Received 8 July 2023, accepted 26 July 2023, date of publication 7 August 2023, date of current version 29 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3303193

RESEARCH ARTICLE

Implementation and Evaluation of Algorithms for Realizing Explainable Autonomous Robots

TATSUYA SAKAI[1], TAKAYUKI NAGAI[1,2], (Member, IEEE), AND KASUMI ABE[2]
[1]Graduate School of Engineering Science, Osaka University, Osaka 560-8531, Japan
[2]Artificial Intelligence Exploration Research Center, The University of Electro-Communications, Tokyo 182-8585, Japan

Corresponding author: Takayuki Nagai (nagai@sys.es.osaka-u.ac.jp)

This work was supported in part by the New Energy and Industrial Technology Development Organization (NEDO), and in part by the Japan Science and Technology Corporation (JST) Moonshot Research and Development under Grant JPMJMS2011.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee, Graduate School of Engineering Science, Osaka University.

ABSTRACT For autonomous robots to gain the trust of humans and maximize their abilities in society, they must be able to explain the reasons for their behavioral decisions. Defining explainable autonomous robots (XAR) as robots with such explanatory capabilities, we can summarize four requirements for their realization: 1) obtaining an interpretable decision space, 2) estimating the user's world model, 3) extracting important information for conveying policy in the robot, and 4) generating explanations based on explanatory factors. So far, these four elements have been studied independently. In this study, we first implement an explanatory algorithm that integrates these four elements. Then, we evaluate the implemented explanatory algorithm by conducting a large-scale subject experiment. The implemented explanation algorithm is shown to generate human-acceptable explanations; the results provide several insights and suggestions for future research on XAR. For example, we found that a robot that can give acceptable explanations to people is more likely to gain their trust. We also found that the questions "Why A?" and "Why not A?" should be explained in different ways.

INDEX TERMS XAI, explainable autonomous robots, world models, human-acceptable explanation.

I. INTRODUCTION

Autonomous robots are being integrated into society at an accelerating pace. Catering robots are being used in restaurants, and the introduction of service robots is being considered in various places. Autonomous robots will continue to expand, and it may not be long before robots are treated as partners on a par with humans. However, their inability to explain the reasons for their decision to act limit the scope of applications of these robots. Current autonomous robots play a specific role in society by faithfully executing commands given by humans to ensure their reliability. In other words, they are required primarily to behave as per the humans' expectations; if they behave unexpectedly, they will be stopped in an emergency without warning. However, robots with advanced sensors and decision-making algorithms can sometimes plan better behavior than humans' expectations. Moreover, if the robot is capable of autonomous online learning, its behavior can change from moment to moment. However, without appropriate explanatory capabilities, robots will find gaining the trust of humans and maximizing their abilities in society difficult.

In recent years, the importance of explainability in artificial intelligence has been widely recognized, and research on explainable artificial intelligence (XAI) [1], [2] is expanding. To endow a real robot with explainability sufficient to explain its reasons for its own action decisions to a user, there are research challenges common to XAI and unique to autonomous robots. The common issue is presenting the explanation in a form that humans can interpret. In particular, XAI focuses on converting the basis for decisions of machine learning models and the models themselves into interpretable

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Liu.

forms. Presenting explanations in human-interpretable forms is also an important research issue for autonomous robots.

Conversely, there are two main research issues specific to autonomous robots. The first is the mismatch in recognizing the features that serve as decision criteria. In XAI and explainable reinforcement learning (XRL), the input features given to the user and the model always coincide. However, autonomous robots and users do not always agree on the environmental model derived from the observations and information they are aware of.[1] Estimating the discrepancy between those perceptions and generating an explanation that considers the discrepancy's content is necessary.

The second point is the explanation of the process leading to the output. Most studies on XAI and XRL present the contribution of each feature to the output as an explanation and do not consider the explanation of why the feature contributes to it. However, for autonomous robots, the process by which each feature contributes to the output is not always obvious; hence, explaining this process is important.

We defined explainable autonomous robots (XAR) as the explainability of autonomous robots in [3] and proposed the four requirements shown in Fig.1 for its realization.

Requirement 1: interpretable decision-making space
Autonomous agents need to have a decision space in which each decision is interpretable by humans.

Requirement 2: the model of others
For autonomous agents to share plans with humans, the first step is to estimate the user's decision space and planning algorithm (model of others).

Requirement 3: information needed for explanation
Extracting information (explanatory factors) is useful for keeping the error between the agent's assumed plan and the human's assumed plan within a certain tolerance or reducing the user's estimation burden.

Requirement 4: presentation to user
The extracted explanatory factors are converted into human-interpretable forms such as language and visual representations.

Each of these requirements has been established as a research issue and has been studied independently [3]. This fact indicates that there are no previous examples of the integration of these elements with respect to the explainability of autonomous robots. In light of the XAR, implementing and evaluating an algorithm that satisfies all these requirements is essential. Investigating how such integration is accepted by people is important for further research on individual elements in the future. Therefore, this paper aims to implement an entire explanatory algorithm for autonomous robots that satisfies these requirements and to evaluate the algorithm through subject experiments to obtain suggestions on the explainability of autonomous robots. We believe that this will provide important insights into the explainability of autonomous robots and contribute to future research.

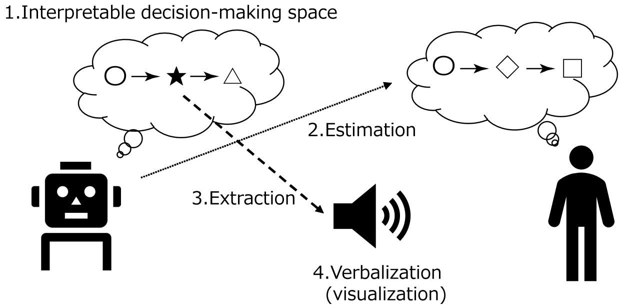


FIGURE 1. Four requirements for XAR [3].

II. RELATED WORKS

We explain existing research on the explainability of autonomous agents by classifying them into the four requirements mentioned earlier. The contributions of this paper are then described.

A. REQUIREMENT 1: INTERPRETABLE DECISION-MAKING SPACE

Among the studies on world models, the work of Zhang et al. [4] is particularly relevant. They identified and represented landmarks on the world model as graphs based on the proximity of representation features, enabling the agent to acquire an abstracted decision space independent of the environment autonomously. Gopalakrishnan and Kambhampati proposed a state-space abstraction method that reduces the number of divergence points of action strategies to increase the predictability of the agents' actions [5]. However, the decision space generated by these methods is not necessarily interpretable by humans, and the results need to be labeled, i.e., assigned meaning, by humans. Reference [6] proposed a method for generating explanations using autonomously learned world models. However, this method assumes an environment where observation information is discrete and interpretable. It has not yet achieved the autonomous acquisition of interpretable decision spaces.

B. REQUIREMENT 2: THE MODEL OF OTHERS

Clair and Matarić proposed a framework for estimating plausible policies based on human actions, assuming that humans and agents hold the same set of policies [7]. Gao et al. proposed a framework for estimating plausible policies currently assumed by the user, considering the user's actions and the interaction history [8]. Huang et al. prepared several policy estimation methods and definitions of plausibility in inverse reinforcement learning and showed that the policies reproduced differ even for the same information presentation depending on these differences [9]. Lage et al. also showed that the accuracy of recovering a sequence from a summary of a sequence of actions varies depending on differences in human policy recovery models [10]. These studies have proposed methods for estimating user policies

[1] In this paper, we consider the explainability for autonomous agents that have acquired policies through reinforcement learning.

and have emphasized the importance of modeling the humans who receive explanations. However, no suitable method for estimating the decision space of users has been proposed.

The authors of this paper proposed an estimation method of the user's model using Graph2Vec and concept activation vector (CAV) [11]. Although some issues remain, the method of [11] is used as the estimation method for the user's model in the integrated XAR system as the experiments in this study are conducted in settings where this method can be used.

C. REQUIREMENT 3: INFORMATION NEEDED FOR EXPLANATION

References [9] and [10] assume that humans use an inverse reinforcement learning or imitation learning framework to recover policies and extract those factors that yield the highest accuracy in recovery as critical information. In [12], the Markov decision process (MDP) is assumed, and the "best-expected reward" and "worst-expected reward" are obtained for a certain state when selecting an action. The significant difference between states is extracted as a critical factor. In [13], in addition to explanations using examples included in the training data, the state in which the value of the action at the next time is particularly high when the action is fixed to the optimal action is presented as a key factor. In [14], the authors calculated the difference between the maximum and minimum Q-values in each state on the MDP. They presented the elements with the most significant values as scenes representing the agent's characteristics. In the study by Sequeira and Gervasio [15], states with a high frequency of occurrence and high variance in action selection frequency are extracted, and the scenes are summarized as video clips. In [16], the explanation is generated by projecting the nodes of an agent's action decision tree with depth constraints onto a causal graph on the input features rather than an MDP.

