ARL-TR-9624 • JAN 2023

Learning to Understand Anomalous Scenes from Human Interactions

by Stephanie M Lukin, Rahul Sharma, and Michael Bellissimo

Approved for public release; distribution is unlimited.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9624 • JAN 2023

Learning to Understand Anomalous Scenes from Human Interactions

by Stephanie M Lukin
Army Research Directorate, DEVCOM Army Research Laboratory

Rahul Sharma
University of Maryland, College Park

Michael Bellissimo
Florida Atlantic University

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* January 2023	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** June 14–September 5, 2022

4. TITLE AND SUBTITLE
Learning to Understand Anomalous Scenes from Human Interactions

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Stephanie M Lukin, Rahul Sharma, and Michael Bellissimo

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLA-IC
2800 Powder Mill Rd, Adelphi, MD 20783

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9624

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Contact author email: <stephanie.m.lukin.civ@army.mil>

14. ABSTRACT

At the US Army Combat Capabilities Development Command Army Research Laboratory, we are studying behavior, building data sets, and developing technology for anomaly classification and explanation, in which an autonomous agent generates natural language descriptions and interpretations of environments that may contain anomalous properties. This technology will support decision-making in uncertain conditions and resilient autonomous maneuvers where a Soldier and robot teammate complete exploratory navigation tasks in unknown or dangerous environments under network-constrained circumstances (e.g., search and rescue following a natural disaster). We detail our contributions in this report as follows: we designed an anomaly taxonomy drawing upon related work in visual anomaly detection; we designed two experiments taking place in virtual environments that were manipulated to exhibit anomalous properties based on the taxonomy; we collected a small corpus of human speech and human–robot dialogue for an anomaly detection and explanation task; and finally, we designed a novel annotation schema and applied it to a subset of our corpus.

15. SUBJECT TERMS
anomaly detection, experimental design, data collection, data annotation, Military Information Sciences

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Stephanie M Lukin
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	30	**19b. TELEPHONE NUMBER (Include area code)** 310-448-5396

Contents

List of Figures

List of Tables

1. Introduction

At the US Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL), we are studying behavior, building data sets, and developing technology for anomaly classification and explanation, in which an autonomous agent generates natural language descriptions and interpretations of environments that may contain anomalous properties. This technology will support decision-making in uncertain conditions and resilient autonomous maneuvers where a Soldier and robot teammate complete exploratory navigation tasks in unknown or dangerous environments under network-constrained circumstances (e.g., search and rescue following a natural disaster). Automatically generated natural language explanations will facilitate the information-overload encountered when sifting through an abundance of low-quality or duplicated visual data by quickly drawing attention to atypical scenarios.

We situate the task of anomaly detection in scenarios where a Soldier is unable to traverse an environment due to conditions that may be dangerous for them. Furthermore, it may be infeasible to receive images or a live stream of the environment due to constraints and limitations on available bandwidth. Therefore it becomes the role of a robot teammate to navigate through the space instead, and communicate information to the Soldier through succinct and informative natural language statements or textual reports. Successful deployment of this envisioned anomaly detection technology must be able to

- identify aspects of the environment that contradict an expectation;
- elaborate on why such an aspect is contradictory, and provide the expected state;
- infer at least one plausible possibility for what might have caused the deviation; and
- infer at least one plausible possibility for what might happen as a result of the deviation.

A team of two interns hosted by ARL and recruited through the National Security Innovation Network X-Force Fellowship spent 10 weeks exploring this problem space. We detail our contributions in this report as follows: we designed an anomaly taxonomy drawing upon related work in visual anomaly detection (Sections 2 and 3); we designed two experiments taking place in virtual environments that were manipulated to exhibit anomalous properties based on the taxonomy (Section 4); we collected a small corpus of human speech and human–robot dialogue for an anomaly detection and explanation task (Section 5); and finally, we designed a novel annotation schema and applied it to a subset of our corpus (Section 6).

2. Related Work

Anomaly detection is the process of determining what is out of place or unexpected with respect to otherwise "normal" observations. This determination relies on a history of patterns or prior behaviors categorized as typical, where violations of the trend are considered anomalous. Anomaly detection has been studied in numerous contexts with diverse data types including social network and spatiotemporal data for misinformation identification[1–4] and surveillance imagery for training self-driving cars to avoid obstacles on the road.[5–7] Obtaining high-quality, labeled training data is key for providing a clear distinction between normal versus anomalous data observations. To this end, many data sets have been built that consist of gold standard, ground truth data points.[8–12] We follow the same procedure in our work and create a novel data set of normal and anomalous linguistic observations, as described in Section 5.

Our pursuit of anomalies lies in visual sensory anomaly detection (i.e., recognizing anomalies from image or video data). Jiang et al.[13] present a three-level classification schema contingent upon the target anomaly and the data type: detecting anomalous objects in images, detecting anomalous scenes in images, and detecting anomalous events in videos. At the object-level, the task is to detect defects in objects (e.g., a tear in a carpet, or a rusty screw).[12] This level assumes that the data type is an image of a single object in isolation against a solid background. A common methodology is to reconstruct the defective portion of the image so that the object is once again normal.[14,15] Anomoly detection at the scene-level is commonly situated in a task, such as obstacle avoidance for self-driving cars. Here, the data type is an image in situ, similar to a photograph taken in the wild which may contain people and natural or artificial scenery. One approach for these task-based scenarios is to determine which portions of the road featured in the image are "free-spaces" and which contain obstacles.[5] The final classification is anomaly detection at the event-level, where anomalies may be unexpected movement over the course of the video. The data type for this level is a video clip, such as spatiotemporal imagery streamed from satellite data as a time-lapse,[16] or videos of people walking or bicycling down crowded streets.[8–11] A successful methodology for event-level anomaly detection attempts to predict the next video frame and determines if a mismatch is present.[17]

