Towards an Ontology of Human Explanations of Robotic Behavior

Maria Rausa[1,*,†], Agnese Augello[1,†] and Antonio Lieto[1,2,†]

[1]Institute for High Performance Computing and Networking of the Italian National Research Council (ICAR-CNR), Via Ugo La Malfa, 153, 90145, Palermo, Italy
[2]Cognition Interaction and Intelligent Technologies Laboratory (CIIT Lab), DISPC, University of Salerno, Via Giovanni Paolo II, 132, 84084, Fisciano (SA), Italy

Abstract

Explainability in AI is essential for fostering trust and effective interaction, particularly in socially sensitive contexts. A conceptual formalization and an empirical study of explanations in Human-Robot Interaction (HRI) scenarios can be informative in the design and implementation of explainable robots. However, the current state of research lacks this type of analysis, especially at a fine-grained level. The HERB (Human Explanation of Robotic Behavior) project aims to collect and analyze explanations of robotic behaviors in social and educational settings. In this work, we present an ontology of explanations designed to analyze semi-structured interviews conducted within the HERB project in order to gather individuals' perspectives on various social scenarios of HRI. This ontology, relying on a taxonomy of explanation types derived from philosophical theories, categorizes explanations into distinct types and incorporates a lexicon of linguistic markers.

Keywords

Explanations, Ontology, Human Computer Interaction, Social Robotics, Explainable AI

1. Introduction

Explainability in Artificial Intelligence (AI) is a critical area of research, driven by the necessity for AI systems to align with ethical principles and user expectations. The ability to provide comprehensible and contextually relevant explanations is essential for fostering trust and enabling effective interactions between users and AI systems. At the basis of the design and implementation of explainable AI systems, it is important to examine the concept of explanation at a fine-grained level, considering both philosophical perspectives and empirical results. This involves analyzing how AI systems and social robots are interpreted and explained by individuals, taking into account their backgrounds, competencies, and experiences.

In the context of the HERB (Human Explanation of Robotic Behavior) project, we are addressing this challenge by collecting and analyzing explanations of robotic behaviors in social and educational settings. Through this process, we identified the need for a systematic procedure to gather and analyze explanations, particularly in providing a structured approach to support and streamline this process. Based on these considerations, this work introduces a formalized ontology of explanations built upon a taxonomy of explanation types derived from philosophical theories. The ontology categorizes explanations into distinct types, such as mechanistic, causal, teleological, and functional, offering a framework that primarily aims at supporting the analysis of explanations provided by individuals during Human-Robot Interaction (HRI) within the HERB project. In addition to this primary goal, we believe that this formalization can also be exploited by AI systems to generate explanations tailored to specific contexts and user's needs. The ontology incorporates a lexicon of linguistic markers that

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy
*Corresponding author.
†These authors contributed equally.
maria.rausa@icar.cnr.it (M. Rausa); agnese.augello@icar.cnr.it (A. Augello); alieto@unisa.it (A. Lieto)
https://docenti.unisa.it/024406/en/home (A. Lieto)
0000-0002-0877-7063 (M. Rausa); 0000-0001-6463-9151 (A. Augello); 0000-0002-8323-8764 (A. Lieto)

is used to guide, using a keyword-based approach, the categorization of human explanations in the defined explanation types.

The article is structured as follows: We begin by exploring the motivations behind this work (Section 2) and the theoretical foundations of explanations, as well as the taxonomy underlying the proposed ontology (Section3). We then describe the implementation process, detailing the use of semantic web tools for ontology modelling (Section 4). Subsequently, we present preliminary examples of the ontology's application (Section 5) by discussing its current limitations (Section 6) and proposing directions for future research, particularly in advancing the semantic analysis of textual explanations and refining the integration of linguistic and conceptual markers (Section 7).

2. Rationale of the work

Research on explanations has gained significant relevance in this period, characterized by extensive discussions on AI and the necessity to adhere to the principles and guidelines established by regulatory frameworks such as the AI Act [1]. Consequently, explainability has become a central focus in AI research, with numerous models and frameworks proposed to make AI systems more transparent and interpretable [2][3][4]. However, it is important to note that despite the growing emphasis on explainability, there is a need for comprehensive analysis regarding how explanations are generated and interpreted. In our opinion, this analysis, appropriately drawing on insights from philosophical theories and supported by empirical data, can provide a foundational perspective to enhance the design of such systems.

The study of explanations is the primary objective of the HERB (Human Explanation of Robotic Behavior) project, within which the proposed work is framed and implemented. In particular, the HERB project investigates human explanations of robotic behaviour in educational and social contexts. The aim is to analyze how individuals interpret and explain robot actions, considering factors such as the type of robot—whether social or non-social, humanoid or non-humanoid—and the level of manipulability. The project examines the relationship between the behavior to be explained (explanandum) and the theoretical assumptions that provide the basis for these explanations (explanans), identifying different types of explanations.

Studying and formalizing explanations is crucial not only for improving interactions with AI systems but also for addressing ethical concerns and contributing to the design of AI systems and social robots that are more comprehensible and trustworthy. In this work, we introduce an explicit formalization of explanations within an ontology, building upon a taxonomy of explanations developed in the context of the HERB project. Our perspective is that establishing an ontological formalization of explanations could support two key processes: first, it may facilitate the analysis of explanations provided by individuals in the interviews conducted within the HERB project to identify the type of explanation, and second, it could enable AI systems to formulate explanations that correspond to a specific explanandum and align with a designated explanation type. The following subsections delve into important considerations regarding the potential applications of this ontology.

