RAGGAE for HERBS: Testing the Explanatory Performance of Ontology-powered LLMs for Human Explanation of Robotic Behaviors

Agnese Augello[1], Edoardo Datteri[3], Antonio Lieto[1,2], Maria Rausa[1], and Nicola Zagni[3]

[1] Institute for High Performance Computing and Networking, National Research Council of Italy
name.surname@icar.cnr.it
[2] University of Salerno
asurname@unisa.it
[3] University of Milano-Bicocca
name.surname@unimib.it

Abstract. In this work we present and test a RAG-based model called **RAGGAE** (i.e. **RAG** for the **G**eneral **A**nalysis of **E**xplanans) tested in the context of **H**uman **E**xplanation of **R**obotic **B**ehavior**S** (**HERBS**). The RAGGAE model makes use of an ontology of explanations, enriching the knowledge of state of the art general purpose Large Language Models like *Google Gemini 2.0 Flash*, *DeepSeek R1* and *GPT-4o*. The results show that the combination of a general LLM with a symbolic, and philosophically grounded, ontology can be a useful instrument to improve the investigation, identification and the analysis of the types of explanations that humans use to verbalize - and make sense of - the behavior of robotic agents.

Keywords: Human-Robot Interaction · Explanations · RAG · Large Language Models · Ontology

1 Introduction

One of the current focuses of Explainable Artificial Intelligence (XAI), a critical area of research, is the necessity for AI systems to make explicit how their underlying processes lead to certain outputs (in particular neural and probabilistic ones). On a more comprehensive note, a wider XAI focus is to enhance the capability of AI systems of becoming more interpretable and transparent to humans [13]. In this context, the ability to provide comprehensible and contextually relevant explanations is essential to foster trust and enabling effective interactions between users and AI systems.

Explanations of machine-driven outputs, however, represent only one of the many possibilities through which to analyze and understand Human-Robot Interaction.

In the context of studying HERBs (Human Explanation of Robotic Behaviors), we reverse focus by taking into account how humans explain (i.e. verbalize) and make sense of the behavior of social robots. In order to do so, we collected and analyzed human explanations of robotic behaviors, collected in social and educational settings. Through this process, we identified the need for a systematic procedure to gather and analyze explanations, particularly to provide a structured approach to support and streamline this process. Based on these considerations, this work first introduces a formalized ontology of explanations built upon a taxonomy of explanation types derived from philosophical theories. Then, we show how the proposed ontology - when used in a Retrieval-Augmented-Generation (RAG) [9] mode with a current state of the art Large Language Models (LLM) - is able to improve the classification capabilities of human explanations when compared with expert humans annotators. In particular, the ontology categorizes explanations into distinct types, such as mechanistic, causal, teleological, deductive-nomological and functional, offering a framework that primarily aims at supporting the analysis of explanations provided by individuals during Human-Robot Interaction (HRI). In the following section, we introduce different types of explanations formalized in the ontological model. Then, we briefly describe the HERB ontology and show how it has been integrated - via RAG - with *GPT-4o* [14], *Google Gemini 2.0 Flash* [6] and *DeepSeek R1* [5] LLMs. Consequently, in an experimental section, we describe the categorization results of our integrated RAGGAE model, comparing it both to the categorization where no ontology was used, and to the categorization provided by two expert human annotators. Discussion and conclusions end the paper.

2 Types of Explanations

The notion of "explanation" has been studied extensively in a number of disciplines starting from philosophy of science, to the early cybernetics to the current approaches in explainable AI. Different types of theories have been proposed to define what is a correct "explanation" from a scientific view point (for details in the context of AI and Cognitive Modelling we remind to [2],[11]). Here, we briefly recall some of the explanatory categories that have been of interest in the context of our study. The first type is the so called Deductive-Nomological (DN) Explanation. According to this view, introduced by Hempel and Oppenheim [8], there are some strict characteristics that an *explanans* (i.e. literally: what explains a certain phenomenon) has to satisfy in order to explain a given phenomenon. In particular, the *explanandum* (i.e. what has to be explained) is seen as something that needs to be logically derived, via deduction, from the explanans. While intuitively this theory adequately addresses a normative notion of explanation, (as it assumes that the explanans provides necessary and sufficient conditions to understand, where understanding is equalized to predicting, the explanandum), this sort of relationship between explanans and explanandum proves to be very strict, focusing exclusively on the general "why" (in line with a strong reductionist view), while many explanations look good to us without satisfying such

