ADEA: An Argumentative Dialog Dataset on Ethical Issues concerning Future A.I. Applications

Christian Hauptmann, Adrian Krenzer, Antonia Krause, Frank Puppe
Julius-Maximilians-Universität Würzburg (JMU)
Am Hubland, 97074 Würzburg, Deutschland
{christian.hauptmann, adrian.krenzer, antonia.krause, frank.puppe}@uni-wuerzburg.de

Abstract

Introducing *ADEA*: a German dataset that captures online dialogues and focuses on ethical issues related to future AI applications. This dataset, which includes over 2800 labeled user utterances on four different topics, is specifically designed for the training of chatbots that can navigate the complexities of real-world ethical AI conversations. The creation of these dialogues is the result of two carefully conducted studies in which university students interacted with an argumentative dialogue system. A fundamental part of our methodology is the use of German argument graphs. These graphs not only form the knowledge base of the dialogue system but also serve as an effective annotation scheme for the dialogues. Apart from the introduction of the dataset and the argument graphs, we provide a preliminary benchmark using GPT-4 via the OpenAI API. This provides researchers with a concrete reference point while demonstrating the potential of our dataset. We make our dataset and argument graphs available at https://github.com/HaupChris/ADEA-Dialogue-Dataset.

Keywords: argumentative dialogue dataset, argument graph, ethics and AI

1. Introduction

The discourse surrounding the applications of artificial intelligence (AI) occupies a prominent place in today's social dialogue. While technical feasibility often dominates these conversations, it is important not to overlook the societal implications and ethical challenges associated with AI. Kahneman, 2011 highlights that humans mainly rely on intuitive thinking for decision-making. However, this way of thought is susceptible to overconfidence and confirmation bias. When it comes to ethically complex issues, such as those raised by certain applications of AI, there is often no clear-cut solution. The method of reflective equilibrium suggests that individuals must weigh their immediate judgments against broader ethical principles, adjusting either as necessary to arrive at a coherent stance (Cath, 2016). This iterative process emphasizes the importance of a comprehensive understanding, enabling individuals to move beyond mere intuition to more considered decisions.

In this paper, we present *ADEA*, a dataset consisting of dialogues on four ethical topics related to future AI applications. These dialogues are the result of two different studies conducted between individuals and a German conversational system to improve knowledge about AI ethics topics through argumentative dialogues. We use an argument graph both as the system's knowledge base and as an annotation scheme for the dataset. Our contributions are:

(1) A two-stage annotated German argumentative dialogue dataset containing more than 2800 user utterances,
(2) four German argument graphs that provide the basis for structured discussions on AI ethics, and
(3) A preliminary benchmark using the GPT-4 API (Bubeck et al., 2023), highlighting the value of the dataset in German argument mining and conversational AI research.

2. Related Work

We structure this section into two subsections: Argument graphs and datasets. The literature review on argument graphs is important as, alongside our dataset, we introduce companion argument graphs.

2.1. Argument Graphs

In recent years, the argument mining community has become increasingly interested in detecting arguments in texts and identifying their interrelationships. Boltužić and Šnajder, 2015a addressed this by identifying and linking arguments in online debates using cluster analysis. In contrast, Reimers et al., 2019 used contextualized word embeddings to classify and group arguments at sentence level. Mayer et al., 2020 used a transformer-based method to detect argument components and their relations, while Trautmann et al., 2020 presented a technique and a dataset for locating and categorizing argument units within sentences.

While these methods focus primarily on the identification and classification of arguments within texts, the development of conversational agents in the argumentation domain requires structured representations that enable them to counterarguments.

LREC-COLING 2024, pages 426–437
20-25 May, 2024.