2.1. Role of Expectations

When formulating an explanation, it is crucial to consider the receiver of the explanation and the social context in which it is provided. As highlighted in the literature, for instance by Hoffman et al.[5], explanations in AI systems are not intrinsic properties; rather, they emerge from the contextual interplay between users and systems. The effectiveness of these explanations hinges on users' prior knowledge and objectives.

The social context shapes expectations, and social expectations significantly influence perceptions of the system's reliability and its ability to act consistently. In this case, expectations are also influenced by personal experiences with technology. Unrealistic expectations often arise, especially within Human-Robot Interaction (HRI), leading to misconceptions about robots' cognitive, emotional, and social

capacities. Vulnerable individuals, such as children, the elderly, or those with mental disabilities, may develop false beliefs about robots possessing human-like thoughts and emotions.

The research conducted by Riveiro et al. [6] underscores the importance of user expectations. Their findings reveal that the nature of the explanation sought by users is heavily shaped by the alignment of the system's output with their expectations.

The proposed ontology could facilitate the creation of explainable AI that tailors and contextualizes explanations to meet users' specific needs. In this way, it would enable the agent to provide explanations that align with the expectations of the user it is interacting with. Imagine if we could relate types of explanations to possible categories of people based on factors such as digital literacy, attitudes, tendencies to anthropomorphize, or the evolution of an interaction. Would an agent that customizes its explanations based on specific interactions or individual users be more explainable? Consequently, could this foster greater trust from the interlocutor and help the agent achieve its goals more effectively, whether through persuasive strategies in assistance scenarios or educational strategies?

2.2. Explanations and Trust

Regarding the previously mentioned influence on trust, it is crucial to delve deeper into how explanations can foster appropriate trust that accurately reflects the capabilities and limitations of AI systems [7]. This involves not only providing clear and accessible information about the AI's functionality but also addressing the potential misconceptions users may have. In the context of robotics, the concept of "epistemic responsibility", as highlighted by Sullins[8], underscores the necessity for users to seek reliable information to assess whether a robotic system is deserving of their trust.

Previous studies suggest that while providing explanations can enhance user trust, this trust can sometimes be misplaced, leading to an over-reliance on systems that do not consistently deliver reliable results, particularly in sensitive areas like healthcare.

The proposed ontology could facilitate the creation of explainable AI systems that recalibrate individual' expectations, by clarifying the effective capabilities and limitations of that system. For example, if a home assistance robot communicates its operational processes and constraints, users can adjust their expectations accordingly.

2.3. Supporting Critical Reasoning in AI Systems

The proposed formalization can be exploited by AI agents to engage in critical reflection on their behavior, making it more comprehensible not only to the users they interact with but also to themselves. This would represent an innovative approach to explainable AI that, in the first case, would allow the agent to analyze its internal processes, focusing on the aspects that led to achieving specific outcomes and improving them through feedback mechanisms. For instance, a well-founded formalization of counterfactual explanations could lead the agent to evaluate various possibilities and enhance the effectiveness of its actions [9].

3. Types of Explanations

In order to provide a first formalization of the types of explanations associated to a certain behavior, it is necessary to introduce the considered vocabulary. First of all, intuitively, the notion of "explanation" is strictly linked to the one of prediction. A good "explanation", should be able to provide predictive models of a certain phenomenon. However, different types of theories have been proposed to define what is a correct "explanation" from a scientific view point (for the details in the context of AI and Cognitive Modelling we remind to [10]). Here we briefly recall some of them that have been of interest in the context of our study. The first type of theory about explanation is the so called Deductive-Nomological (DN) Explanation. According to this view, introduced by Hempel and Oppenheim [11], there are some strict characteristics that an explanans have to satisfy in order to explain a given phenomenon. In particular, the explanandum is seen as something that needs to be logically derived, via deduction, from

the explanans. While, intuitively, this theory adequately address a normative notion of explanation, (since it assumes that the explanans provides the causes, i.e. the necessary and sufficient conditions, to understand the explanandum) such requirement is very strict since there are many good explanations, also in the scientific fields, where the elements having an explanatory role (the explanans) is not able to completely "derive" a deductive account of the explanandum phenomenon (i.e. all the empirical laws do not work in this way). Another type of explanation is the so called "functional" one where explaining consists in providing "a function that a system is believed to possess" [12]. In other words: functional explanations explain the capacities of a system in terms of its sub-components and capacities (e.g. one can explain that a computer is able to produce a certain output since it is made by a certain hardware or software architecture where each component plays a certain function contributing to the final output). To a certain extent, this explanation is given by the how a certain system of model is build, not by the computations performed by itself. Other explanatory theories developed in the literature concerns the so called "teleological", "evolutionistic" and "mechanistic" explanations. We briefly describe them by using a running example coming from the biological domain. Let us suppose that we aim at explaining the phenomenon according to which chameleons change their skin color. This usually happens in presence of a predator (they assume different color configurations based on the different predators they perceive) or potential mating partners. Now, if we are interested in an explanation about why chameleons assume more often the color configuration associated to a particular predator (e.g. birds etc.) a possible answer could be that "the number of bird predators in major with respect to other animals and thus this has determined a stronger selective pressure". This is a typical example of evolutionistic explanation, a type of explanation that plays an important role in many scientific theories. If we suppose, however, that the focus of our interest is just to understand why chameleons, in general, change their color skin we could have other types of explanation. For example: a teleological explanation (from the greek "telos": scope). This type of explanation assumes that, in order to explain a phenomenon F one has to point out which is the ultimate scope that F allows one to achieve. In the example, if someone tells us that "chameleons change their skin color to mimetize themselves and escape from predators" she simply provides an explanation about the scope of the phenomenon intended to explain. However, can we say that this kind of explanation helps us to understand "why" chameleons change their color? Of course if we suppose to be interested to the mechanisms determining that phenomenon, we cannot really declare ourself satisfied by that answer. On the other hand, if we receive the following explanation "the skin color change in chameleons is due to the response of some cells contained in the animal pigments (cromatofores) to nervous and endocrinous stimuli" we would probably be satisfied by this answer. In particular, our satisfaction would probably by derived by the fact that this kind of explanation shows the "mechanisms" determining the phenomenon we want to understand. This kind of explanation is called "mechanistic" and represents the kind of explanation able to shade lights on the inner mechanisms determining the behaviour of a given system. In the example provided, the very simple mechanistic explanation was also a causal explanation.