Sakai et al. proposed a method to identify actions that are important for reaching a target state by approximating the causal effect of each action on the probability of reaching the target state [6]. Although this method can generate explanations even when important factors do not appear in the action values, it has two issues: 1) it requires random search, and the results of the importance calculation depend on the depth of the search, and 2) the semantics of the behavior may not match the user.

D. REQUIREMENT 4: PRESENTATION TO USER

In [17], the set of states in which an agent chooses a particular action on the MDP is obtained. The agent's policy is verbalized by presenting a set of languages representing those states. In addition, Waa et al. generated an explanation for the question "Why did you choose a^+ instead of a^-?" by inferring the states and consequences that would result from the action and presenting the language associated with the states and consequences [18]. Furthermore, a framework for generating verbal explanations directly from an agent's state sequence using an encoder-decoder model was proposed in [19] and [20].

In [15], important scenes are presented as video clips to generate human-acceptable explanations. Huber et al. applied a saliency map, which is conventionally used to improve the interpretability of image classification tasks, to highlight important parts of an image representing a specific scene to improve the interpretability of the policy [21].

These methods can present information using interfaces such as language and images that are easy for humans to understand. However, in interpersonal explanation, it is desirable to transform the information to be explained and how the explanation is presented according to the user. Yeung et al. proposed a framework that enables the selection of the optimal explanation presentation method by incorporating the processes of explanation presentation and user comprehension into a reinforcement learning framework [22].

E. CONTRIBUTIONS OF THIS PAPER

The contribution of this paper is as follows: we constructed an entire explanation algorithm that satisfies requirements 1-4 and conducted a large-scale subject experiment. No implemented algorithm satisfies requirements 1–4 so far to the best of our knowledge. The experimental protocol follows [6], but the number of subjects is large, and the relationship between good and bad explanations and subjective impressions of the agent is also investigated. We evaluated the algorithm and obtained valuable insights for future research from these results.

III. ALGORITHMS FOR XAR

A. OVERVIEW OF THE PROPOSED XAR

The overview of the XAR algorithm realized in this paper is shown in Fig.2. The red dashed line in this figure illustrates the realization of the entire XAR defined in [3]. Not all elements are necessarily involved, and it is also possible to generate the explanation by excluding some elements from the overall algorithm. The blue and yellow dashed lines in Fig.2 show the algorithm configurations without some of the features as a comparison method in the experiments described later in this paper.

B. DETAILS OF EACH ALGORITHM

1) INTERPRETABLE DECISION-MAKING SPACE [REQUIREMENT 1]

This paper assumes that the robot represents its actions and the external world in a discrete state space, such as the grid environment [23]. First, the robot learns a policy corresponding to the environment in which it acts by reinforcement learning. Then, using the data obtained in the search process during policy learning, the relationship between actions and state transitions in the discrete state space is represented as a graph, which is used as the world model. Through this series of searches and learning, we assume that the autonomous robot satisfies requirement 1.

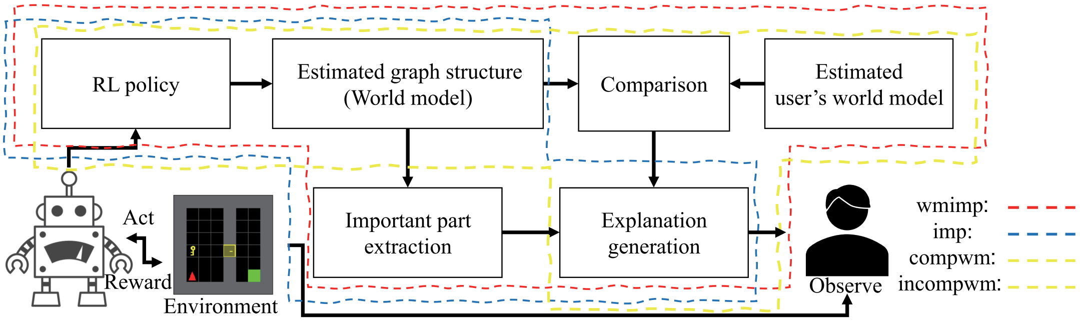


FIGURE 2. Overview of the XAR algorithms used in the experiments.

However, other problems in acquiring an interpretable world model require further discussion. For example, there is no guarantee that the state space autonomously acquired by the robot is comprehensible to humans. In this paper, we consider a discrete environment called a grid world, and the state space is consistent with human intuition, so this is not a significant problem. Conversely, this problem is severe when dealing with continuous spaces. It is necessary to consider methods for constructing world models that consider human interpretability, such as including language and human feedback.

2) MODEL OF OTHERS [REQUIREMENT 2]

As mentioned earlier, existing research focused on policies and planning algorithms. However, robots in the real world are designed to behave as humans expect them to behave, and their ultimate goals are shared with the user. In such situations, questions about the results of action decisions often stem from discrepancies in environmental awareness. In this paper, we implement a method to estimate the user's world model from the robot's world model and the questions (queries) given by the user.

The method implemented in this paper first creates a space with a distributed representation of graphs of its world models using Graph2Vec. Then, it estimates the CAV [24] from the opponent's question and estimates it by translating its world model. For algorithm details, please refer to [11].

3) EXTRACTION OF INFORMATION NEEDED FOR EXPLANATION [REQUIREMENT 3]

We proposed a method for identifying scenes that are important for reaching the goal state by approximating the causal effects of actions in each scene on the plan on the probability of reaching the goal state [6]. This study uses probabilistic policy and subjective observation information based on the method of [6]. Further, we also introduce a conceptualization of actions. These improvements improve the accuracy and versatility of important scene extraction. The details of the algorithm are described in the appendix.

4) EXPLANATION TO THE USER [REQUIREMENT 4]

This paper focuses on image information as the final explanatory interface (for details, please refer to section IV). In addition, we investigate a simple linguistic method of explanation in conjunction with estimating the user's model. Using different methods of presenting explanations according to the user is essential in interpersonal explanation but is outside the scope of this study.

C. METHODS OF COMPARISON IN EXPERIMENTS

The following six methods are algorithms that we evaluate in later experiments.

1) **WMIMP CONDITION**: ESTIMATION OF USER's WORLD MODEL AND EXTRACTION OF IMPORTANT INFORMATION (RED DASHED LINE IN FIG. 2)

The agent's world model is compared with the user's world model (estimated one), and the most recent scene with a different state transition structure is presented. However, only important scenes are presented, and only scenes that occur during the transition from the query state to the target state are presented. The world model obtained during policy learning is used to understand the state transition structure.

2) **IMP CONDITION**: QUERY AND EXTRACTION OF IMPORTANT INFORMATION (BLUE DASHED LINE IN FIG. 2)

No estimation of the user's world model is performed. Instead, the state in which the query is given is used as a cue. In other words, it presents the most recent important scene among the scenes that occur during the transition from the state in which the query is given to the target state.

3) **COMPWM CONDITION**: ESTIMATION OF USER's WORLD MODEL (YELLOW DASHED LINE IN FIG. 2 WITH IDEAL WORLD MODEL)

The agent's world model is compared with the user's world model (estimated model), and the most recent scene with a different state transition structure is presented. However, only scenes during the transition from the state given by the query to the target state are presented. The ideal world model, a perfect copy of the state transitions in the real environment, is used to understand the state transition structure.

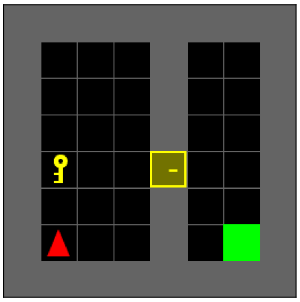


FIGURE 3. **Environment used in the experiments. Keys and doors are in different positions.**

4) INCOMPWM CONDITION: ESTIMATION OF USER's WORLD MODEL (YELLOW DASHED LINE IN FIG. 2 WITH LEARNED WORLD MODEL)

The agent's world model is compared with the user's world model (estimated model), and the most recent scene with a different state transition structure is presented. However, only the scenes during the transition from the state given by the query to the target state are presented. The world model obtained during policy learning is used to understand the state transition structure.

5) RANDOM CONDITION: RANDOM PRESENTATION

This method presents a random scene from among the processes during the transition from the state given the query to the target state.

6) ALL CONDITION: ALL ACTIONS PRESENTED

The method presents all the processes during the transition from the state given the query to the target state.