Unlike the prior work that treats each anomaly level as distinct, non-overlapping problems requiring different methodological approaches, we argue that our problem space requires a unification of the three levels to conduct anomaly detection at an all-encompassing environment-level. The anomalies we expect in our work span across all three levels (object, scene, event) and are self-contained within the same data type (real-time video streaming). Our problem also differs in terms of the task. Whereas some works focus on specific tasks like self-driving cars and obstacle detection, our task is open-ended. It is a general anomaly detection for anything out of place that is situated in the observer's own interpretation of what is or is not "normal" in the environment.

Furthermore, our work requires subsequent explanation of the anomalies detected, whereas the prior works already knew how the anomaly occurred (e.g., the object was put into the road), did not concern themselves with it (e.g., it is a binary classification task if a carpet is torn or not), or did not subsequently concern themselves with the after-action (e.g., what might happen next.) Our work is thus a new exploration of visual sensory anomalies of a general environment, which may involve entities, relationships between entities, and predicting movement or activities through a space.

3. Defining a Taxonomy of Anomalies

We defined a taxonomy of anomalies with respect to entities in an environment. Our taxonomy is divided into two major branches (pictured in Fig. 1). The upper branch classifies the *properties of individual objects* (e.g., a large [size], pink [color] mug on its side [orientation]). The lower branch classifies the *relationship between objects* (e.g., a large, pink mug on its side on top of [co-location] an office desk). Furthermore, the state or activity of an object can be inferred (e.g., a sink with the faucet open but no water flowing out after a period of time [odd action], or a refrigerator door left open for a period of time [odd state]). The taxonomy was formed in a bottom-up observation-driven approach after looking at various environments and the properties and compositions of their objects. The taxonomy is data type agnostic, and therefore overlaps with Jiang et al.'s[13] images at the object-level (e.g., a single mug), images at the scene-level (e.g., situating the mug within its surroundings), and videos at the event-level (e.g., water not flowing.)

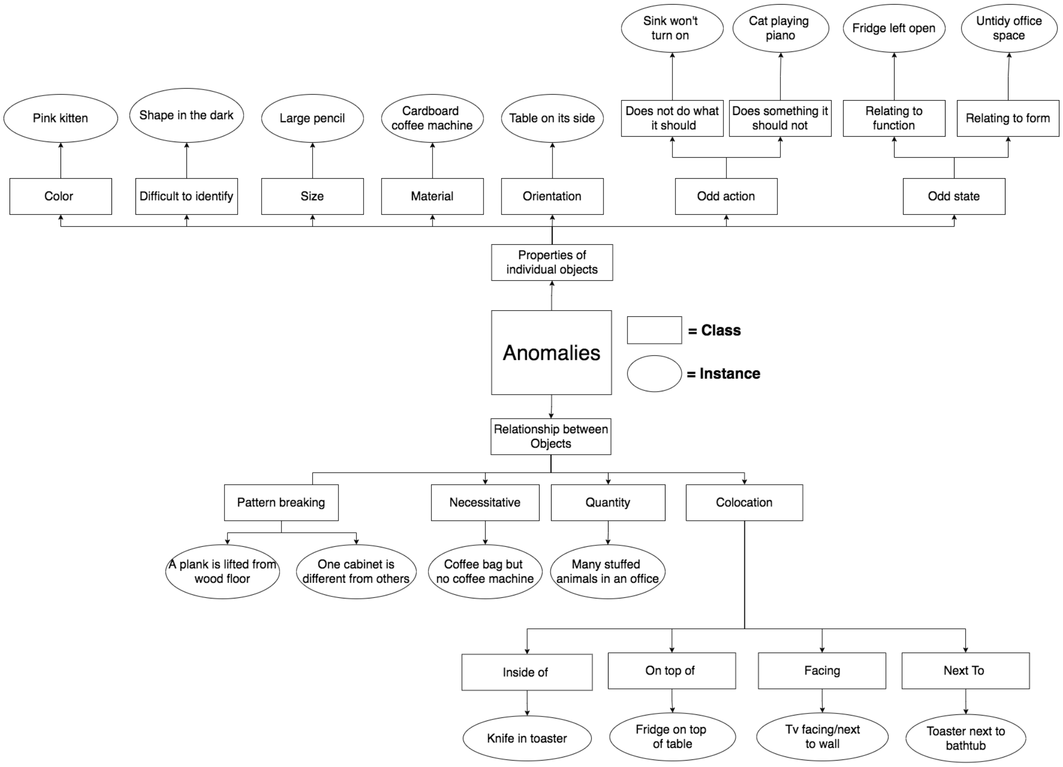


Fig. 1. Anomaly taxonomy of classes with instance examples. The rectangular boxes contain the class type (e.g., the "color" of an entity), and the ovals are an instance or example of the class property (e.g., "*pink* kitten").