This different types of explanations (and their specializations) have been the ones in focus during our study and formalized in our preliminary ontology.

4. The HERB Ontology

The developed ontology provides a preliminary formalization of the above introduced different types of explanations, with a particular focus on distinctions such as nomological-deductive, mechanistic, causal, functional, evolutionistic, teleological (and their subclasses that will be introduced below). The ontology (Figure 1) has been implemented in OWL using the Protégé software [1], integrating SWRL rules [2] to enhance semantic inference and explicitly define the concepts, relationships, and governing rules behind these categorizations.

[1] https://protege.stanford.edu/software.php
[2] https://www.w3.org/submissions/SWRL/

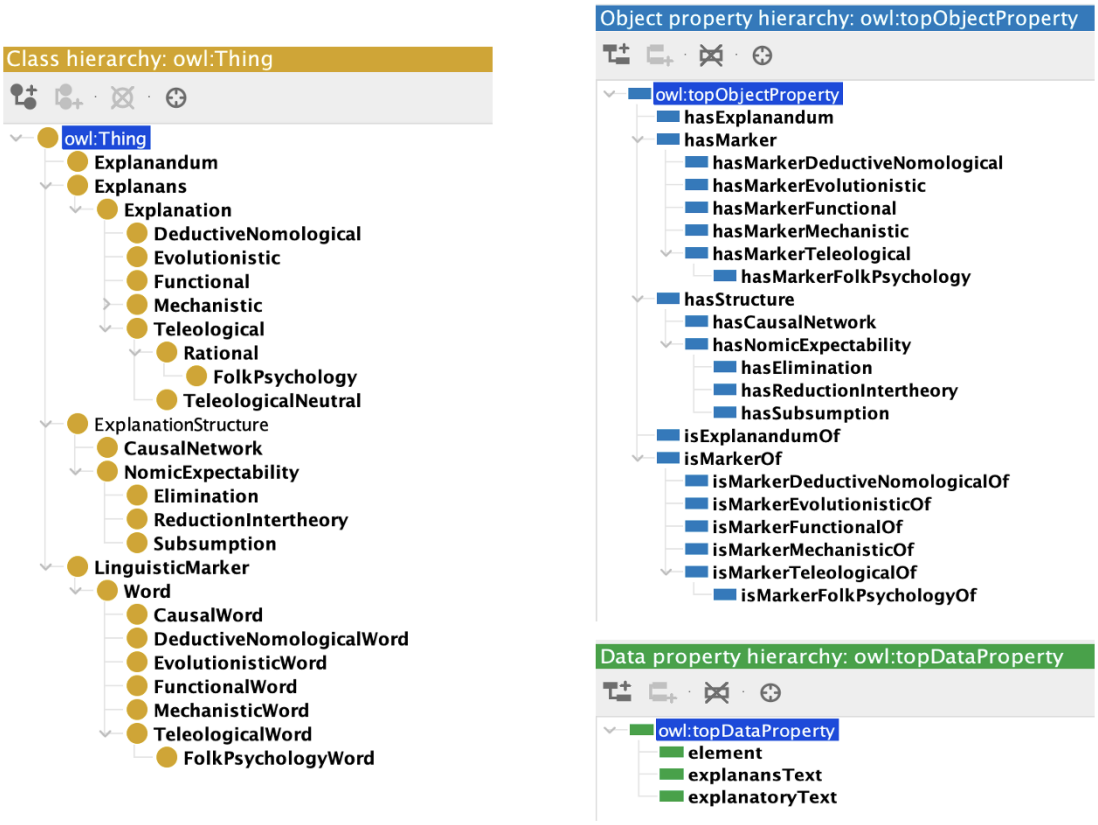

Figure 1: Classes, Object Properties and Data Properties of the HERB Ontology.

4.1. Classes, Object Properties and Data Properties

The core Classes include *Explanandum*, which represents the phenomenon or behavior that requires explanation, and *Explanans*, which captures the general concept of explanation regardless of its specific type, representing the statements or concepts used to elucidate a phenomenon. The *Explanation* class categorizes specific types of explanans into subclasses, including *DeductiveNomological, Mechanistic* (and its subclass *Causal*), *Evolutionistic, Functional, Teleological* (and its subclass *Neutral, Rational* and *FolkPsychology*).