IV. EXPERIMENTS

A. TEST ENVIRONMENT

We apply the proposed method to an agent that learns its policy using proximal policy optimization (PPO) [25] in a simulation environment and evaluate its usefulness through subject experiments. A partially modified grid environment [23] with multiple objects, as shown in Fig.3, is used for the experiments. In this environment, the agent (triangle) gets rewarded by taking a key, opening a door, and reaching a goal in the lower right corner. The maximum value of the reward is 1, and it decreases with the number of actions required to reach the goal. The position of the goal is unchanged, but the positions of the key and the door change every trial. The agent has five actions: go straight, turn left, turn right, take the key from the grid in front of it, and open the door. The agent observes the absolute position of the key (x, y-coordinate), the absolute position of the door (x, y-coordinate), its absolute position (x, y-coordinates) and orientation, holding/not holding the key, and opening/closing the door, for a total of 9 dimensions.

B. OVERVIEW OF THE EXPERIMENT

Using the grid environment shown in Fig.3, we analyze whether the explanations generated by the implementation algorithm are useful as interpersonal explanations. The experiment was conducted online on 99 male and female subjects aged 18 or older who agreed to participate in the experiment and passed a screening to ensure the reliability of their responses. To understand the explanations provided by the presentation of the grid map, one must understand the meaning of the grid map. This screening includes questions regarding the comprehension of the grid map. The subjects are as follows: 2 subjects aged 18–19 (0 males, 2 females), 2 subjects aged 20–29 (1 male, 1 female), 11 subjects aged 30–39 (6 males, 5 females), 27 subjects aged 40–49 (20 males, 7 females), 27 subjects aged 50–59 (19 males, 8 females), 21 subjects aged 60–69 (19 males, 2 females), and 9 subjects aged 70 and older (9 males, 0 females).

In this verification, all subjects will evaluate six scenarios of explanation. Each scenario proceeds as follows. The "robot" in this experiment refers to the red agent.

(1) Each subject is given the robot's current position, the state of holding the key, the state of the door, and a map of the room (Fig.4). At the same time, the robot's next action is also displayed.
(2) As the presented next action is not optimal given the map, the robot is given the query "action a should be chosen in the presented state" (this process is fixed, and the subject is shown a predefined query).
(3) The robot explains based on the given query. The explanation presented varies from subject to subject, and one of the explanations generated by the six methods described in the previous section is presented at random.
(4) Subjects answer questions that measure the quality of the explanation.

An example of the explanation generated by the wmimp condition (III-C1) is shown in Fig.5. In the upper part, possible state transitions in the robot's recognition of its environment are displayed, and only the points where state changes are allowed are shown on the grid. Only the key is shown in the scene where the robot takes a key, and no doors or walls are shown. Similarly, only the red agent is shown in the scene where the robot moves straight ahead. The explanations generated under other conditions are in the same format, differing only in the scene presented. Even-numbered scenarios are those in which the correct other-world model (i.e., the room map presented to the subject) can be estimated from the query. Odd-numbered scenarios are those in which the estimation fails. Subjects evaluate all scenarios 1 through 6, but each scenario provides only one type of explanation. The questions are as follows, all using the seven-point scale.

Q1 Which of the following four maps do you think is the map of the room the robot is supposed to be in? The correct answer is always one of these, but if you do not know, choose "I don't know".
Q2 Was the robot's description concise (simple and free of unnecessary information)?
Q3 Was the robot's explanation a "good explanation" for you?

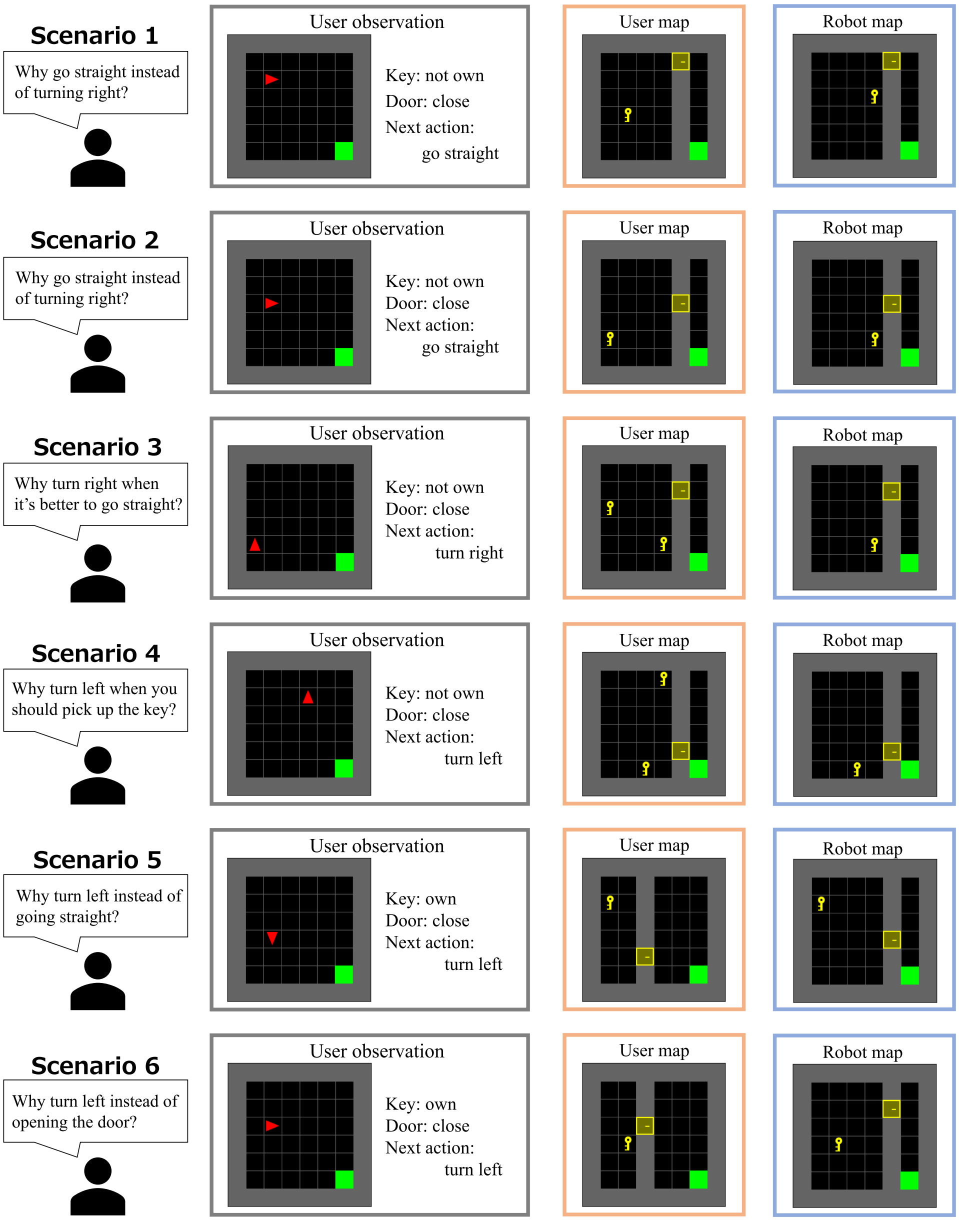


FIGURE 4. Information for each scenario. Subjects are not presented with the robot's perception of the environment.

Q4 Assume you have been doing surveillance work for a long time and are familiar with the task. Every time you ask a question, the robot offers this explanation. In this situation, do you think the robot's explanation is a "good explanation"?

Q5 Impression evaluation using the Godspeed questionnaire method [26].

Question 1 measures the likeliness of the explanation, i.e., whether the world model assumed by the robot is correctly conveyed to the subject. Question 2, limited to those who answered question 1 correctly, measures whether the explanation is concise and free of redundant information. Question 3 measures the overall satisfaction with the explanation, and question 4 measures the change in the evaluation when long-term use is assumed. The impression evaluation in question 5 uses Godspeed [26] with a seven-point scale. This method employs the SD method [27], which uses adjective pairs with opposite meanings and can measure anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety among people's impressions of the robot. Several authors pointed out that the adjective pairs of "calm – upset" and "calm – surprised" in the "safety" evaluation item were reversed in terms of the adjectives used for positive and negative evaluations [28], [29]. Therefore, the adjectives used for positive and negative evaluations are swapped in this verification.