A widely accepted strategy is to adopt a graph-based knowledge representation: Many studies have developed dialogue agents based on argument graphs. For example, Hadoux et al., 2018 constructed an argument graph that integrates beliefs and emotions, on the topic 'annual flu vaccination for hospital staff' with 50 arguments. Chalaguine and Hunter, 2019 build an extensive crowdsourced graph that was later used to debate university fees via a chatbot (Prakken et al., 2020) and another on meat consumption (Chalaguine et al., 2019). During the COVID-19 pandemic, argumentative dialogue systems supported by argument graphs became popular for persuasion. (Chalaguine and Hunter, 2021; Fazzinga et al., 2021). Finally, Aicher et al., 2022 presented a voice-activated dialogue system anchored by a 72-component argument graph.

It's worth noting that these studies predominantly use English. While there are German-language works on argument graphs (Dumani et al., 2021; Mirko et al., 2020), to the best of our knowledge, this is the first study to create and use them for argumentative dialogues on ethically challenging topics.

2.2. Datasets

Developments in the past years have brought datasets that focus on the structure and components of arguments. For example, Stab and Gurevych (2014) provide a collection of 90 persuasive essays from an online student forum, focusing on the annotation of argument components and relationships in persuasive texts. Similarly, the DART dataset from Bosc et al. (2016) systematically breaks down arguments on X (formerly known as Twitter), detailing their interrelationships. There is also the work of Wambsganss et al. (2020), which annotates argument components and relations in a dataset of 1000 persuasive German student reviews. Another contribution is the Internet Argument Corpus of Abbott et al. (2016), which captures arguments in various debates from online forums on topics ranging from gun control to the death penalty.

Moving from pure argumentation to dialogue-based datasets, there are several corpora of dialogues, e.g. Xu et al. (2021) with over 3 million context-response pairs from Ubuntu's IRC channels, the Wizard of Wikipedia, which contains open domain dialogues grounded on Wikipedia (Dinan et al., 2018), the Daily Dialogues Corpus (Li et al., 2017) or the 'FinChat' dataset by Leino et al. (2020), which explores Finnish chat-based conversations.

However, there are far fewer resources for argumentative dialogues and even fewer for German. The closest to our work is the 'ArgSciChat' dataset by (Ruggeri et al., 2022), which contains 41 dialogues between scientists about research papers and annotates argumentative and explorative units within these conversations. Additionally, Romberg and Conrad (2021) created a dataset that captures public participation in urban planning, allowing users to suggest changes on certain parts of the city map and other citizens to comment on them. The corpus annotates argumentative and non-argumentative components.

To our knowledge, there are no resources that focus on ethical AI issues that arise in a dialogue setting. We fill this gap by providing such a resource that can be used for the development of conversational agents.

3. Conducting Ethical AI Dialogues

Our user studies are based on a German-specific argumentative dialogue system designed to explore the ethical considerations of AI applications. The system integrates a mobile-optimized web interface, a backend server, a text processing unit, and a core knowledge base. Upon a user's selection of an AI ethics topic and his stance on the question of discussion, the chatbot adopts a counter-viewpoint to stimulate diverse discussions. We employ the system of Hauptmann et al., 2024, which is designed for argumentative dialogue, allowing users to respond with written text. For the identification of user arguments, it combines a fine-tuned Sentence-BERT (Reimers and Gurevych, 2019) approach with knowledge-based filtering to use dialogue context to map user utterances onto nodes of an *argument graph (AG)*. The identified graph node is used to build a response in natural language from predefined sentence components.

This AG is essential for the system's ability to engage in meaningful dialogues since it serves as both a knowledge base and an annotation scheme. This paper focuses on the AG and the dialogue dataset. Each AI ethics topic's knowledge base includes:

- **Scenario and Question of Discussion**: A hypothetical future setting, e.g. "Imagine a future where AI systems autonomously make medical diagnoses and therapeutic decisions within medical centers.", with a central question to determine user stance, e.g. "Should AI systems be allowed to make autonomous medical decisions without human oversight?"
- **Argument Graph**: A hierarchically structured representation of pro and con arguments, curated by PhD students specializing in internet and computer science law.
- **FAQs**: Common queries about the scenario, derived from preliminary studies.