Our work utilized the taxonomy in two ways. First, we used the classes to methodologically design environments for testing anomaly detection that vary in type, complexity, and combination within the taxonomy. Second, we posited that the data we collected would exhibit properties from the taxonomy, for example, the spoken question "is the lamp plugged in?" seeks to assess the functional state of an object. Thus, the taxonomy served as a guide in understanding the most salient or informative elements in anomaly detection in environments.

4. Experimental Design and Setup

In order for a robot to support a human teammate in anomaly detection, we envision two operating conditions: 1) the robot observes alone through monologue reporting, and 2) the robot observes together with the human through dialogue reporting. In the first condition, the robot will analyze the environment by itself to detect anomalies, and then generate a complete natural language report of the environment for its human teammate. In the second condition, the robot will be guided through a cooperative dialogue by a human to analyze the environment. The human is not present in the environment with the robot in either condition. Thus in the first condition, the robot's report is the only access the human has the the environment and must be complete, and in the second condition, the robot must understand the human's questions and report back succinctly.

Based on these conditions, we designed two experimental protocols that showcased these varied interaction configurations: a stream of consciousness human monologue data collection experiment (Experiment-Monologue) and a human–robot dialogue experiment (Experiment-Dialogue). The data collected from both experiments would serve as training data for a robotic reasoning and explanation system.

Under the Experiment-Monologue condition, human participants mimicked an autonomous robot, focusing on the real-time thought processes as the participant examined and explained the environment. The participant had full visual and navigation control, yet no prior situation knowledge.[*] They were tasked with verbally describing the environment, stating if anything was anomalous, the reason for that classification, and a possible explanation for the anomaly or group of anomalies.

Under the Experiment-Dialogue condition, human participants fulfilled the role of teammate instructing a remotely located robot to investigate the environment. The participant had no prior knowledge about the task, whether situational or visual, and streaming video and sending images from the robot was not supported. Because an autonomous robot system does not yet exist (i.e., the robot behavior described is the desired outcome of this research), robot behavior in this experiment was carried out using a Wizard-of-Oz methodology (WoZ). WoZ is a well-established experimental protocol used in human–robot research to explore new and desired robotic

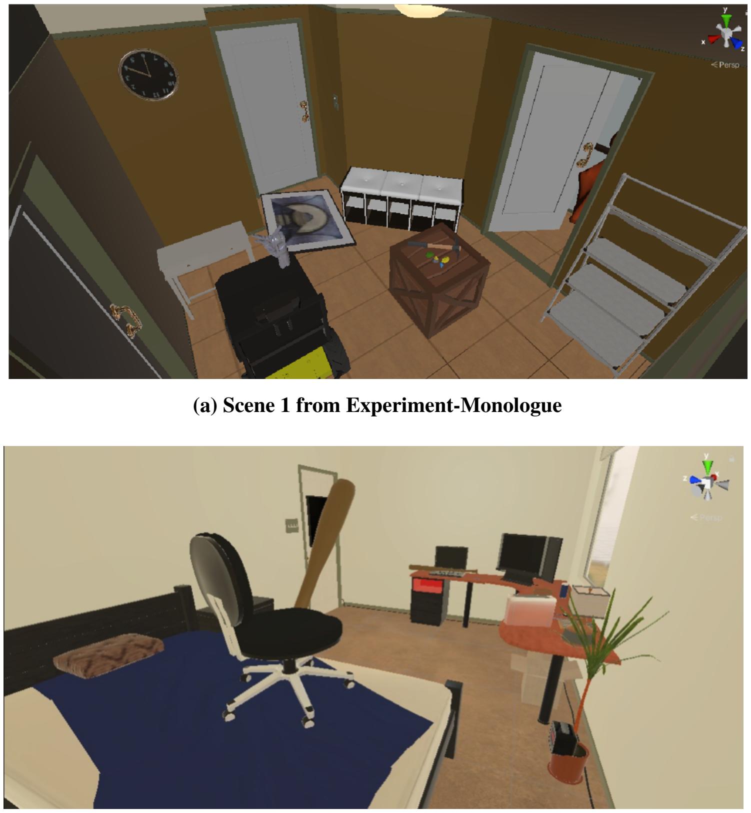
[*] In future experiments, situational knowledge will be a key component of the experimental protocol and thoroughly tested.

behaviors that do not yet exist.[18] Behavior in this experiment included dialogue management and navigation. For navigation, a human experimenter, referred to as a "Wizard," used a joystick to move the robot through the environment. For dialogue management, the same Wizard followed a set of predetermined guidelines and verbally spoke responses to the participant. The data collected from this experiment focused on the conversational aspect of the human–robot interaction, and how a human would naturally ask questions to investigate an unknown and unseen space. The participant verbally instructed the robot to move through the space while asking questions about what the robot saw, and the robot reported its answers until the participant was satisfied that they had completed a comprehensive assessment of the environment.

Anomalous scenes were designed and implemented in Unity using the Robot Interaction in Virtual Reality (RIVR) framework.[19,20] RIVR supports human–robot interaction with a robot in high fidelity environments allowing for a simulated robot to navigate through the virtual environment using the Robot Operating System (ROS). Both Experiment-Monologue and Experiment-Dialogue took place in RIVR in order to have full control over the environments and allow for remote data collection. Environments were modified in RIVR to create anomalous situations according to properties in the anomaly taxonomy. Pre-created scenes from AI2-Thor[21] were selected and then manually and automatically manipulated in RIVR. Additionally, point and click transportation was implemented so that navigation through the virtual environment could take place while the participant in Experiment-Monologue and the WoZ robot in Experiment-Dialogue were seated at their workstations.