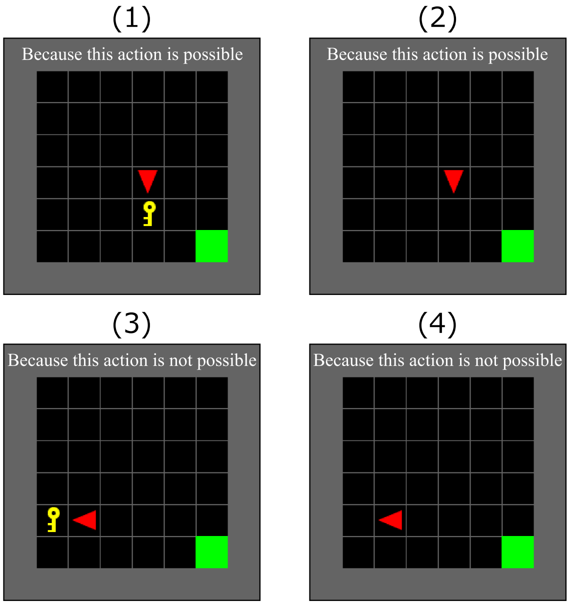

(1)
Because this action is possible
(2)
Because this action is possible
(3)
Because this action is not possible
(4)
Because this action is not possible

FIGURE 5. Example of the explanation presented. Images (1) to (4) are presented in order.

Because question 1 is a categorical result in this experiment, a statistical hypothesis test was conducted using the Steel method, which is nonparametric and suitable for testing multiple groups.

C. OVERALL ANALYSIS

The mean values for all scenarios are shown in Tab. 1. Values in bold are the highest values, and values with a double star (**) indicate a significant difference at the significance level of $\alpha = 0.01$ relative to the highest value.

The percentage of respondents who answered the correct environment in question 1 was high in the wmimp, imp, and all conditions and significantly low in the other conditions. In particular, the values for the incompwm and random conditions were similar to those obtained by selecting a random response, indicating that many of the scenarios did not make sense as explanations. In addition, the compwm condition produced significantly lower average values than the top three methods, as some scenarios could generate meaningful explanations while others could not. From this, it can be said that even if a complete world model can be learned, simply presenting the differences in the state transition structure of the world may not be sufficient as an explanation. For example, in scenarios 5 and 6, the x-coordinates of the door (in the direction of the horizontal axis) are different. The important scene is not the scene where the door is opened, but the scene where the agent moves in the x direction, and the compwm condition could not present the location of the door to the user. This indicates that the extraction of important scenes is indispensable for explanation generation. In the wmimp condition, the world model obtained during policy learning was used. Nevertheless, in all scenarios, the explanations generated were identical to those generated using a world model that completely copied the state structure of the real environment. This indicates that the important scene extraction absorbed a certain degree of error in the world model.

Question 2 was administered only to those who answered Question 1 correctly. However, there was no correlation between the number of scenes presented and brevity; the evaluation value of the ‘‘random’’ condition was significantly lower. This may be because, compared to the other explanations, the correct answers for the random condition were given without any certainty and thus were not evaluated as presenting the minimum necessary information. Although the ‘‘all’’ condition presented significantly more scenes than the other conditions, it is thought that certainty was evaluated, and the explanation of the intermediate transitions was not necessarily regarded as unnecessary information. From these results, it can be said that whether the explanation is concise depends not only on the number of elements presented but also on various factors, such as whether it helped the user understand the event. This is consistent with the findings of loveliness, suggesting that it is essential to transform the content of the explanation according to the user's skill level in the task and the ability to understand the explanation [30].

Like Question 1, Question 3 also obtained high evaluation values in the wmimp, imp, and all conditions. Since this experiment was conducted on a wide range of subjects, it was expected that they would have more difficulty understanding the experimental setup and grasping the content of the explanations than the more experienced participants. However, even among such subjects, the explanations generated in the wmimp and imp conditions obtained as high evaluation values as those in all conditions. In other words, the explanations generated by the proposed method obtained the same level of evaluation as the whole-sequence explanations by presenting significantly fewer scenes than the whole-sequence explanations. In the real environment, reducing the explanation time is important for safety management and productivity improvement, suggesting the superiority of the explanations generated by the proposed method.

Question 4 was intended to measure the explanation's quality in long-term use, and it yielded almost the same results as question 3. However, the evaluation value may change when the user uses the system for a long time since the user's proficiency in the task and the ability to read the necessary information from the explanation are expected to improve significantly. In particular, under all conditions, there is a possibility that the evaluation will decrease if the efficiency of information transfer is emphasized since much information is unnecessary for estimating the correct map. Detailed experiments are needed to examine changes in the evaluation over a long period of use.

Next, we present the results of Godspeed. The reliability of the scale scores is recognized when Cronbach's α coefficient exceeds 0.7. In this experiment, $\alpha = 0.84$ for anthropomorphism, $\alpha = 0.88$ for animacy, $\alpha = 0.95$ for likability, $\alpha = 0.97$ for perceived intelligence, and $\alpha = 0.94$ for perceived

TABLE 1. Averages of all scenarios. Values in boldface indicate the highest value. Values with a double star (**) significantly differ from the highest value at the significance level $\alpha = 0.01$.

	question 1 accuracy	question 2 brevity	question 3 goodness	question 4 long-term
wmimp	0.61	5.14	4.53	**4.47**
imp	**0.68**	**5.27**	**4.55**	4.44
compwm	0.47**	4.93	4.15	4.08
incompwm	0.22**	4.57	3.63**	3.59**
random	0.20**	4.00**	3.20**	3.29**
all	0.66	5.01	4.43	4.46

TABLE 2. Results of Godspeed impression evaluation. The highest values are in bold, with a star (*) indicating a significant difference at a significance level of $\alpha = 0.05$ from the highest value. Values with a double star (**) indicate a significant difference at a significance level of $\alpha = 0.01$ from the highest value.

	anthropo.	animacy	likability	intelligence	safety
wmimp	3.77	3.94	**4.16**	**4.43**	**4.59**
imp	**3.83**	**3.95**	4.07	4.31	4.45
compwm	3.70	3.87	3.96*	4.29	4.40
incompwm	3.44**	3.64**	3.64**	3.96**	4.01**
random	3.39**	3.52**	3.53**	3.79**	4.03**
all	3.76	3.88	4.04	4.26	4.50

TABLE 3. Correlation coefficients between each item and the rating of Q3 (goodness of explanation). Values with a double star (**) indicate that a correlation was found at a significance level of $\alpha = 0.01$ due to an uncorrelated test.

question	anthropo.	animacy	likability	intelligence	safety
corr. coeff	0.59**	0.61**	0.78**	0.75**	0.65**

safety, and $\alpha > 0.7$ for all evaluation items. Therefore, all groups of questions were analyzed as is. The results for each explanation method are shown in Tab. 2. Values in bold are the highest values; values with a star (*) indicate that significant differences were found at a significance level of $\alpha = 0.05$ from the highest value. Values with a double star (**) indicate that significant differences were found at a significance level of $\alpha = 0.01$ from the highest value. As a result of the analysis, the explanation method with a higher evaluation (the value of question 3) obtained a higher impression evaluation in all evaluation items. In addition, as shown in Tab.3, there was a correlation between each evaluation item and the evaluation value of Explanation 3. Particularly strong correlations were found for likability and perceived intelligence. The distribution of the evaluation of each item is shown in Figs.6 (a), (b), (c), (d), and (e). Note that R^2 represents the coefficient of determination of the regression line.

The results of the no-correlation test by the corresponding two-tailed t test showed a positive correlation at a significance level of $\alpha = 0.01$ between the evaluation of explanation 3 and each evaluation item. This indicates that the quality of the explanation has a significant impact on the users' impression of the robot. In particular, it is important to present a good explanation of the robot's behavior for the user to feel a biological closeness to the robot and to place trust in the performance of the robot itself.

D. ANALYSIS BY SCENARIO

The overall analysis suggests that the wmimp, imp, or all conditions should be used for explanation generation. Therefore, in this section, we focus on these three conditions and compare the evaluation values of each scenario. The questions to be compared are question 1 (accuracy of environmental estimation) and question 3 (evaluation of the quality of the explanation). The analysis results for each scenario are shown in Tabs.4 and 5.

1) SCENARIOS 1, 2, 5, AND 6

In scenarios 1, 2, 5, and 6, the explanations generated by the imp condition consisted only of information necessary for environmental estimation. In contrast, the explanations generated by the other conditions included information not necessary for environmental estimation. The analysis results show that the imp condition produced a relatively stable high accuracy rate and explanation evaluation value. However, there were no significant differences in accuracy rate and explanation evaluation. The "all" condition produced many highly accurate scenarios, but the explanation evaluation tended to be lower than the "imp" condition. The only scenario with a higher explanation evaluation value in all conditions than in the imp condition was scenario 5. Users were confused and requested a more detailed explanation since the door was in a different environment from scenarios 1 to 4, whereas the key was in a different location in scenario 5. In addition, since the wmimp condition included the "presentation of a situation in which the key cannot be removed," it is thought that the user was burdened with accepting the explanation, resulting in many scenarios with a low accuracy rate. From the above, it is considered appropriate to present explanations in the imp condition when the imp condition can generate sufficient explanations. For tasks that users are unfamiliar with, it is also effective to generate explanations using all conditions.

