cámara, G. Moreno, and D. Garlan, "Reasoning about Human Participation in Self-Adaptive Systems," Proc. - 10th Int. Symp. Softw. Eng. Adapt. Self-Managing Syst. SEAMS 2015, no. i, pp. 146–156, 2015.

[13] N. Li, C. Javier, D. Garlan, and B. Schmerl, "Hey ! Preparing Humans to do Tasks in Self-adaptive Systems ." In Proceedings of the 16th Symposium on Software Engineering for Adaptive and Self-Managing Systems, Virtual, 18-21 May 2021.

[14] N. Li, J. Cámara, D. Garlan, and B. Schmerl, "Reasoning about When to Provide Explanation for Human-in-the-loop Self-Adaptive Systems," In Proceedings of the 2020 IEEE Conference on Autonomic Computing and Self-organizing Systems (ACSOS), Washington, D.C., 19-23 August 2020.

[15] P. D. Kwiatkowska M., Norman G., "Probabilistic Model Checking: Advances and Applications," Form. Syst. Verif. Springer, Cham., pp. 73–121, 2018.

[16] C. Baier, "Probabilistic model checking," Dependable Softw. Syst. Eng., vol. 45, no. August, pp. 1–23, 2016.

[17] S. W. Cheng and D. Garlan, "Stitch: A language for architecture-based self-adaptation," J. Syst. Softw., 2012.

[18] M. Kwiatkowska, G. Norman, and D. Parker, "PRISM 4.0: Verification of probabilistic real-time systems," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011.

[19] M. Kwiatkowska, G. Norman, D. Parker, and G. Santos, "PRISM-games 3.0: Stochastic Game Verification with Concurrency, Equilibria and Time," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020.

[20] R. Sukkerd, R. Simmons, and D. Garlan, "Towards explainable multi-objective probabilistic planning," Proc. - Int. Conf. Softw. Eng., pp. 19–25, 2018.

[21] N. Li, S. Adepu, E. Kang, and D. Garlan, "Explanations for human-on-the-loop: A probabilistic model checking approach," Proc. - 2020 IEEE/ACM 15th Int. Symp. Softw. Eng. Adapt. Self-Managing Syst. SEAMS 2020, pp. 181–187, 2020.

[22] T. Chen, V. Forejt, M. Kwiatkowska, D. Parker, and A. Simaitis, "Automatic verification of competitive stochastic systems," Form. Methods Syst. Des., vol. 43, no. 1, pp. 61–92, 2013.