In particular, as indicated before, *DeductiveNomological* explanations based on general laws or principles, explaining phenomena by logically deriving them from an explanans; *Mechanistic* explanations focus on the processes and functionalities of complex systems, explaining phenomena through their subcomponents and interactions; *Causal* explanations, a subclass of Mechanistic explanation, concentrate on cause-effect relationships between the components of a system; *Evolutionistic* explanations analyze phenomena in term of change and adaptation over time; *Functional* explanations that highlight a phenomenon's function within a broader system; *Teleological* explanations are goal-oriented and they can be further divided into *Neutral*, which refers on general goals, and *Rational*, that explain behavior in terms of goals, beliefs, and rationality. In turn, Rational Teleological explanations have a subclass, *FolkPsychology* explanation, which employ concepts from folk psychology.

The ontology incorporates linguistic markers, represented by the *LinguisticMarker* class, which identifies significant linguistic elements associated with different types of explanations. These markers are further specialized in the *Word* subclass, capturing terms that are characteristic of specific explanatory styles. For instance, *DeductiveNomologicalWord* includes terms like "law" or for explanations grounded in law or general principles, while *MechanisticWord* encompasses terms like "mechanism" or "structure", relevant to explanations referring to processes or systems. Similarly, *CausalWord* contains terms like "cause" or "determine" *EvolutionisticWord* includes phrases such as "evolved for" or "selected for", and *FunctionalWord* captures terms like "function as" or "role." For teleological explanations,

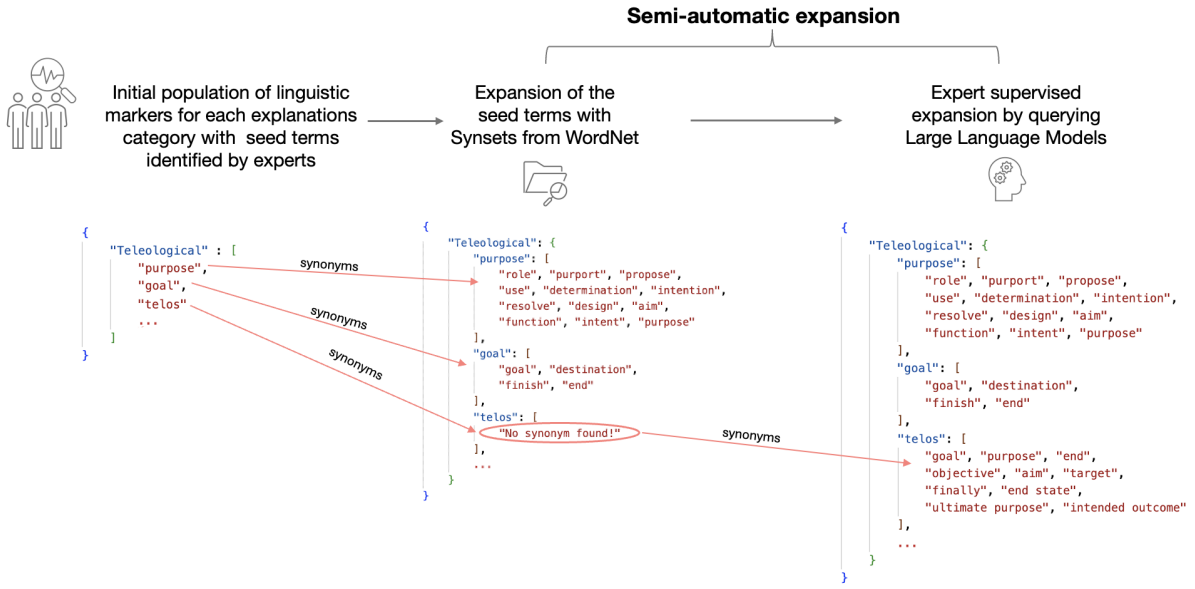

Figure 2: Creation of lexicon through a semi-automatic process: First Expansion with the Wordnet Synset and Second Expansion with LLMs.

TeleologicalWord represents goal-oriented terms like "purpose" or "objective", while *FolkPsychologyWord* (subclass of Teological Words and markers) encapsulates vocabulary tied to Folk Psychology, such as "intention" or "desire." All the above mentioned linguistic markers are typically associated to (and adopted within) the different types of explanations investigated in this work. In our work they are essential for identifying and categorizing explanation types in natural language processing contexts through SWRL rules.

The relationships between classes and instances in the ontology are captured through Object Properties. For example, *hasExplanandum* links an explanation to the phenomenon it seeks to explain, with the inverse property *isExplanandumOf*. The *hasMarker* property associates an explanation with its linguistic markers, and its sub-properties (*hasMarkerDeductiveNomological, hasMarkerMechanistic, hasMarkerCausal, hasMarkerEvolutionistic, hasMarkerFunctional, hasMarkerTeleological*, and *hasMarkerFolkPsychology*) specify markers for particular explanatory types, ensuring precision in categorization. Additionally, *hasStructure* connects an explanation to its structural framework, with sub-properties like *hasNomicExpectability* (further detailed with *hasSubsumption, hasReductionIntertheory*, and *hasElimination*) and *hasCausalNetwork*, which describe relationships relevant to nomological-deductive and mechanistic explanations respectively.

The ontology also leverages Data Properties to describe intrinsic attributes of its entities. In fact, the *element* property links instances of the *Word* class to their representative textual strings, enabling precise annotation of linguistic elements. Meanwhile, *explanansText* and *explanatoryText* provide natural language descriptions for instances of the *Explanans* and *Explanation* classes, respectively.

4.2. Lexicon of Linguistic Markers

The ontological model developed for the HERB project incorporates a carefully structured lexicon of linguistic markers, created through an hybrid process combining manual and semi-automatic semantic expansion techniques (Figure 2).