2) SCENARIO 3

In scenario 3, none of the explanations lack the information necessary to estimate the environment. They can tell which coordinate has the key but cannot adequately indicate that the other coordinate does not. Under these conditions, the correct response rate was higher for the imp and all conditions, and the explanation evaluation value was higher for all conditions. The wmimp condition presented "infeasible transitions," but since it was based on an incorrect estimation of the user's world model, it is considered to have increased unnecessary information, resulting in lower accuracy. The explanation of the "all" condition obtained a high explanation evaluation value because the user requested a more detailed explanation because the environment with two keys was used from Scenario 3. In contrast, only one key was used in Scenarios 1 and 2, and a high explanation evaluation value was obtained for Scenario 5.

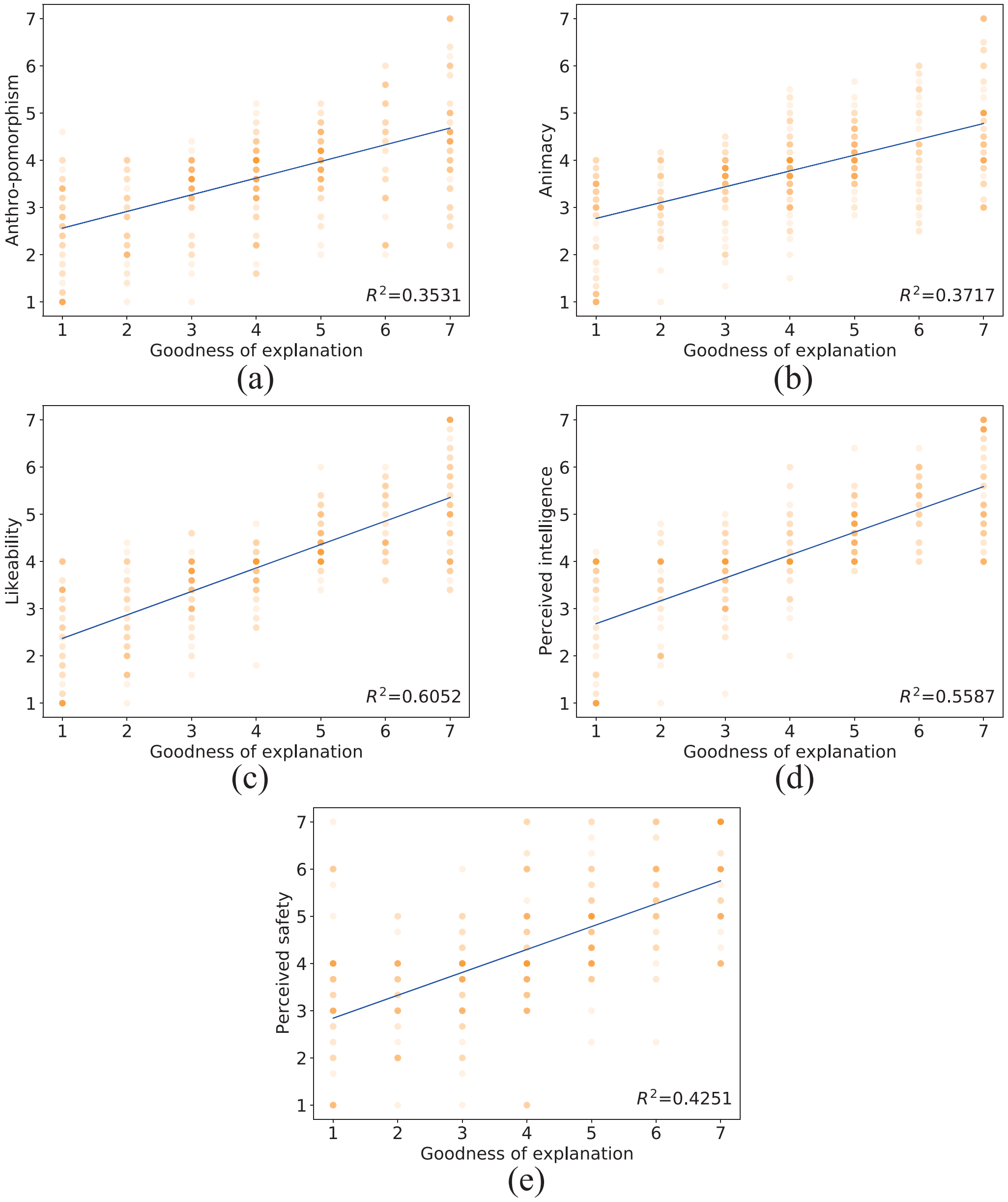


FIGURE 6. Scatter plots of the rating values for the goodness of explanation and Godspeed questionnaires: (a) anthropomorphism, (b)animacy, (c)likability, (d)perceived intelligence, and (e)perceived safety.

From the above, it can be said that, given a query that requires correct estimation of the user's assumed world model, it is effective to provide an explanation in the imp and all conditions when the user's world model cannot be estimated correctly.

TABLE 4. Accuracy by scenario.

scenario method	1	2	3	4	5	6
wmimp	0.67	0.74	0.48	0.61	0.45	**0.63**
imp	**0.71**	**0.90**	**0.65**	0.50	0.60	0.60
all	0.53	0.82	0.63	**0.70**	**0.61**	0.60

3) SCENARIO 4

In scenario 4, only the wmimp condition presents the necessary information. In other words, the wmimp condition adequately indicates that the key is not in the user's expected location. However, the correct response rate was the highest for all conditions. This may be because some users confused ''infeasible transitions'' presented in the wmimp condition with ''feasible transitions'' and selected the wrong map. On the other hand, the wmimp condition had the highest explanation evaluation value. These results suggest that the wmimp condition is the best explanation when the world model can be correctly estimated for queries that require correctness (negation) of the world model the user

TABLE 5. Evaluation by scenario.

method \ scenario	1	2	3	4	5	6
wmimp	**4.93**	**5.13**	3.83	**4.64**	4.55	4.16
imp	4.90	5.10	4.53	3.70	4.00	**4.50**
all	3.94	4.41	**4.96**	4.13	**4.89**	4.07

assumes. Although all conditions had the highest rate of correct answers, it is thought that the interface, not the content of the explanation, greatly impacted the rate of correct answers. In the next section, we will examine this hypothesis by presenting the results of a simplified version of Scenario 4, in which the explanatory interface is changed to language.

E. IMPRESSION EVALUATION EXPERIMENT WITH VERBAL EXPLANATION PRESENTATION (ADDITIONAL EXPERIMENT)

In Scenario 4, the interface is simplified, and the wmimp, imp, and all conditions are compared again. In this verification, the subject is presented with the same object arrangement as in Scenario 4, the environment shown in Fig.7. In this experiment, we consider handing a key to a user in the center of a room. We also assume a situation in which there is no key in key location A, located near the robot, and the robot moves toward the direction of key location B. The subject is asked to explain the reason for the robot's action. The following verbal explanation is provided for the question asking the reason for the action and is evaluated.

wmimp condition

I am going to get the key at location B because there is no key at location A now.

imp condition

I am going to get the key at location B.

all condition

After going straight south, I am going to pick up the key at location B, bring the key to you, and hand it over to you.

In the wmimp condition, the agent's world model is compared with the user's model (estimated model). The most recent important scene with a different state transition structure is presented. Thus, the Wmimp condition explains state transitions that the user assumes are feasible but are not feasible (i.e., the key cannot be obtained at location A) and state transitions that the user assumes are not feasible but are feasible (i.e., the key can be obtained at the location B).[2]

The imp condition presents the most recent important scenes when transitioning from the query state to the target state. Therefore, it explains taking the key at location B. In all conditions, all the processes that pass through when transitioning from the state where the query is given to the target state are presented. Therefore, it explains all actions, including those other than the key-obtaining scene.

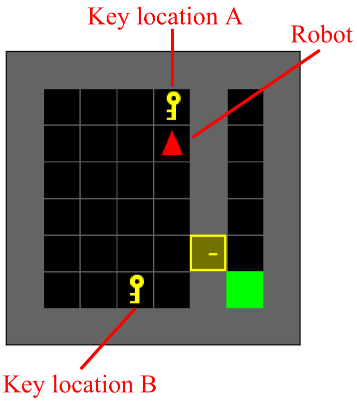


FIGURE 7. Environment used for additional user experiment.

In this experiment, each subject was presented with all of the above explanations and asked to choose from the following options the current situation in the key location area for each explanation. This question corresponds to Question 1 in the previous experiment, and the percentage of subjects who chose option 2 is the rate of correct responses.