tight constraints, such as singular causal explanations (e.g. "the impact of my knee on the desk caused the tipping over of the inkwell" [16]). Another type of explanation is the so called "functional", where explaining consists of providing "a function that a system is believed to possess" [3]. 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 how a certain system of model is built, 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 classical running example from the biological domain. Let us suppose that our aim is to explain why chameleons change their skin color. This usually happens when a predator is present (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 the color configuration more often associated to a particular predator (e.g. birds), a possible answer could be that "the number of bird predators in chameleons' environment is major in 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 evolutional 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 [10] (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 mimetise themselves and escape from predators" she is simply providing an explanation about the scope of the phenomenon intended to explain. If we suppose to be interested to the mechanisms determining why chameleons change their color the above explanation is not sufficient. On the other hand, a satisfactory explanation (in this respect) would be the following "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". In particular, our satisfaction would probably be 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" [12] a kind of explanation able to shed light on the inner componential functioning that determine the behavior of a given system. In the example provided, the very simple mechanistic explanation was also a causal explanation. These different types of explanations (and their specializations) have been the ones in focus during our study and formalized in our ontology.

3 The HERB Ontology

The HERB (Human Explanation of Robotic Behavior) ontology provides a first 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 [4], integrating SWRL rules [5] to enhance semantic inference and explicitly define the concepts, relationships, and governing rules behind these categorizations.

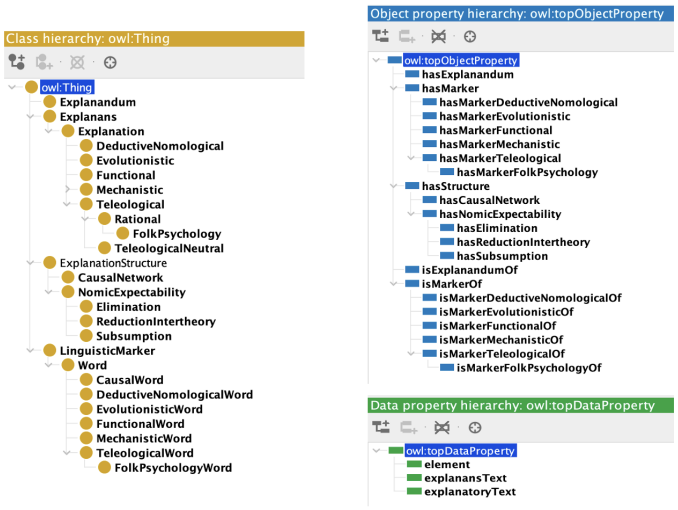

Fig. 1. A Taxonomy of Classes (in yellow), Object Properties (in blue) and Data Properties (in green) of the HERB Ontology.

3.1 Classes, Object Properties and Data Properties

The core Classes of the HERB ontology 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*).

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

In particular, as indicated before, *DeductiveNomological* explanations are 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 (e.g. beliefs, desires, intentions) [4].

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, *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 (see for details [15]).

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 subproperties 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 Experimental Setup

In order to acquire data consisting in verbally expressed accounts of robotic behaviors, we recruited participants that were requested to explain the behavior of robots in different scenarios (the different scenarios were provided by showing videos of different robotic behaviors). Afterwards, we built RAGGAE by used the HERB ontology as a symbolic component to extend and deepen the knowledge of LLMs about explanations[6]. The results of RAGGAE were compared with those obtaineed by the LLMs (without RAG) and with a baseline represented by the categorization, of the same explanandum, provided by two expert human annotators (i.e. two philosophers of science working on the epistemology of the different types of explanations). These different steps are described below.

4.1 Participant Recruitment and Data Collection

In our study, we involved 74 participants, recruited through mailing lists, social networks, and word of mouth. The inclusion criteria require participants to be over 18 years old and fluent in Italian. Participation was entirely voluntary. All provided signed informed consent.