The initial identification of markers was carried out by the authors in collaboration with a team of philosophers of science, employing a seed-based approach. This method, widely adopted in the construction of lexical resources [13, 14], involves the manual selection of a set of keywords, *seed terms*, that represent the semantic core of various explanatory categories. The seeds were selected for their theoretical relevance and recurrence in descriptive contexts of explanatory types.

Figure 3: a) Individuals of the *WordList*. b) The most relevant words represented with a Word Cloud Generator (https://www.freewordcloudgenerator.com).

Once the seeds for each explanatory category were defined, they were integrated into the HERB ontology, where linguistic markers are represented as *Individuals* named *DeductiveNomologicalWordList, MechanisticWordList, CausalWordList, EvolutionisticWordList, FunctionalWordList, TeleologicalWordList* and *FolkPsychologyWordList* (Figure 3a), which belong respectively to the classes *DeductiveNomologicalWord, MechanisticWord, CausalWord, EvolutionisticWord, FunctionalWord, TeleologicalWord* and *FolkPsychologyWord*. Each *WordList*, using the ObjectProperty *element*, contains the seeds that reflect the different categories of explanations.

Below are some of the most relevant seeds for each *WordList* (Figure 3b):

- *DeductiveNomologicalWordList*: “law”, “principle”, “rule”, “consequence”, “condition”, “deduce”
- *MechanisticWordList*: “mechanism”, “process”, “gears”, “component”, “engine”, “system”
- *CausalWordList*: “cause”, “effect”, “provoke”, “trigger”, “generate”, “lead to”, “consequence”
- *EvolutionisticWordList*: “adaptation”, “evolution”, “advantage”, “selection”, “development”
- *FunctionalWordList*: “function”, “role”, “utility”, “advantage”, “purpose”, “system”, “maintain”
- *TeleologicalWordList*: “purpose”, “end”, “objective”, “reason”, “goal”, “design”
- *FolkPsychologyWordList*: “intention”, “desire”, “will”, “belief”, “emotion”, “plan”, “behavior”

To ensure broader semantic coverage and improve the effectiveness of the lexicon, a two-phase semi-automatic expansion process was implemented. In the first phase, to organize words into synonym sets (synsets) and maps their semantic relationships, we used WordNet [15]. With a Python script, the seeds were processed to identify synonyms and related terms, thereby expanding the vocabulary associated with each category. However, in some cases, WordNet failed to provide results for some terms. We therefore used a second manual phase of supervised expansion leveraged large language models. These models enabled the automatic generation of terms semantically related to the seeds, enriching the lexicon in cases where WordNet returned *“No synonym found!”*.

The enriched lexicon is being verified by experts and subsequently integrated into the HERB ontology with the DataProperty *element*.

The *WordLists* were then used to support a system of rules based on the Semantic Web Rule Language (SWRL) that allow for the automatic inference of the explanatory type associated with an *Explanans* entity by analyzing the textual content of the *DataProperty explanansText*.

The rules compare the provided text with the seeds in the *WordLists* and, in case of a match, automatically assign the entity to the corresponding explanatory class.

For instance, Figure 4, if the *explanansText* of *EngineExplanans* is “The cooling unit works through a heat exchange mechanism” and the *MechanisticWordList* contains the term “mechanism”, an SWRL rule recognizes the semantic match and classifies the entity as belonging to the *Mechanistic* category.

This process relies on a built-in function (*swrlb:contains*) that verifies the presence of the term in the text. This approach enables automated categorization based on linguistic criteria, harnessing the ontology’s inferential capabilities to classify explanations efficiently and accurately.

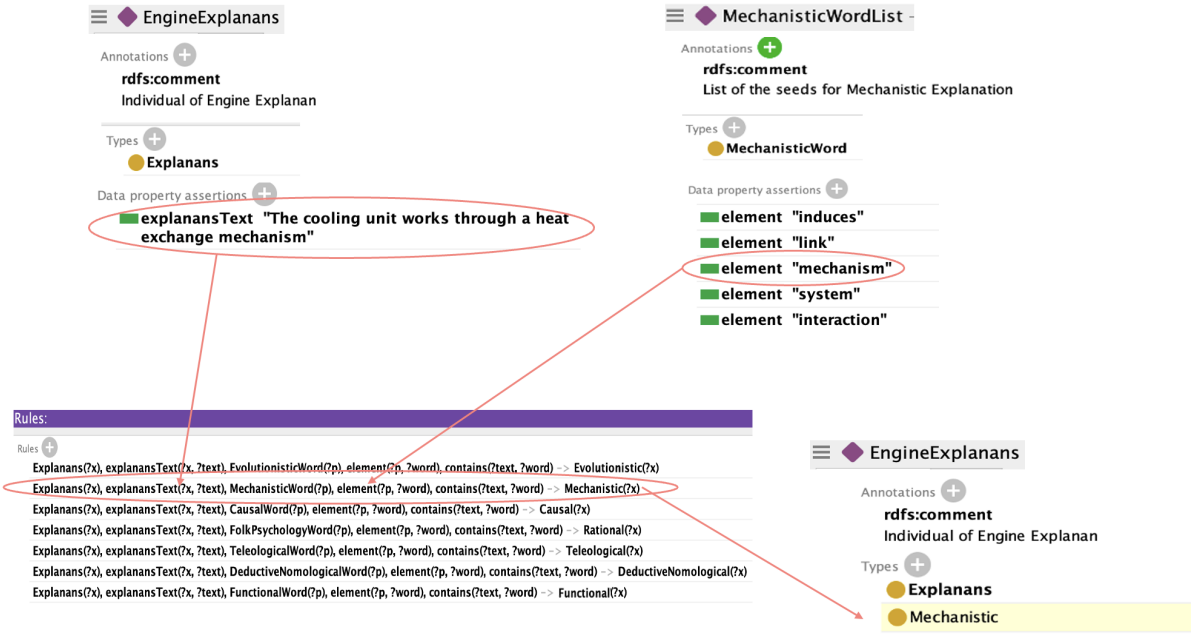

Figure 4: Example of the application of the SWRL Rules

5. Examples

In this section, we present examples for each category of explanation, illustrated through the robot's behaviour in the following scenario: "*A robot walks toward a door, stops upon reaching it, and then changes direction*".