(1) Location A has a key, but location B does not.
(2) Location A has no key, but location B does.
(3) There is a key to locations A and B.
(4) There are no keys to either location A or B.
(5) I do not understand the explanation.

Subjects were also asked to rank the explanations generated by the three conditions according to the explanation they felt was the best. This question corresponds to Question 3 of the previous experiment and measures the quality of the explanations. The experiment was conducted online on 24 male and female subjects aged 18 and older: 21 subjects aged 20–29 (11 males and 10 females), 1 subject aged 30–39 (0 males and 1 female), and 2 subjects aged 50–59 (1 male and 1 female). There was no overlap between the subjects in the previous experiment and those in the present.

Tables 6 and 7 show the results. The wmimp condition was the best for the correct response rate and the explanatory evaluation value. Statistical hypothesis testing using the Steel method revealed that $t(23) = 5.16$, $p < 0.001$ for the wmimp and imp conditions in accuracy and $t(23) = 5.43$, $p < 0.001$ for the wmimp and all conditions in accuracy. A comparison of the explanatory ratings of the wmimp and imp conditions showed that $t(23) = 5.61$ and $p < 0.001$, and a comparison of the explanatory ratings of the wmimp and all conditions showed that $t(23) = 5.63$ and $p < 0.001$. These results suggest that the explanation of the wmimp condition is the best explanation for queries that need to correct (negate) the user's assumed world model, if the world model can be correctly estimated.

Presenting explanations in language rather than images is important in promoting user acceptance of explanations.

[2] In the experimental setup, the user understands that they can get the key at the location B. However, in this experiment, the world model in which the key cannot be taken at location B was estimated as the user's world model, which also explains that the key can be taken at location B.

TABLE 6. Accuracy in an impression evaluation experiment using verbal explanation presentation.

method	accuracy
wmimp	0.92
imp	0.17
all	0.13

TABLE 7. Evaluation of the impression evaluation experiment using verbal explanation presentation.

goodness / method	best	second best	worst
wmimp	22	1	1
imp	1	12	11
all	1	11	12

More systematic knowledge is required to identify explanatory factors and design interfaces that encode and present these factors.

F. SUMMARY OF USER EXPERIMENTS

The following is a summary of the analysis results obtained by the user experiment.

- The average accuracy rate and evaluation value for all scenarios are high for the wmimp, imp, and all conditions. Considering the practical aspect, the wmimp and imp conditions are superior because they can provide explanations in a short time.
- It is important to present a good description of the robot's behavior, so the user can feel a biological closeness to the robot and place trust in the performance of the robot itself.
- For the explanation "Why A?", generating an explanation in the imp condition is effective.
- For the explanation "Why A instead of B?", generating an explanation in the wmimp condition is effective. Nevertheless, it is necessary to obtain a correct reason about the user's world model.
- Since the acceptability of explanations varies greatly depending on the method of presentation, an appropriate interface design is required.

V. DISCUSSION

A. CONSIDERATION OF DIFFERENCES IN USER MODELS

In this paper, we focused on the user's world model, but there are user models to consider in addition to the world model. In particular, estimating the states and policies the user assumes is very important from the viewpoint of explanation generation. Even if the user and the robot share the same world model, different states assumed by the user and the robot will result in different behaviors. Even if the user and the robot share the same world model and state, the different policies will naturally result in different behaviors. Therefore, in addition to the possibility of different world models, agents must consider the possibility of different assumed states and policies.

The most important aspect about this problem is that all these possibilities must be considered simultaneously. Even if a user provides a query in response to an action taken by a robot, it is not easy to determine whether the query is due to a difference in the world model, a misperception of the robot's situation, or because the user has a better policy. To make this judgment, it is necessary to integrate a model of the user's knowledge obtained through interaction with the user, social norms, common sense, and a vast amount of other information.

However, we humans always execute such reasoning unconsciously to achieve smooth communication. Considering the explainability of autonomous robots is equivalent to considering the communication between robots and humans [3]. For a robot to become an equal partner to a human, it is essential to handle a vast amount of information in an integrated manner and to infer the internal states of others appropriately, and it is necessary to find a way to do so.

B. ESTIMATION OF STATE REPRESENTATION OF OTHERS

In this study, we assumed that the state representation of the robot's world model and the user's state representation are equivalent. We also realized the inference of the user's world model based on the robot's state representation. In real environments, however, the state representation of robots and humans with different physical characteristics may differ. Furthermore, the state representation differs for each user, which may affect the user's decision-making tendencies. Therefore, it is necessary to estimate the user's state representation to realize a more user-friendly explanation in a real scenario.

Although the method of estimating the state representation is not obvious, assuming that the robot and the user observe the same state through different feature extractors and learn the state representation, the robot's state representation and the user's state representation can be converted using some nonlinear mapping function. Using several samples, it is possible to obtain a mapping function from the robot's state representation to the user's state representation. The same argument can be applied not only to the state representation of the world model but also to the state representation of the internal state, including policy (or objective function) and the state of the external world if such a representation exists. This mapping function can be thought of as a kind of model of others, and it may be possible to efficiently estimate the behavior of others based on the robot's knowledge.

C. LEARNING OF WORLD MODEL IN REAL ENVIRONMENT

One of the key issues in utilizing world models in real-world environments is the setting of the range to be represented by the world model. For example, in the case of a home robot, the cost of maintaining world models for multiple rooms increases exponentially compared to maintaining a world model for a single room with different object arrangements.

An approach that maintains multiple partial world models and integrates them as necessary is considered effective in addressing this problem. In the same way that we focused on important features when we extracted important scenes in this study, there are differences in world models that are important or unimportant for the current goal (objective function). It is desirable to establish a method to construct a world model by considering the significance of each partial world model in terms of policy and integrating them.

D. DEVELOPMENT OF EXPLANATION INTERFACE

Although the proposed implementation algorithm does not include an explanation interface, the results of user experiments suggest that the acceptability of explanations varies greatly depending not only on the content but also on the method of presentation. Currently, explanation interfaces are designed by humans according to the task. However, verifying which interface is effective for what kind of explanation presentation is being actively promoted is hard. To present useful explanations to humans, it is important to systematically organize the findings of cognitive science and the results of user experiments and establish a general theory of interfaces. Establishing such a systematic theory will be an important stepping stone for robots to autonomously determine the explanation interface in the future.

E. TEMPORAL VARIATION IN EXPLANATION EVALUATION

In evaluating the explainability of autonomous robots and agents, including this paper, there are few examples of long-term interactions with users. However, it is conceivable that the user's internal state certainly changes as they receive explanations and that the user's evaluation of the explanations also changes over time. An explanation initially perceived as "polite" may be "too time-consuming" or trust in the robot may be undermined by presenting an incorrect explanation multiple times. Investigating how impressions of robots change through long-term interaction between robots and users is also very important in explainability.

F. EXPLANATION AS COMMUNICATION

In addition to the issues described so far, many other research issues must be resolved when considering explanation as communication. One important issue is determining the necessity of explanations. In this study, explanations were generated based on queries provided by users, but in the real world, users may not explicitly provide queries. In such cases, the necessity of an explanation must be determined based on the user's behavior and past interaction history. Especially in situations where the user does not continuously monitor the robots, such as when one user operates multiple robots, spontaneous explanation from the robot can prevent the user from acting contrary to the user's intention.

Another important issue is to analyze the effect of explanations. Investigating the change in trust with the user when the robot presents an explanation and the positive and negative effects when the robot generates a false explanation is important when considering explanation strategies. These investigations span cognitive science, and interdisciplinary progress in the research area is desirable. Moreover, explainability can be applied to "education" to make users understand what is correct, "persuasive dialogue" to make users feel as if they understand, not whether or not they are correct, and "creativity" to add values to things that are not thought of by common sense.

VI. CONCLUSION

In this study, we implemented an explanation generation framework that meets the requirements for realizing XAR. These robots can explain the reasons for their action decisions; we verified their effectiveness through experiments on human subjects. The experimental results suggest that the implemented explanation framework can generate explanations acceptable to humans. It was also found that how the explanations should be generated depends on the question's content and the explanation's interface. Future work includes developing an explanation framework for more complex situations such as continuous state spaces. Further validation through experiments using more complex tasks and actual robots is also needed. Examining how different media affect impressions of explanations is also expected in future work.

APPENDIX

A. EXTRACTION OF INFORMATION NEEDED FOR EXPLANATION (REQUIREMENT 3)

The algorithm implemented in this paper extends the method of [6] and identifies important scenes by the following procedure.

(1) **Acquisition of action concepts:** Based on the robot's subjective observation, the policy implications of each action in the plan are identified. This meaning of action is defined as an action concept.