Figure 2 shows two manipulated environments. Figure 2a was designed with three target anomalies fulfilling different classes in the taxonomy (noted as follows in parenthesis): an upside-down clock (orientation), mining tools inside the room (co-location), and a painting on the floor (co-location). Figure 2b was designed with five target anomalies: a large bat on door (size, orientation, odd state), a chair on the bed (co-location, odd state), an alarm clock in the plant (co-location, odd state), a bulky phone (size), and a bat on top of a laptop (co-location, odd state). The full list of scenes and anomalies is detailed in the Appendix.

We note a tension that formed between the affordances that a virtual modeling environment like Unity entailed, and the plausibility of the chosen anomalies. For



(a) Scene 1 from Experiment-Monologue

(b) Scene 1 from Experiment-Dialogue

Fig. 2. Screenshots of Unity scenes in RIVR

example, the bulky phone in Fig. 2b was created by distorting the shape of a regular phone through Unity. Therefore, while it fulfilled the size property in the taxonomy, the realism was skewed with respect to real-world environments. We additionally note that while the environments were designed to be anomalous following the taxonomy, with certain assumed or given prior knowledge, anomalies may be "explained away." For example, the large quantity of basketballs on the floor in Fig. 2b may not be anomalous if the room's inhabitant is known to play or coach basketball. As no such prior knowledge was given in the experiments, such impromptu interpretations are of great interest in the collected results.

5. Data Collection

A total of three participants took part in the prepilot experiments: one participant completed five trials (where one trial was one unique scene) for Experiment-Monologue, one participant completed five trials for Experiment-Dialogue, and one participant completed five trials for Experiment-Monologue and five trials for Experiment-Dialogue. Each unique scene was experienced twice by different participants. Table 1 provides a summary of the number of trials and statistics about the environments.

Table 1. Experimental statistics showing environment ("Env."), participant ("Parts."), anomalies, and experimental counts

Parameter	Experiment-Monologue	Experiment-Dialogue
# Unique Env.	5	5
# Unique Parts.	2	2
# Trials	10	10
Avg. # Anomalies / Env.	4	5
Avg. Length of Experiment	5 min	10 min

The data collection resulted in a small corpus of 10 human monologues for Experiment-Monologue and 10 human–robot WoZ dialogues for Experiment-Dialogue. All trials have screen recordings from the perspective of the person wearing the virtual reality headset in Unity (i.e., the participant in Experiment-Monologue and the WoZ robot in Experiment-Dialogue).

6. Data Analysis

From the screen recordings of both Experiment-Monologue and Experiment-Dialogue, we created a pipeline to extract the participant and WoZ robot utterances into a time-aligned Excel spreadsheet. Following that, we conducted two preliminary analyses.

6.1 Preprocessing

We used the "ffmpeg" Ubuntu package to extract the audio stream from the screen recording, and then an off-the-shelf Hugging Face model for Automated Speech Recognition (ASR). The ASR tool did not segment the audio stream into distinct speakers or utterance segments, therefore it required manual segmentation and correction of misrecognized words. Once verified, the corrected transcripts were converted into an Excel file with one utterance per speaker per row.

As of the publication of this report, 2 of 10 transcripts from each Experiment-Monologue and Experiment-Dialogue have been preprocessed and subsequently analyzed.

6.2 Translation to First-Order Logic (FOL)

The participant stream-of-consciousness transcript from Experiment-Monologue was first annotated according to if the utterance was 1) a description of something in the environment, 2) a statement classifying an anomaly, or 3) a possible or plausible explanation for the anomaly, if one exists. Table 2 shows this annotation in the middle column for the transcription excerpt in the left-most column.

Table 2. Experiment-Monologue transcription, annotation, and FOL

Participant utterance	Utterance function	First-order logic
Okay so this looks like a living room	Description	ISA(ENVIRONMENT1, LIVING ROOM)
The first anomaly I see is a large tv remote on this table	Description	ISA(LARGE REMOTE, REMOTE), SIZEOF(LARGE REMOTE), ONTOPOF(LARGE REMOTE, TABLE2)
It is anomalous because its very large	Anomaly classification	SIZEOF(LARGE REMOTE)
Much larger than normal remotes	Anomaly classification	SIZEOF(LARGE REMOTE)
A possible cause could be a decoration or someone is just collecting large remotes	Possible explanation	ISA(LARGE REMOTE, DECORATION), COLLECTS(OCCUPANT, OBJECTSOFTYPE(LARGE REMOTE))
Another anomaly is this very small couch	Object description	ISA(LITTLE COUCH, COUCH), SIZEOF(LITTLE COUCH)
This is anomalous because couches usually aren't that small	Anomaly classification	SIZEOF(LITTLE COUCH)
A possible cause could be that it's a toy or a doll house or a small alien	Possible explanation	ISA(LITTLE COUCH, TOY), ISA(LITTLE COUCH, MODEL), BELONGSTO(LITTLE COUCH, SMALL ALIEN)

This data was further transformed into FOL shown in the right-most column of Table 2. The purpose was to compare a priori expected observations against the observations of the current environment through a representation that encodes information about properties of an object. To narrow in on one example of how the expected or unexpected properties of an object can be harnessed, consider the size of remote controls. Based on prior observations, their typical size has been observed and may be recorded in some kind of existing knowledge graph. This statement can be encoded in FOL as IsA(obj1, Remote), SizeOf(obj1) = 0.3 m^3. However, if a "large, 4 cubic meter" remote control is observed in the current environment, its FOL representation, IsA(obj1, Remote), SizeOf(obj1) = 4m^3, contradicts the prior observations. Depicted graphically, if a knowledge base of some kind is able to capture enough observations of SizeOf for remote controls, a distribution can be plotted—hypothesized in Fig. 3—and if an attribute falls outside the normal distribution, a flag can be set to its anomalous property.