We begin with the *Deductive-Nomological Explanation*, which relies on laws or general principles. An example is the explanans with the explanatory text: *"The robot avoided hitting the door because, according to the principle of motion dynamics, its sensors detected the door within its safety radius, and the control system applied the rule to stop and change direction"*. In this case, the presence of the terms *"according to", "principle", and "rule"*, belonging to the *DeductiveNomologicalWord* class, allows the explanation to be inferred as *Deductive-Nomological* (Figure 5a).

Moving on to the *Mechanistic Explanation*, which focuses on internal mechanisms and their functioning, we consider the explanans with the explanatory text: *"The robot's navigation mechanism relies on a system of interconnected components: sensors detect the door, the motor's gears adjust the wheels, and the processing unit calculates a new path to avoid a collision"*. Here, the terms *"mechanism", "system", "components"* and *"process"*, belonging to the *MechanicalWord* class, identify the explanation as *Mechanistic* (Figure 5b).

For the *Causal Explanation*, which interprets behavior in terms of cause-effect relationships, we examine the explanans with the explanatory text: *"The robot avoided hitting the door due to the detection of the obstacle by its sensors, triggering a signal that stopped the motors and recalibrated its direction"*. The presence of the terms *"due to"* and *"trigger"*, belonging to the *CausalWord* class, allows this explanation to be classified as *Causal* (Figure 5c).

In the case of the *Evolutionistic Explanation*, which analyzes phenomena in terms of change and adaptation over time, we consider the following example of the explanans with explanatory text: *"The robot has evolved its door-avoidance capabilities through iterative training and adaptation. Its algorithm was selected for its ability to improve collision prevention in complex environments"*. The terms *"evolved", "selected for"* and *"adaptation"*, associated with the *EvolutionisticWord* class, suggest that this explanation falls into the *Evolutionistic* category (Figure 5d).

The *Functional Explanation* highlights the function of a behavior within the context of a broader system. An example is the explanans with the explanatory text: *"The robot's avoidance function is*