Each participant is asked to watch a series of short videos, each lasting no more than two minutes, depicting various robotic behaviors in social and educational settings. In these videos, the humanoid robot Pepper interacts with a human counterpart in scenarios specifically designed to elicit explanations from the observer. The situations are inspired Strange Stories by Happé [7] , a classic tool used to assess Theory of Mind (ToM), and they differ in terms of complexity, everyday familiarity, mentalistic content, and the nature of the robot's behavior. Some videos show the robot entering a half-empty room and moving around in an apparently random way, pausing briefly in front of an object – either a box or a plush toy – inviting different interpretive responses. Other scenes depict more socially complex interactions, such as an encounter in a hallway between a woman carrying a box and the robot, which may respond either by politely yielding the way or by acting in an ambiguous, socially uncooperative manner. In another scenario, Pepper serves as a receptionist for students looking for internships, reacting differently depending on the appropriateness of the student's behavior – in one case failing to intervene in response to an inappropriate attitude, and in another, calmly redirecting the person to a human operator.

After each video, participants are invited to describe what they saw, highlighting the aspects that captured their attention and, more specifically, answering questions aimed at explaining what the robot did, why it did it, and how.

[6] The system is exposed at `https://www.ciitlab.org/agent.html`.

These verbal explanations are then transcribed and serve as the foundational dataset for the subsequent classification phase.

4.2 Classification Methodology

We focused our analysis on the participants' responses to the question *"Why did it do that?"*, coding each of the 74 explanations according to categories derived from the philosophy of science. The categories used were: Deductive-Nomological, Mechanistic, Causal, Evolutionary, Functional, Neutral Teleological, and Folk Psychology Teleological. The classification process was carried out in three distinct phases:

1. **LLM** – In this phase, a language model was provided with a prompt that required classifying the explanations of the robot's behavior according to the theoretical categories listed above. The prompt included definitions and specific examples for each category. The model was asked to interpret and assign each explanation to the predominant category, even in the presence of long texts.
2. **LLM+RAG** – Here, the same prompt from Phase 1 was used, but with the addition of knowledge derived from a file containing an ontology. This allowed the creation of **RAGGAE**, a system that integrates the symbolic component of the HERB ontology to expand and deepen the model's understanding of epistemological explanations.
3. **Human Annotations** – Two experts (philosophers of science with specific expertise in the epistemology of explanations) independently classified each response. When an explanation was missing, the label *ExplanationMissing* was assigned. Explanations that did not fit into the predefined categories were labeled as *Other*, or classified under a new category, if deemed relevant.

The classification from *Phase 3* serves as the baseline for evaluating the performance of computational models.

4.3 Baseline and Model Performance

Once the three sets of classifications were obtained —those produced by RAGGAE, the Large Language Model (LLM) without ontological support, and the two human annotators — their outputs were compared using a baseline based on the labels assigned independently by two expert annotators (Annotator1 and Annotator2). The inclusion of two experts aimed to reduce the influence of individual subjectivity and to enhance the reliability of the reference labels used for evaluating the automated models. To this end, we calculated the Inter-Annotator Agreement (IAA) [1] using Cohen's Kappa coefficient, a statistical measure that quantifies the level of agreement between two raters for qualitative classifications.

The construction of the baseline followed a clearly defined procedure. Instances where both annotators provided either generic or null responses (such as *"explanation missing"*, *"other"*, or *"unclassifiable"*) were excluded from the

analysis, as they offered no informative reference for automatic evaluation. In cases where only one annotator provided a valid classification, the available label was adopted as the reference. Finally, when both annotators assigned valid but potentially different labels, both were retained for model comparison.

Based on this baseline structure, we evaluated how closely the labels assigned by the model aligned with the annotations provided by the human experts. We considered two metrics, a *strict accuracy* to evaluate how often the label assigned by the model matches the annotations provided by both annotators, and a *partial accuracy* to evaluate how often label assigned by the model matches at least one of the two annotations provided by the annotators, using the following formulas:

$$\text{Strict Accuracy} = \frac{\#\left\{i \mid L_i = A_i^{(1)} \wedge L_i = A_i^{(2)}\right\}}{N}$$

$$\text{Partial Accuracy} = \frac{\#\left\{i \mid L_i = A_i^{(1)} \vee L_i = A_i^{(2)}\right\}}{N}$$

Where:

- L_i: label assigned by the model for the i-th explanation.
- $A_i^{(1)}$, $A_i^{(2)}$: labels assigned by the two annotators.
- $\#\{\cdot\}$: number of cases satisfying the condition.
- N: total number of explanations.