(2) **Calculation of the importance of each action concept:** We extend the method of [6] to be used with probabilistic policy. Important scenes are identified by calculating the average treatment effect on the expected reward of the feasibility of each action concept.

The main differences between the proposed method and the method in [6] are as follows.

- **Use of probabilistic policy**
  The method in [6] requires a random search to identify important scenes and the result of the importance calculation changes depending on the depth of the random search. This proposal uses a probabilistic policy instead of a deterministic one to eliminate the random search process and reduce the number of hyper-parameters.
- **Use observations from subjective viewpoints**
  The method in [6] considers the causal effect of taking a specific action in a certain state on reaching the target state. However, there is a problem that even when the relationship between the object and the target of the action is the same, the difference in absolute

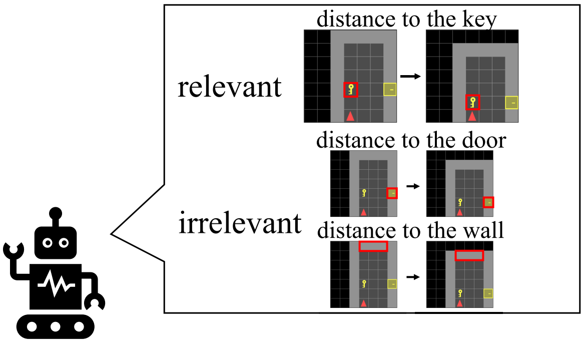


FIGURE 8. Schematic diagram of the action concept. Among the changes in the observed information due to the robot's actions, only the change in the approach to the key is important for the policy. Therefore, we do not focus on the change in the relative coordinate to the door or wall but only on the change in the relative coordinate to the key to obtaining the action concept.

coordinates causes the object to be recognized as a different state. The proposed method solves this problem by using observations from a subjective viewpoint.

- **Using the difference between the states of before and after action**
  In the method in [6], actions are classified at the command level, and causal effects are calculated. However, even if the actions differ at the command level, the actual environmental effects may be the same. For example, even if a person tries to pick up a key or open a door in a space with no key nor door, it is equivalent to choosing the action of "doing nothing". In other words, it is an excessive classification to recognize "picking up the key" and "opening the door" as different actions in this situation. Therefore, the proposed method defines the action concept using the difference between states before and after the robot's action.
- **Focus only on policy-significant state changes**
  As mentioned earlier, the method in [6] considers the causal effects of certain actions in certain states. However, the meaning of the state changes depending on the policy. For example, when preparing to go out, a person goes to get the key on the desk at the end of the room. This action aims to get the key, not to go toward the desk. Therefore, in this paper, we define the concept of action by focusing only on the state change that is important for the policy (in the previous example, approaching the key).

B. ACQUISITION OF ACTION CONCEPTS

The proposed method identifies critical situations by calculating the average treatment effect of the feasibility of each action concept on the expected reward. Therefore, it is important to obtain appropriate action concepts. There are two ways of thinking about action concepts: one is to focus on the action commands (e.g., drive a certain motor, bend an arm) as the subject, and the other is to focus on the state changes as a result of the action. When considering interpersonal explanation, the user and the robot generally have different action commands as subjects, and the user is unaware of the robot's internal control methods. Here, action is defined by the state change δs.

$$\delta s = s_{t+1} - s_t \tag{B.1}$$

The state s_t, s_{t+1} is defined from the subjective viewpoint, as in partial observation. This aims to prevent the exact meaning of state change from being recognized as different action concepts depending on the placement of objects.

Another important perspective when considering action concepts is the meaning of actions regarding the policy. For example, when considering the meaning of a route to a shopping street, if the purpose is to shop at a grocery store, the action is "going to the grocery store," and if the purpose is to shop at a fish shop, the action is "going to the fish shop". If the purpose is to shop at more than one store, the action could be "head for the shopping district". It is desirable to conceptualize actions considering the meaning of the action in terms of the policy. Therefore, we define the concept of action by focusing only on the important characteristics of each state as illustrated in 8 as follows.

$$CA(s_t, s_{t+1}) = (\boldsymbol{w}^1, \boldsymbol{w}^2, \dots, \boldsymbol{w}^n), \tag{B.2}$$

$$\boldsymbol{w}^j = \begin{cases} (z_t^j, z_{t+1}^j)\ \mathit{for}\ \ z_t^j \neq z_{t+1}^j\ \ \mathit{and} \\ \qquad val(z_m^j) > \alpha \cdot \max_j val(z_m^j),\ \exists m \in \{t, t+1\} \\ \mathit{None}\ \ \mathit{for}\ \ z_t^j = z_{t+1}^j\ \ \mathit{or} \\ \qquad val(z_m^j) \leq \alpha \cdot \max_j val(z_m^j),\ \forall m \in \{t, t+1\}, \end{cases}$$

where $CA(\cdot)$ is an action concept, $val(z_m^j)$ is a function (feature extractor) that outputs the importance of the jth feature z_m^j in the state s_m at time m. n represents the dimensionality of the features, and $\alpha \in (0, 1)$ is a parameter that adjusts the region of the important features. In the case of $\boldsymbol{w}^j = (z_t^j, z_{t+1}^j)$, the change of the j-th feature from z_t^j to z_{t+1}^j is included in the action concept. When $\boldsymbol{w}^j = None$, the change of the j-th feature is not included in the action concept, and the change of z_t^j, z_{t+1}^j to any value does not affect $CA(s_t, s_{t+1})$. Using Eq.(B.2), the action concept is defined by a change in a feature that is considered an important feature in at least one of the states s_t, s_{t+1} defined from the subjective viewpoint. The features that do not change among the important features are not included in the action concepts.

We use Shapley Additive exPlanations (SHAP) [31] as a feature extractor. SHAP is a method for quantifying the contribution of input features to output results using the Shapley value of game theory. By applying kernel SHAP to the robot policy, we identify the important features in determining actions and use them to form action concepts. Specifically, $val(z_m^j)$ is the kernel SHAP value of the feature z_m^j. When forming action concepts using image information, $val(\cdot)$ can also be designed using a saliency map that visualizes which parts of the image are focused on and which results are output.

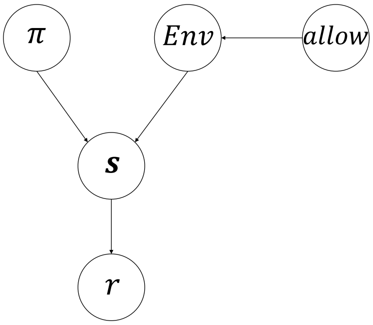


FIGURE 9. Causal graphs of variables in formulations using ATE.

C. CALCULATION OF THE IMPORTANCE OF EACH ACTION CONCEPT

An action concept is acquired for each state transition until the target state is reached. The average treatment effect (ATE) for the expected reward of being able to perform the action concept is calculated. This ATE is defined as the importance I of each action concept as in Eq.(C.3). In addition, as in the method of [6], we define the state transition corresponding to the action concept CA_f such that the value of $I(CA_f, \pi)$ exceeds a specific threshold value as an important scene.

$$I(CA_f, \pi) = \mathbb{E}[r|\pi, do(allow(CA_f) = 1)] - \mathbb{E}[r|\pi, do(allow(CA_f) = 0)], \qquad \text{(C.3)}$$

where $do()$ is Pearl's do operator [32], which indicates intervention. $allow(CA_f)$ takes the value 1 when an action concept CA_f can be executed and 0 when it cannot. The causal graph in this formulation is shown in Fig.9. Env is a variable representing the dynamics of the environment, and π, s, and r are the agent's policy, the state sequence it goes through, and the reward it acquires, respectively.

D. SIMULATION EXPERIMENT: COMPARISON OF IMPORTANCE CALCULATION RESULTS

We compare the importance obtained by the proposed method with that of the method in [6]. The initial state is shown in Fig.3. In this experiment, we use partial observations of 7 × 7 squares in front of the agent, as shown in Fig.10. The key holding/not holding is expressed in the coordinates of the agent's position, and the state of the door opening/closing is also expressed in the observation information. The hyperparameter in Eq.(B.2) is $\alpha = 0.8$.

First, an example of the behavioral concepts obtained by the proposed method is shown in Fig.10. If the agent and the key are drawn overlapping, the agent holds the key. In the scenes of acquiring the key (Step 2) and opening the door (Step 7), the action concept is focused on the object's coordinates. In the actions toward the goal coordinates (Steps 12 and 13), the action concept is focused only on the change in coordinates relative to the goal.

Next, the importance of each step is shown in Fig.11. Compared to the conventional method, the proposed method shows a significant difference in the importance of the steps near the key, door, and goal, which must be passed through, and the importance of the other steps. In particular, the key-obtaining step (Step 2) is not highly important in the conventional method because there are multiple possible directions to obtain the key. Still, it has very high importance in the proposed method.