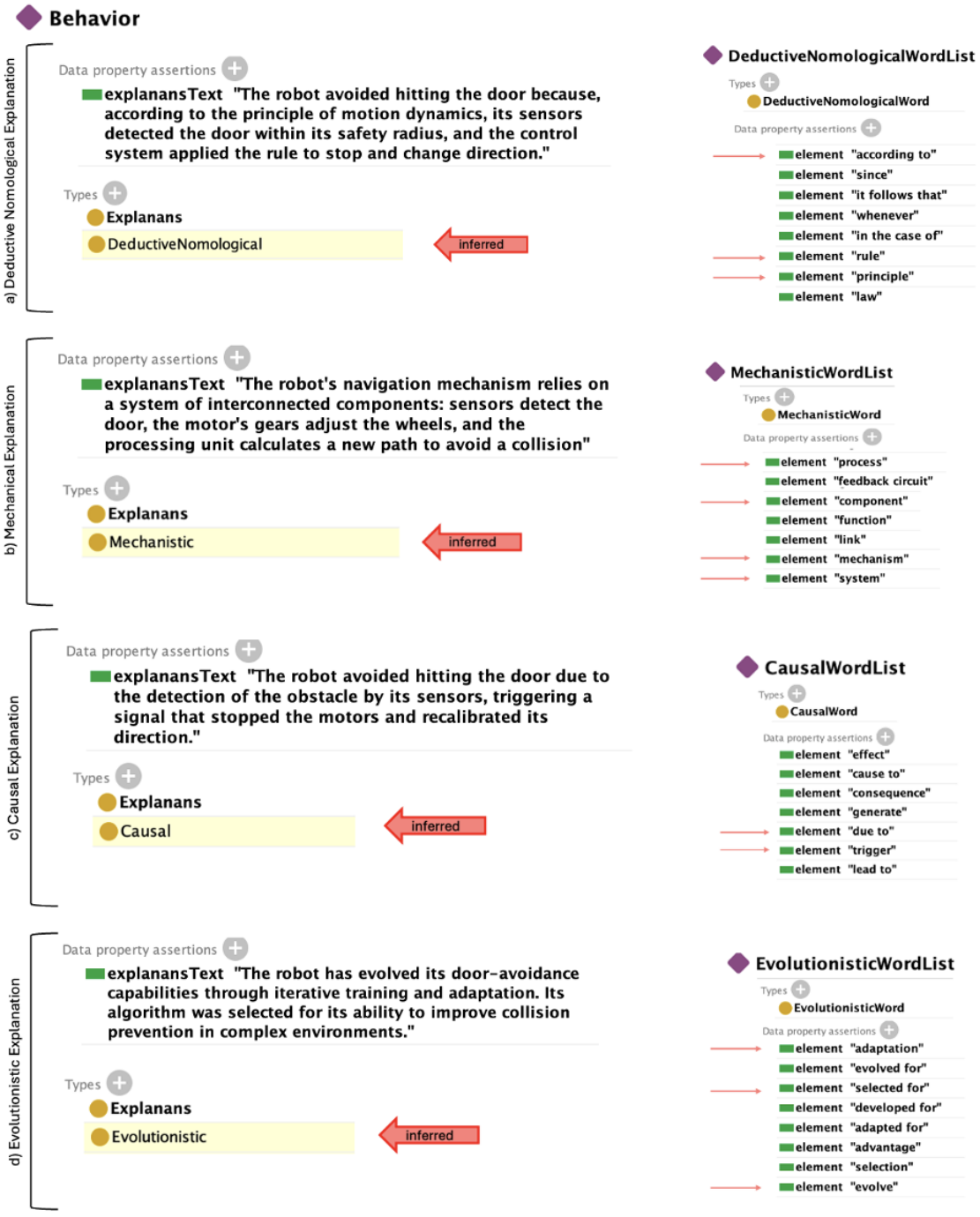

Figure 5: Example of a) Deductive-Nomological Explanation; b) Mechanistic Explanation; c) Causal Explanation; d) Evolutionistic Explanation.

essential for preventing collisions with doors and maintaining smooth navigation, contributing to its broader role of efficient task execution in dynamic environments". The terms *"function"* and *"role"*, belonging to the *FunctionalWord* class, identify the explanation as *Functional* (Figure 6a).

Regarding the *Teleological Explanation*, which focuses on the goals of behavior, we distinguish two variants. The *Neutral Teleological* is represented by the explanatory text: *"The robot avoids hitting the door with the purpose of maintaining its objective: completing its task efficiently without damage or delays"*. The presence of *"purpose"* and *"objective"* terms from the *TeleologicalWord* class, places this explanation in the *Teleological* category (Figure 6b). Conversely, the *Rational Teleological*, which explains behavior as if it were guided by rationality, is exemplified by the text: *"The robot stopped in front of the door.*

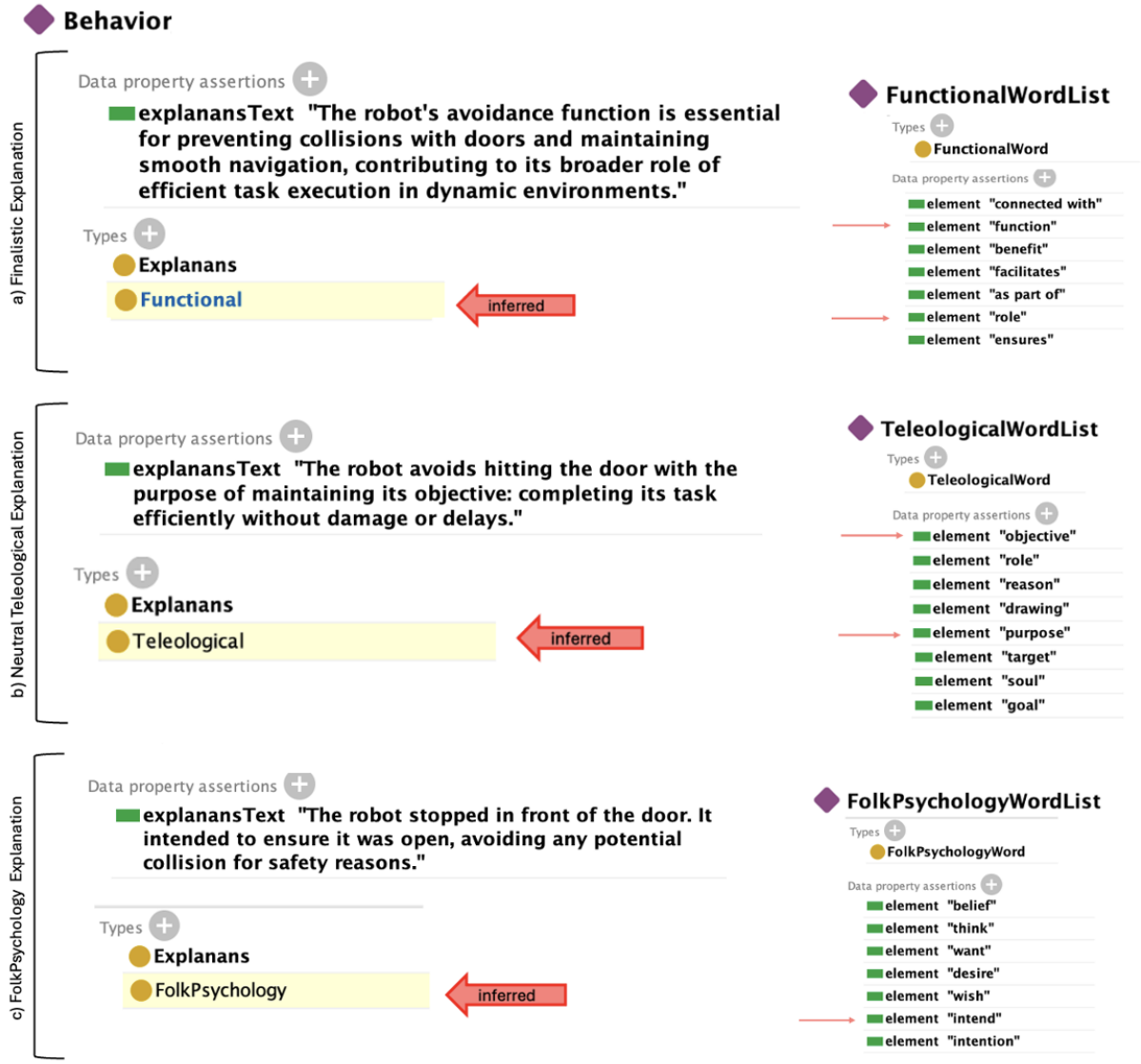

Figure 6: Example of a) Functional Explanation; b) Neutral Teleological Explanation; c) Folk Psychology Explanation.