Based on these accuracy scores, we identified the most reliable model. For this model, confusion matrices were generated in relation to each annotator's classifications to further analyze classification patterns and mismatches.

The results of this analysis are reported in the following section.

5 Results

The analysis of the Inter-Annotator Agreement (IAA) shows a moderate level of agreement between the two evaluators, with a Cohen's Kappa coefficient of 0.25. This value reflects some variability in the assignments, further emphasizing the importance of a structured comparison between multiple annotations.

Using the LLMs *Google Gemini 2.0 Flash*, *DeepSeek R1* and *GPT-4o*, we computed accuracy scores under two settings: with and without the integration of the symbolic component RAGGAE. The results, shown in Figure 2 (strict accuracy) and Figure 3 (partial accuracy), highlight performance improvements when the models are supported by the HERB ontology via RAGGAE.

The comparison reveals that: for *Google Gemini 2.0 Flash*, the integration of RAGGAE significantly improves both strict accuracy, from 16.1% to 24.2%, and partial accuracy, from 43.5% to 56.5%; in the case of *DeepSeek R1*, the model is unable to generate any valid classifications without RAGGAE (0% for both accuracy types), but successfully classifies when RAGGAE is applied (strict: 12.9%, partial: 37.1%); while, for *GPT-4o*, RAGGAE leads to an improvement

in strict accuracy from 14.5% to 17.7%, but a decrease in partial accuracy, from 43.5% to 35.5%.

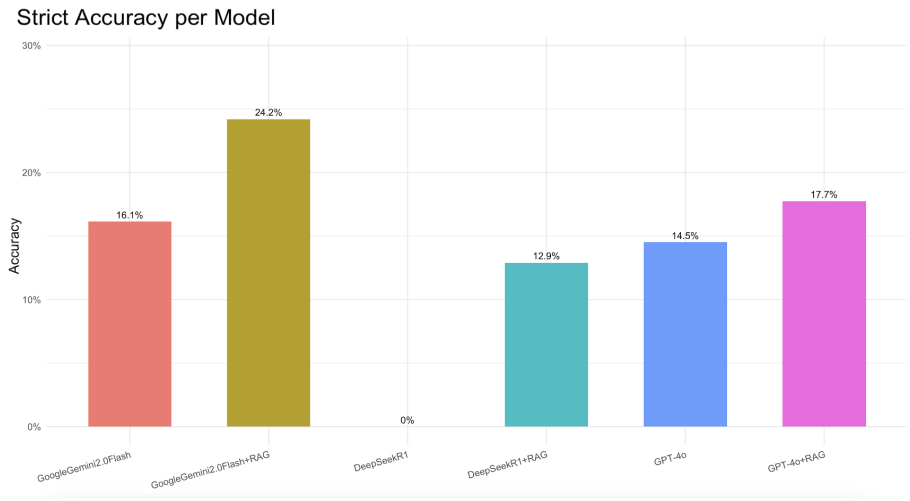

Fig. 2. *Strict accuracy* scores for LLMs and their RAGGAE-enhanced versions (LLM+RAG) across the models: *Google Gemini 2.0 Flash*, *DeepSeek R1*, *GPT-4o*.

These results confirm that, overall, the use of RAGGAE enhances model accuracy. Particularly, among all evaluated models, Google Gemini 2.0 Flash with RAGGAE achieves the best overall performance and is therefore selected as the reference model for the in-depth analysis. As an additional analysis, two confusion matrices were generated comparing the labels produced by the top performing RAGGAE model (i.e. the one with Google Gemini 2.0 Flash) with two annotators (Figure 4 and 5). These matrices provide a detailed view of the areas of convergence and disagreement between the automatic model and the human evaluators. Specifically, the matrix in Figure 4 shows a fair alignment for the *TeleologicalNeutral* class (8 matches) and *Unclassifiable* (9 instances). However, numerous overlaps with other categories emerge, particularly among *TeleologicalFolkPsychology*, *Functional*, and *Mechanistic*. For example, some instances labeled as *Functional* by the annotator were often classified by the model as *Mechanistic* and *TeleologicalNeutral*, suggesting a conceptual overlap. Additionally, the *ExplanationMissing* class frequently overlaps with *TeleologicalNeutral*, indicating a possible tendency of the model to assign teleological interpretations even in the absence of an explicit explanation. The second matrix, in Figure 5, displays a different distribution. The model shows strong agreement with the annotator in the classification of the *TeleologicalNeutral* category (15 matches), *Mechanistic* (13 matches), and *Unclassifiable* (8 matches). Nonetheless, several misclassifications occur between *Mechanistic* and *TeleologicalNeutral*: as many