These results suggest that the proposed method forms an action concept focusing on state changes important in acquiring rewards. It can calculate importance more accurately than the method in [6].

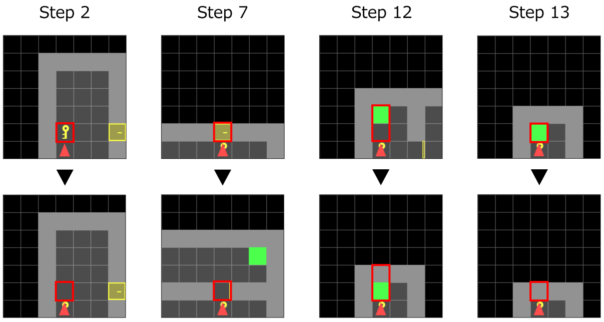


FIGURE 10. Acquired action concepts. The state change of the grid surrounded by a red frame represents the action concept.

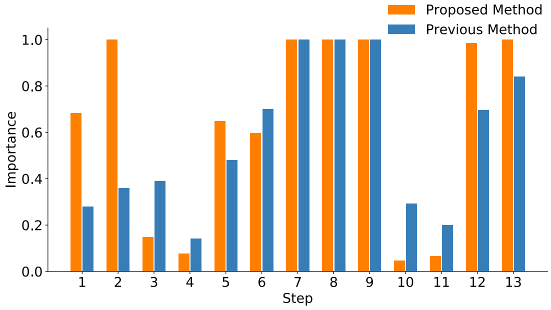


FIGURE 11. The importance of each step is as follows: step 2 is the scene of taking the key, steps 7 to 9 are the scenes of opening the door and passing through it, and step 13 is the scene of reaching the goal.

REFERENCES

[1] 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.

[2] C. Molnar. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[3] T. Sakai and T. Nagai, "Explainable autonomous robots: A survey and perspective," *Adv. Robot.*, vol. 36, nos. 5–6, pp. 219–238, Mar. 2022, doi: 10.1080/01691864.2022.2029720.

[4] L. Zhang, G. Yang, and B. C. Stadie, "World model as a graph: Learning latent landmarks for planning," 2020, *arXiv:2011.12491*.

[5] S. Gopalakrishnan and S. Kambhampati, "Minimizing robot navigation-graph for position-based predictability by humans," 2020, *arXiv:2010.15255*.

[6] T. Sakai, K. Miyazawa, T. Horii, and T. Nagai, "A framework of explanation generation toward reliable autonomous robots," *Adv. Robot.*, vol. 35, no. 17, pp. 1054–1067, Sep. 2021, doi: 10.1080/01691864.2021.1946423.

[7] A. St. Clair and M. Mataric, "How robot verbal feedback can improve team performance in human–robot task collaborations," in *Proc. 10th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, Mar. 2015, pp. 213–220.

[8] X. Gao, R. Gong, Y. Zhao, S. Wang, T. Shu, and S.-C. Zhu, "Joint mind modeling for explanation generation in complex human–robot collaborative tasks," in *Proc. 29th IEEE Int. Conf. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2020, pp. 1119–1126.

[9] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Auto. Robots*, vol. 43, no. 2, pp. 309–326, Feb. 2019.

[10] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir, "Toward robust policy summarization," in *Proc. 18th Int. Conf. Auto. Agents MultiAgent Syst.* Richland, SC, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 2081–2083.

[11] T. Sakai and T. Nagai, "Estimation of user's world model using Graph2vec," 2023, *arXiv:2301.03793*.

[12] O. Khan, P. Poupart, and J. Black, "Minimal sufficient explanations for factored Markov decision processes," in *Proc. Int. Conf. Automated Planning Scheduling*, Jan. 2009, pp. 194–200.

[13] T. Dodson, N. Mattei, and J. Goldsmith, "A natural language argumentation interface for explanation generation in Markov decision processes," in *Proc. Int. Conf. Algorithmic Decis. Theory*, Oct. 2011, pp. 42–55.

[14] D. Amir and O. Amir, "Highlights: Summarizing agent behaviors to people," in *Proc. 17th Int. Conf. Auto. Agents Multiagent Syst.*, Stockholm, Sweden, Jul. 2018, pp. 1168–1176. [Online]. Available: https://scholar.harvard.edu/files/oamir/files/highlightsmain.pdf

[15] P. Sequeira and M. Gervasio, "Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations," *Artif. Intell.*, vol. 288, Nov. 2019, Art. no. 103367.

[16] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Distal explanations for explainable reinforcement learning agents," 2020, *arXiv:2001.10284*.

[17] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *Proc. 12th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI*, Mar. 2017, pp. 303–312.

[18] J. Waa, J. Diggelen, K. Bosch, and M. Neerincx, "Contrastive explanations for reinforcement learning in terms of expected consequences," 2018, *arXiv:1807.08706*.

[19] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: A technique for explainable AI and its effects on human perceptions," in *Proc. 24th Int. Conf. Intell. User Interfaces*, Mar. 2019, pp. 263–274.

[20] D. Das, S. Banerjee, and S. Chernova, "Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery," 2021, *arXiv:2101.01625*.

[21] T. Huber, B. Limmer, and E. André, "Benchmarking perturbation-based saliency maps for explaining Atari agents," 2021, *arXiv:2101.07312*.

[22] A. Y. Yeung, S. Joshi, J. J. Williams, and F. Rudzicz, "Sequential explanations with mental model-based policies," 2020, *arXiv:2007.09028*.

[23] M. Chevalier-Boisvert, L. Willems, and S. Pal, "Minimalistic gridworld environment for OpenAI gym," Tech. Rep., 2018.

[24] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. ICML*, vol. 80, J. G. Dy and A. Krause, Eds. 2018, pp. 2673–2682. [Online]. Available: http://dblp.uni-trier.de/db/conf/icml/icml2018.html#KimWGCWVS18

[25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[26] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. J. Social Robot.*, vol. 1, no. 1, pp. 71–81, Jan. 2009.

[27] C. E. Osgood, "The nature and measurement of meaning," *Psychol. Bull.*, vol. 49, pp. 197–237, May 1952.

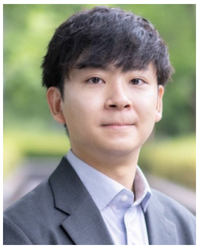
[28] G. Schillaci, S. Bodiroža, and V. V. Hafner, "Evaluating the effect of saliency detection and attention manipulation in human–robot interaction," *Int. J. Social Robot.*, vol. 5, no. 1, pp. 139–152, Jan. 2013.

[29] D. Karreman, G. S. Bradford, B. van Dijk, M. Lohse, and V. Evers, "What happens when a robot favors someone? How a tour guide robot uses gaze behavior to address multiple persons while storytelling about art," in *Proc. 8th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, Mar. 2013, pp. 157–158.

[30] T. Lombrozo, "Simplicity and probability in causal explanation," *Cognit. Psychol.*, vol. 55, no. 3, pp. 232–257, Nov. 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010028506000739

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[32] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.



TATSUYA SAKAI received the degree from the School of Engineering Science, Osaka University, in 2020, and the master's degree from the Department of Systems Innovation, Graduate School of Engineering Science, Osaka University, in 2022. During his studies, he engaged in research on autonomous agents and explainability of robots. He is currently with NEC Corporation.

TAKAYUKI NAGAI (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from the Department of Electrical Engineering, Keio University, in 1993, 1995, and 1997, respectively. From 2002 to 2003, he was a Visiting Scholar with the Department of Electrical Computer Engineering, University of California, San Diego. Since 1998, he has been with The University of Electro-Communications (UEC). Since 2018, he has been a Professor with the Graduate School of Engineering Science, Osaka University. He is also a specially-appointed Professor with the Artificial Intelligence Exploration Research Center (AIX), UEC, and a Visiting Researcher with the Brain Science Institute, Tamagawa University. His research interests include intelligent robotics, cognitive developmental robotics, and robot learning. He aims at realizing flexible and general intelligence like humans by combining AI and robot technologies. He was the IROS Best Paper Award Finalist. He received the Advanced Robotics Best Paper Award and the JSAI Best Paper Award.

KASUMI ABE received the B.Eng., M.Eng., and Ph.D. degrees from The University of Electro-Communications (UEC), in 2009, 2011, and 2015, respectively. She has been a JSPS Research Fellow, since 2015, and has been a Project Assistant Professor with AIX, UEC, since 2020. She is a Visiting Researcher with the Department of Agent Interaction Design, Advanced Telecommunications Research Institute International (ATR), and a Chief Researcher with ChiCaRo Company Ltd. Her research interests include human–robot interaction, robotics for children and childcare, and childcare support systems.