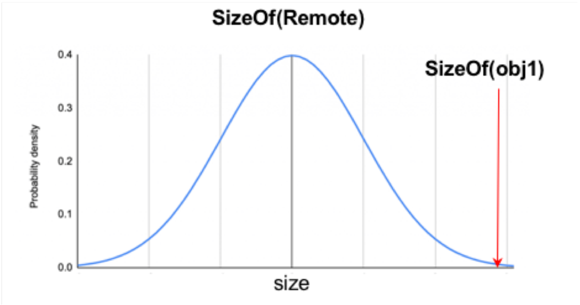


Fig. 3. Sample distribution of remote control size as encoded in a hypothetical knowledge graph

6.3 Coding Dialogue Informativeness

The participant and WoZ-Robot dialogue from Experiment-Dialogue was annotated at two levels. First, participant and WoZ-Robot utterances were grouped together at the Transaction Unit (TU) level. Used in prior dialogue research, a TU represents a group containing the initiation and potential fulfillment of an intent.[22–24] Here, TUs were grouped around the entity under examination, where the intent refers to an assessment for particular anomalous attributes. Future work will examine the selected level of granularity of TU annotation. Table 3 shows the TU annotation, where each TU is visually separated by a double line in the table. TU-1 assesses the type of room the robot is seeing, TU-2 and TU-3 inquire about the bed, and TU-4 inquires about the lamp.

The second annotation conducted was labeling participant utterances according to what property in the taxonomy they were assessing, in other words, what the participant was hoping to learn by asking the question. At the end of TU-3, the participant

Table 3. Experiment-Dialogue transcription and TU annotation

Participant utterance	Robot utterance	TU	Identification
What kind of room are we in now?		1	Room
	A bedroom.	1	
Is there a bed?		2	Co-Location / Inside of
	Yes.	2	
Ok. Is the bed made like neatly?		2	
	Decently.	2	
Is there anything else on the bed? Like other items that have been placed there?		3	Co-Location / On top of
	There's a lamp.	3	
	And there's a sniper rifle leaning on the bed.	3	
Oh ok. So the gun should not be in the bedroom.		3	
It should be locked away. Not a good place to keep that.		3	
The lamp, is it like a desk lamp?		4	Size
	It could be a desk lamp. It's one of the lamps you keep on a dresser or nightstand.	4	
Is it plugged in?		4	Pattern Breaking
	No.	4	
So there's just a lamp in the middle of the bed?		4	Co-Location / On top of
	Yes.	4	
Ok. So that's obviously not where lamps usually go.		4	

was able to make a deduction about the (inappropriate, in their interpretation) presence of the weapon based on the "room" question in TU-1, yet the participant did not learn of its presence until asking about the objects on the bed at the start of TU-3. We also observe that the participant started a line of questioning about the size and state of the lamp in TU-4, until it was abandoned and a more direct co-location line of questioning adopted instead.

6.4 Interpretations

These analyses represent the first step in understanding how anomalies are observed at an environment level using our proposed taxonomy. Our analysis of the Experiment-Monologue data segmented the users' utterances into three types of behaviors. The description utterances serve as salient visual filtering, indicating which part of the environment contains the anomaly, while the classification serves as the evidence for it. Formalizing this into FOL provides one approach for unifying the representation of different types of anomalies within a single computational space.

Our analysis of the Experiment-Dialogue reveals the types of questions within the taxonomy that the user determined were the most important in assessing the environment. The questions probe at properties that are hypothesized to be anomalous and informative in subdividing the search space within the environment. Both the singular object and relationship between objects in the taxonomy were used for understanding, showing the flexibility and coverage of the taxonomy.

7. Conclusion and Ongoing Work

This work presented a shallow yet complete pass through the problem space of determining how a robot and a human–robot team can detect and describe anomalies in environments. The proposed taxonomy is multipurpose in supporting the development of varied anomalous environments, as well as strategies for detecting anomalies and which may be more salient.

Ten environments were curated in Unity to support anomaly exploration and supplemented with the capability for point and click teleportation for virtual reality navigation. Twenty trials were conducted resulting in a small data set of 10 transcripts in the Experiment-Monologue condition and 10 dialogues in the Experiment-Dialogue condition. Two transcripts from each experiment were annotated with novel annotation schemas tying back to the developed taxonomy.

Moving forward, the preprocessing pipeline will be refined and the remaining transcripts created. The annotation schema will be applied to the rest of the corpus. The behaviors observed in annotation will form the basis of training data and serve as a baseline in developing automated systems to follow the process of detecting and describing anomalies. Furthermore, more environments will be designed to test different properties of the taxonomy, and additional data will be collected.

8. References

1. Thom D, Bosch H, Koch S, Wörner M, Ertl T. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. 2012 IEEE Pacific Visualization Symposium; 2012. p. 41–48.

2. Cao N, Shi C, Lin S, Lu J, Lin YR, Lin CY. Targetvue: visual analysis of anomalous user behaviors in online communication systems. Transactions on Visualization and Computer Graphics. 2015;22(1):280–289.