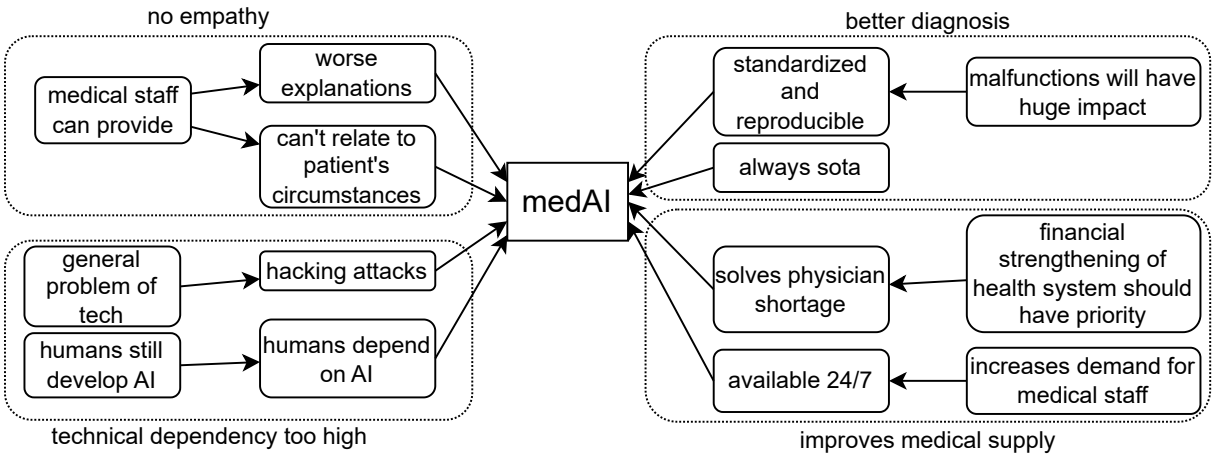

Figure 1: Translated excerpt from the medAI argument graph. Argument graph nodes are in rectangles with rounded corners. Arrows between nodes represent the "counters" relationship, with counterarguments pointing at the attacked argument. The boxes with dotted lines represent argument groups.

Each topic's graph has nodes for both main and counterarguments. Main arguments stand alone, while counterarguments refer to another, even possibly another main argument. Table 1 shows the sizes of the argument graph for each topic. This structure follows the definition of Hadoux and Hunter (2019):

Definition 1. *An argument graph* $\mathcal{G} = (\mathcal{A}, \mathcal{R})$ *consists of nodes* $\mathcal{A}$ *and arcs* $\mathcal{R}$ *where* $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$. *In this directed graph, each element* $A \in \mathcal{A}$ *is an argument. If* $(A_i, A_j) \in \mathcal{R}$ *then* A_i *counters* A_j

Each node encompasses a *label*, *summary text*, *a full text*, and a *rating* (integer between 0 and 5). An example is shown in figure 2. Arguments with higher ratings are more likely to be used as a response by the bot. Additionally, each argument is assigned to an argument group, gathering arguments that address similar aspects, like social or technical arguments. Figure 1 shows an excerpt from the argument graph on the topic medical AI.

Label: P1
Summary: MedAI will improve medical care.
Full Text: A medAI is available around the clock; at night or on weekends, even if only in emergency staffing with medical support personnel.
Rating: 5

Figure 2: Example of an argument node's content, translated into English. "P" indicates a pro argument, with "1" serving as an identifier.

The argument graph has the following three roles:

1. The system attempts to map user utterances to graph nodes to identify the user's intent.
2. Once a successful mapping is achieved, arguments from the graph are employed to further the dialogue. This can be in the form of a direct counterargument or, in the absence of a counter, a main argument. The bot therefore leverages texts from the graph nodes to construct a response.
3. After completion of the study, the graphs function as an annotation scheme that annotators employ to assign arguments from the graph to user utterances (as explained in section 5).

Table 1: Number of arguments for each topic's argument graph.