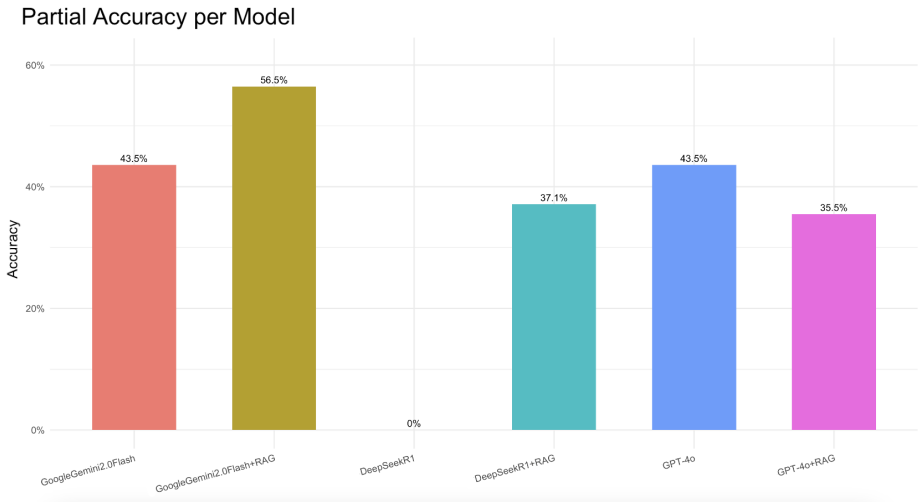

Fig. 3. *Partial accuracy* scores for LLMs and their RAGGAE-enhanced versions (LLM+RAG) across the models: *Google Gemini 2.0 Flash*, *DeepSeek R1*, *GPT-4o*.

as 7 instances labeled as *Mechanistic* were classified by the model as *TeleologicalFolkPsychology*. This once again reflects the difficulty in distinguishing explanations influenced by subtle linguistic nuances. It is also worth noting the poor alignment in the *ExplanationMissing* category, which the model struggles to identify correctly in both comparisons. Overall, the two matrices confirm the findings from the IAA analysis, showing that while model performance improves with the integration of the HERB ontology, RAGGAE still exhibits significant ambiguity in conceptually related classes, reflecting both model limitations and potential divergences between annotators.

6 Conclusions and Future Works

The obtained results show how the adoption a philosophically grounded ontology of human explanations of robotic behavior (HERBs), when used in a RAG model (RAGGAE), improves the explanatory performance of AI systems based on human verbalization of the behavior of social robots. While the current datum is of interest, even if it deserves further investigations with a larger number of LLMs - it is worth-noticing how, for this complex task, the performance of AI systems are still very far from being comparable to human expert annotations. As future works we plan to better axiomatize (via knowledge specialization and extension when needed) the current version of the ontology. This task will allow to improve the formal structure that can be superimposed to LLMs and, as a consequence, its categorization accuracy. In addition, we plan to extend our evaluation both acquiring and analyzing more verbal data and by extending the