3. Zhao J, Cao N, Wen Z, Song Y, Lin YR, Collins C. #FluxFlow: visual analysis of anomalous information spreading on social media. Transactions on Visualization and Computer Graphics. 2014;20(12):1773–1782.

4. Anand K, Kumar J, Anand K. Anomaly detection in online social network: a survey. 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT); 2017. p. 456–459.

5. Pinggera P, Ramos S, Gehrig S, Franke U, Rother C, Mester R. Lost and found: detecting small road hazards for self-driving vehicles. 2016 International Conference on Intelligent Robots and Systems (IROS); 2016. p. 1099–1106.

6. Santhosh KK, Dogra DP, Roy PP. Anomaly detection in road traffic using visual surveillance: a survey. ACM Computing Surveys. 2020;53(6):1–26.

7. Wang H, Fan R, Sun Y, Liu M. Dynamic fusion module evolves drivable area and road anomaly detection: a benchmark and algorithms. 2022 IEEE Transactions on Cybernetics. 2022;52(10):10750–10760.

8. Adam A, Rivlin E, Shimshoni I, Reinitz D. Robust real-time unusual event detection using multiple fixed-location monitors. Transactions on Pattern Analysis and Machine Intelligence. 2008;30(3):555–560.

9. Mahadevan V, Li W, Bhalodia V, Vasconcelos N. Anomaly detection in crowded scenes. 2010 IEEE Computer Vision and Pattern Recognition (CVPR); 2010. p. 1975–1981.

10. Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab. 2013 International Conference on Computer Vision (ICCV); 2013. p. 2720–2727.

11. Luo W, Liu W, Gao S. A revisit of sparse coding based anomaly detection in stacked rnn framework. 2017 International Conference on Computer Vision (ICCV); 2017. p. 341–349.

12. Bergmann P, Fauser M, Sattlegger D, Steger C. MVTec AD–a comprehensive real-world dataset for unsupervised anomaly detection. 2019 IEEE Computer Vision and Pattern Recognition (CVPR); 2019. p. 9592–9600.

13. Jiang X, Xie G, Wang J, Liu Y, Wang C, Zheng F, Jin Y. A survey of visual sensory anomaly detection. arXiv preprint arXiv:2202.07006; 2022.

14. Zavrtanik V, Kristan M, Skočaj D. Reconstruction by inpainting for visual anomaly detection. Pattern Recognition. 2021;112:107706.

15. Li J, Xu X, Gao L, Wang Z, Shao J. Cognitive visual anomaly detection with constrained latent representations for industrial inspection robot. Applied Soft Computing. 2020;95:106539.

16. Cao N, Lin C, Zhu Q, Lin YR, Teng X, Wen X. Voila: visual anomaly detection and monitoring with streaming spatiotemporal data. Transactions on Visualization and Computer Graphics. 2017;24(1):23–33.

17. Liu W, Luo W, Lian D, Gao S. Future frame prediction for anomaly detection–a new baseline. 2018 IEEE Computer Vision and Pattern Recognition (CVPR); 2018. p. 6536–6545.

18. Bonial C, Marge M, Foots A, Gervits F, Hayes CJ, Henry C, Hill SG, Leuski A, Lukin SM, Moolchandani P, Pollard KA, Traum D, Voss C. Laying down the yellow brick road: development of a wizard-of-oz interface for collecting human-robot dialogue. arXiv preprint arXiv:1710.06406; 2017.

19. Higgins P, Kebe GY, Darvish K, Engel D, Ferraro F, Matuszek C. Towards making virtual human-robot interaction a reality. 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions; 2021.

20. Higgins P, Barron R, Matuszek C. Head pose for object deixis in VR-based human-robot interaction. 31st IEEE International Conference on Robot and Human Interactive Communication; 2022.

21. Kolve E, Mottaghi R, Han W, VanderBilt E, Weihs L, Herrasti A, Gordon D, Zhu Y, Gupta A, Farhadi A. AI2-THOR: An interactive 3D environment for visual AI. arXiv preprint arXiv:1712.05474; 2017.

22. Sinclair JM, Coulthard M. Towards an analysis of discourse: the English used by teachers and pupils. Oxford University Press; 1975.

23. Grosz BJ, Sidner CL. Attention, intentions, and the structure of discourse. Computational Linguistics; 1986.

24. Bonial C, Abrams M, Baker AL, Hudson T, Lukin SM, Traum D, Voss CR. Context is key: annotating situated dialogue relations in multi-floor dialogue. Workshop on the Semantics and Pragmatics of Dialogue; 2021.