It intended to ensure it was open, avoiding any potential collision for safety reasons". Here, the term *"intend"*, belonging to the Folk Psychology language present in the *FolkPsychologyWord* class, allows this explanation to be inferred as *FolkPsychology* (Figure 6c).

6. Limitations of The Work

The current preliminary work has the advantage of enabling - in a theoretically grounded way - the categorization of the types of explanation used by humans when observing robotic behaviors. There are, however, at the current state of affairs some issues that needs to be dealt with and that represent a current limitation of this approach. Namely: in some cases, the term based classification adopted can lead to assigning to the same Explanadum to different types of explanation due to the presence of shared terms across multiple *WordLists*. However, the specific meaning of these terms depends on the context of the sentence.

For example, the term *"function"* can be used both in *Mechanistic* explanations, where it describes the operation of a component within a system, and in *Functional* explanations, where it highlights the

function of a behavior within a broader context. Specifically, when analyzing the Explanans with the following explanatory text: *"The cooling function of the robot's processor is achieved through an internal system of heat pipes and fans that regulate the temperature during operation."* the term *"function"* refers to the internal and specific operation of the cooling system, which is an integral part of the robot's mechanism. Therefore, the explanation is *Mechanistic*.

On the other hand, in the Explanans: *"The cooling function of the robot's processor is essential to ensure its long-term performance and reliability in harsh environments"*, the word *"function"* refers to the general role of cooling in maintaining the robot's performance over time. Hence, the explanation becomes *Functional*.

This example demonstrates a current limitation of the HERB Ontology, as the use of terms alone is insufficient to classify an explanation. Instead, the context and explanatory intent determine the category to which it belongs.

7. Conclusion and Future Works

In order to mitigate the above mentioned limitations we are exploring two possible paths. The first one concerns the enrichment of formalization of the ontology with knowledge about the structure of the arguments used (i.e. the *Explanation Structure* category). The second one relies on the exploration of a dual process approach (see e.g. [16]) where the two options under consideration are wether to submit the results of the ontology to the assessment of a Large Language Model or vice versa.

8. Acknowledgments

This work is supported by funding by the European Commission - Next Generation EU - PNRR M4 - C2 -investimento 1.1: Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN) - PRIN 2022. Prot. 20224X95JC, as part of the research project titled HERB - Human Explanation of Robotic Behaviour. CUP H53D23004060006.

References

[1] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, et al., The role of explainable ai in the context of the ai act, in: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, 2023, pp. 1139–1150.

[2] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. URL: https://www.sciencedirect.com/science/article/pii/S1566253524000794. doi:https://doi.org/10.1016/j.inffus.2024.102301.

[3] A. Augello, Unveiling the reasoning processes of robots through introspective dialogues in a storytelling system: A study on the elicited empathy, Cognitive Systems Research 73 (2022) 12–20.

[4] C. F. Longo, P. M. Riela, D. F. Santamaria, C. Santoro, A. Lieto, A framework for cognitive chatbots based on abductive–deductive inference, Cognitive Systems Research 81 (2023) 64–79.

[5] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance, Frontiers in Computer Science 5 (2023) 1096257.

[6] M. Riveiro, S. Thill, The challenges of providing explanations of ai systems when they do not behave like users expect, in: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, 2022, pp. 110–120.

[7] M. M. De Graaf, B. F. Malle, How people explain action (and autonomous intelligent systems should too), in: 2017 AAAI Fall Symposium Series, 2017.
[8] J. P. Sullins, Trust in robots, in: The Routledge handbook of trust and philosophy, Routledge, 2020, pp. 313–325.
[9] K. Sokol, P. A. Flach, Glass-box: Explaining ai decisions with counterfactual statements through conversation with a voice-enabled virtual assistant., in: IJCAI, 2018, pp. 5868–5870.
[10] A. Lieto, Cognitive design for artificial minds, Routledge, 2021.
[11] C. G. Hempel, P. Oppenheim, Studies in the logic of explanation, Philosophy of science 15 (1948) 135–175.
[12] R. Cummins, Functional analysis, Journal of Philosophy 72 (1975) 741–765.
[13] E. Riloff, J. Shepherd, A corpus-based approach for building semantic lexicons, 1997. URL: https://arxiv.org/abs/cmp-lg/9706013. arXiv:cmp-lg/9706013.
[14] M. Darwich, S. A. Mohd, N. Omar, N. A. Osman, Corpus-based techniques for sentiment lexicon generation: A review., J. Digit. Inf. Manag. 17 (2019) 296.
[15] G. A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (1995) 39–41. URL: https://doi.org/10.1145/219717.219748. doi:10.1145/219717.219748.
[16] A. Augello, I. Infantino, A. Lieto, U. Maniscalco, G. Pilato, F. Vella, Towards a dual process approach to computational explanation in human-robot social interaction, in: Proceedings of the 1st CAID workshop at IJCAI, 2017.