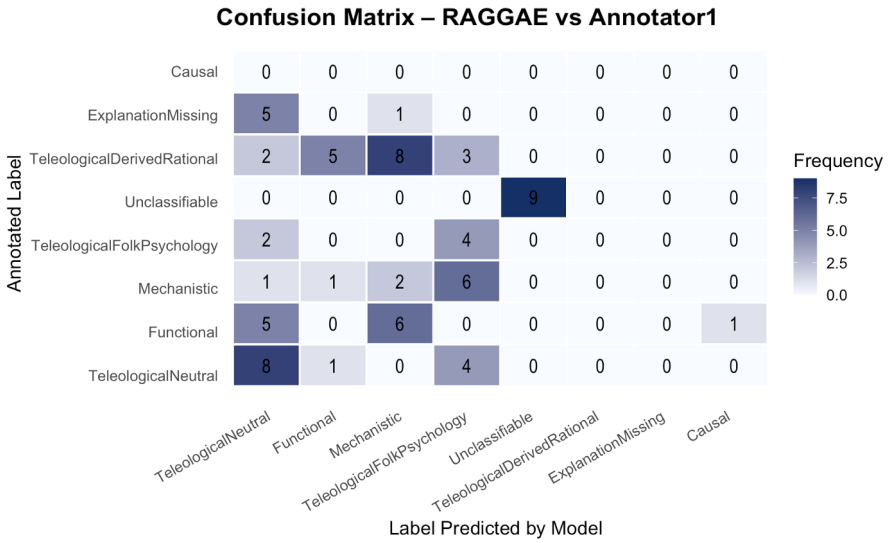

Fig. 4. Confusion matrix of the best RAGGAE model (*Google Gemini 2.0 Flash*) vs the labels provided by *Annotator1*.

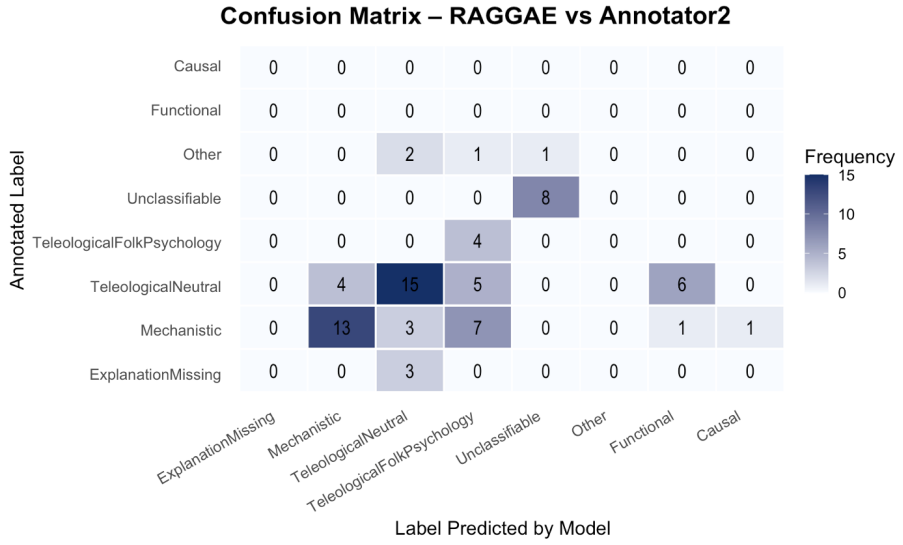

Fig. 5. Confusion matrix of the best RAGGAE model (*Google Gemini 2.0 Flash*) vs the labels provided by *Annotator2*.

number of human annotators in order to have a more robust ground truth upon which to compare the results of RAGGAE.

References

1. Ron Artstein. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313, 2017.
2. Roberto Cordeschi. *The discovery of the artificial: Behavior, mind and machines before and beyond cybernetics*, volume 28. Springer Science & Business Media, 2002.
3. Robert Cummins. Functional analysis. *Journal of Philosophy*, 72(20):741–765, 1975.
4. Daniel C Dennett. Intentional systems. *The journal of philosophy*, 68(4):87–106, 1971.
5. DeepSeek-AI et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, 2025.
6. Google. Gemini 2.0 Flash: https://gemini.google.com/app.
7. Francesca GE Happé. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154, 1994.
8. Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175, 1948.
9. Yucheng Hu and Yuxing Lu. Rag and rau: A survey on retrieval-augmented language model in natural language processing. *arXiv preprint arXiv:2404.19543*, 2024.
10. Mariska Leunissen. *Explanation and teleology in Aristotle's science of nature*. Cambridge University Press, 2010.
11. Antonio Lieto. *Cognitive design for artificial minds*. Routledge, 2021.
12. Peter Machamer, Lindley Darden, and Carl F. Craver. Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25, 2000.
13. Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.
14. OpenAI. GPT-4o: https://chatgpt.com/?model=gpt-4o.
15. Maria Rausa, Agnese Augello, and Antonio Lieto. Towards an ontology of human explanations of robotic behavior. In *Proceedings of the Fifth Workshop on SOcial and Cultural IntegrAtion with PersonaLIZEd Interfaces (SOCIALIZE) at the 30th Annual ACM Conference on Intelligent User Interfaces Cagliari, Italy · March 24-27, 2025 ACM IUI 2025*, 2025.
16. Michael Scriven. Explanations, predictions, and laws, 1962.