Appendix. Log of Scenes and Anomalies

Anomalies in Scene 1 from Experiment-Monologue, pictured in Fig. A-1:

1. Upside down clock (orientation)

2. Mining tools (co-location, inside of room)

3. Painting on floor (co-location, on top of)

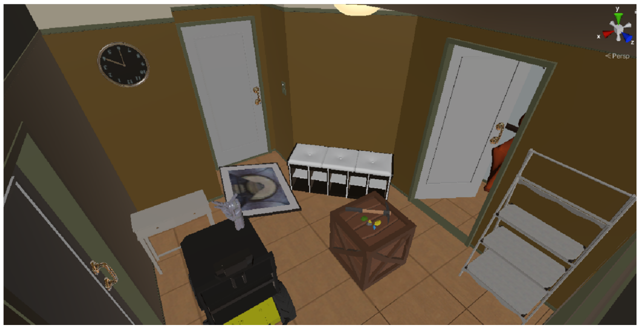


Fig. A-1. Scene 1 from Experiment-Monologue

Anomalies in Scene 2 from Experiment-Monologue, pictured in Fig. A-2:

1. Giant lotion bottle (size)

2. Toaster in bathtub (co-location, inside of)

3. Garbage can on top of toilet (co-location, on top of)

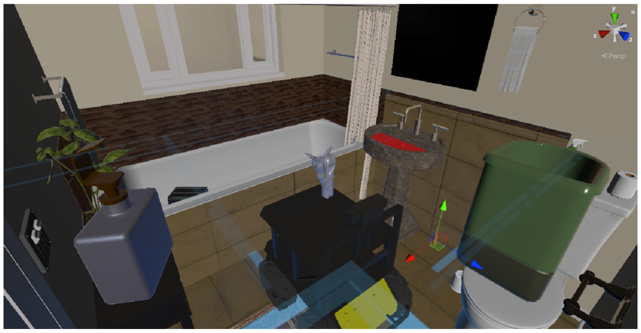


Fig. A-2. Scene 2 from Experiment-Monologue

Anomalies in Scene 3 from Experiment-Monologue, pictured in Fig. A-3:

1. Giant coffee mug (size)
2. Full roll of toilet paper in the trash (co-location, inside of)
3. Table on its side (orientation)
4. Chemistry equipment (co-location, inside of)
5. Coffee machine on top of stove (co-location, on top of)

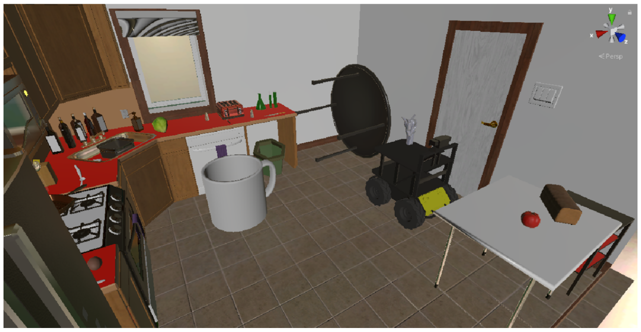


Fig. A-3. Scene 3 from Experiment-Monologue

Anomalies in Scene 4 from Experiment-Monologue, pictured in Fig. A-4:

1. Toilet in living room (co-location, inside of)
2. Gun inside of living room (co-location, inside of)
3. Giant remote (size)
4. Tiny couch (size)

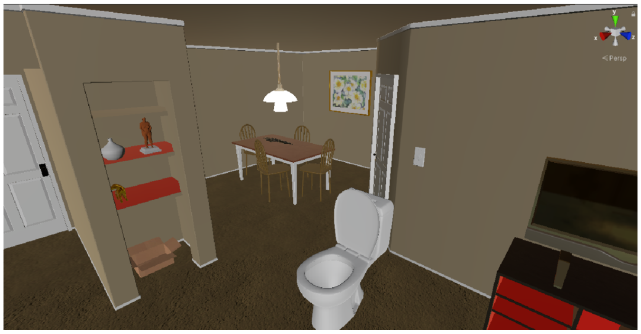


Fig. A-4. Scene 4 from Experiment-Monologue

Anomalies in Scene 5 from Experiment-Monologue, pictured in Fig. A-5:

1. Giant watch (size)
2. Gun inside of living room (co-location, inside of)
3. Table upside down (orientation)
4. Couch on its side (orientation)

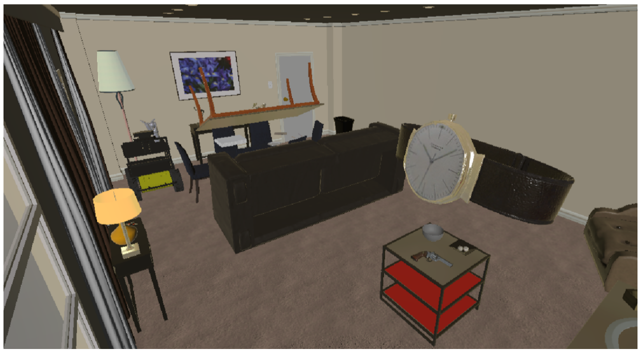


Fig. A-5. Scene 5 from Experiment-Monologue

Anomalies in Scene 1 from Experiment-Dialogue, pictured in Fig. A-6:

1. Large bat on door (size, orientation, odd state)
2. Chair on bed (co-location, odd state)
3. Alarm clock in plant (co-location, odd state)
4. Very thick phone (size)

5. Bat on laptop (co-location, odd state)

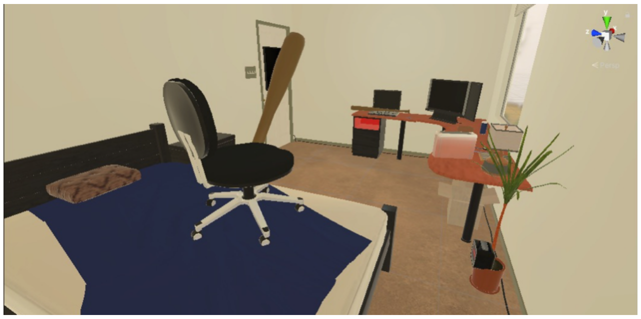


Fig. A-6. Scene 1 from Experiment-Dialogue

Anomalies in Scene 2 from Experiment-Dialogue, pictured in Fig. A-7:

1. Bat in large mug on bed (co-location, size)
2. A lot of basketballs (quantity)
3. Knocked over trash can in front of door (material, color, orientation)

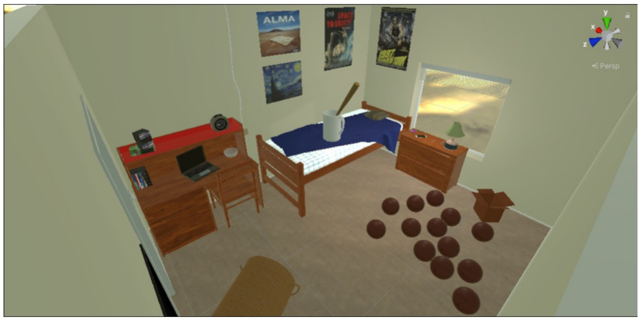


Fig. A-7. Scene 2 from Experiment-Dialogue

Anomalies in Scene 3 from Experiment-Dialogue, pictured in Fig. A-8:

1. Laptop and phone on sofa on couch (co-location)
2. Knocked over statue (orientation, odd state)
3. Large remote (size)
4. Knocked over plant (orientation, odd state)

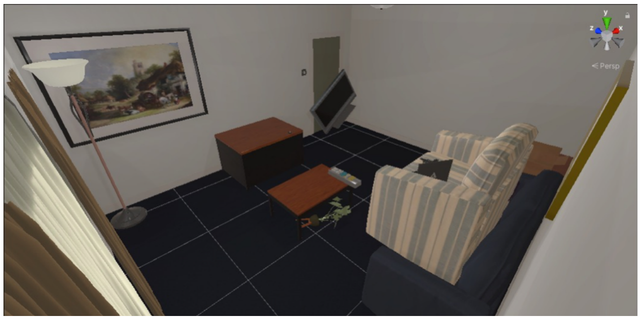


Fig. A-8. Scene 3 from Experiment-Dialogue

5. TV in front of door (co-location, odd state, orientation)

Anomalies in Scene 4 from Experiment-Dialogue, pictured in Fig. A-9:

1. Many plants on table (quantity)
2. Very tall box (size, difficult to identify)
3. Painting in front of couch (location, odd state)
4. Lamp in front of couch (location, orientation, odd state)
5. Two TV's holding a third TV (quantity, orientation, odd state)

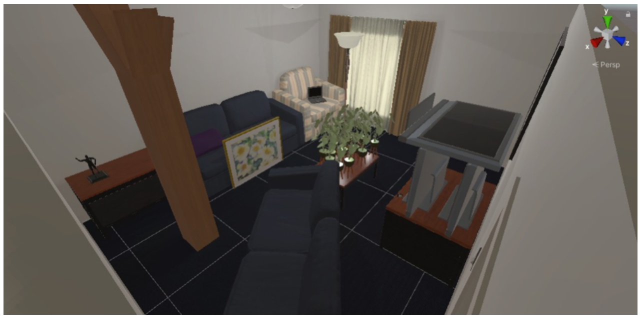


Fig. A-9. Scene 4 from Experiment-Dialogue

Anomalies in Scene 5 from Experiment-Dialogue, pictured in Fig. A-10:

1. Bat on door handle (co-location)
2. Laptop on chair knocked over (co-location, orientation, odd state)
3. Lamp on crooked bed (co-location, odd state)
4. Sniper rifle on bed (co-location, odd state)
5. Pillow on box (co-location)
6. Drawers left open (odd state)
7. Large bowl on floor (size, odd state)
8. Pencil and pen on floor (odd state)
9. Phone on trash can (co-location, material, color)

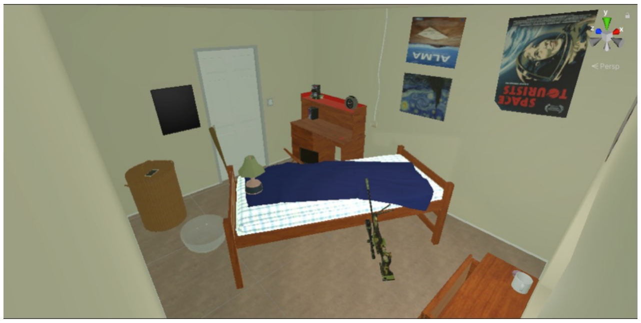


Fig. A-10. Scene 5 from Experiment-Dialogue

List of Symbols, Abbreviations, and Acronyms

AI2-Thor – A near photo-realistic interactable framework for embodied AI agents

ARL – Army Research Laboratory

ASR – Automated Speech Recognition

DEVCOM – US Army Combat Capabilities Development Command

ffmpeg – A framework for streaming, encoding, and decoding multimedia files

FOL – first-order logic

RIVR – Robot Interaction in Virtual Reality

ROS – Robot Operating System

TU – Transaction Unit

WoZ – Wizard of Oz

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 DEVCOM ARL
(PDF) FCDD-RLB-CI
TECH LIB

1 DEVCOM ARL
(PDF) FCDD-RLA-IC
S M LUKIN

1 FLORIDA ATLANTIC UNIVERSITY
(PDF) M BELLISSIMO

1 UNIVERSITY MARYLAND
(PDF) R SHARMA