Topic	Main Arguments	Counterarguments
MedAI	25	33
LawAI	22	45
CarAI	29	50
RefAI	20	58

4. Dataset Collection

To analyze AI ethical discussions, we conducted two user studies with our German-specific dialogue system, discussing the following AI ethics topics:

- **MedAI**: Permission of a future AI system to decide on diagnoses and therapies on its own.
- **LawAI**: AI judges in civil law processes.
- **CarAI**: Cars that drive completely autonomously without human intervention.
- **RefAI**: Replacing referees in soccer matches (introduced only in the second study).

The system identifies and counters user arguments, leveraging a template-based approach for responses. If it maps a user's argument to a graph node, the response utilizes the node's summary text for alignment with the user intent. Subsequently, the system then offers a counterargument or a contrasting proactive argument. In the first study, unrecognized arguments prompted users to rephrase. Feedback led to a refined approach in the second survey, where users chose their argument from a dropdown list, improving dialogue flow.

We refined the dialogue system for the second study, using data from the first, and introduced the RefAI topic. The goal was to evaluate the system's knowledge dissemination and influence on user attitudes, as well as to curate a chatbot training dataset. The first study yielded 178 dialogs, and the second produced 200. Students from the University of Würzburg were recruited as participants for both studies. Each participant was given a short instruction sheet containing a QR code with a link to the chatbot website, as well as a short explanation of how to start and end the chat and the instruction to chat for at least 6 conversation turns or 10 minutes. To ensure a realistic application scenario, the study could be conducted on a smartphone and participants did not have to stay on site. A typical session began with participants scanning the QR code, leading to the web interface. After selecting a discussion topic and stance, participants engaged in a dialogue with the system, culminating in a post-dialogue survey.

Central to the system is the argument graph, which is used for mapping user utterances and guiding the dialogue. Despite its structured nature facilitating most dialogs, challenges emerged when users introduced out-of-scope arguments, highlighting areas for future improvement.

Having detailed our dataset collection method, which yielded a total of 378 dialogs across two studies, we now turn our attention to the specifics of how this data was annotated.

5. Annotation

5.1. Annotation Guidelines

Similar to Stab and Gurevych (2014), our annotation scheme labels every phrase within a user utterance. To handle diverse textual units within utterances, annotators use sentence boundaries as separators, splitting a sentence into multiple units only if it contains multiple arguments. For label annotation, we differ between:

1. **Argumentative Units**: Annotators first identify if a segment contains an argument. If so, they categorize it as:
   - **Well-Founded Arguments**: Logically sound claims aligned with a graph node.
   - **Unfounded Arguments**:
     - Violate scenario assumptions (e.g., AI should only support physicians).
     - Lack a clear premise (e.g., "AI makes life easier").
     - Present a false claim (e.g., "Human reaction time is shorter than that of AI").

     Some frequent unfounded arguments, such as the notion that AI should act only in a supportive role, are mapped onto a dedicated graph section. If a direct node match for any argument isn't feasible, annotators opt for a broader semantic group label. If this isn't possible, the argument should be classified as well-founded or unfounded, and further as pro or con.
2. **Non-Argumentative Units**: For non-argumentative segments and queries tied to the discussion scenario, annotators tried to match them to a specific FAQ label or use a general FAQ group label if a direct match isn't possible.
3. **Miscellaneous Units**: All other segments are labeled as expressions of agreement/disagreement or *miscellaneous*.

Annotators note reasons for decisions between two specific arguments for a certain text unit. This approach ensures comprehensive and granular annotation of user utterances, capturing both semantic and argumentative essence. Figure 3 shows a dialog dataset excerpt. The system labels bot utterances, while our annotators label user utterances.

5.2. Inter Annotator Agreement

Inter Annotator Agreement (IAA) evaluates the consistency and clarity of our annotation guidelines, being crucial for tasks with subjective interpretations. A high IAA indicates that the annotation process is consistent and the guidelines are clear, while a low IAA suggests potential ambiguities in the guidelines or differences in annotators' interpretations.

Our study assesses the IAA on two fronts: the labels assigned to textual units in user utterances and the boundaries of these textual units.

Table 2 presents the IAA for labels assigned to textual units. We used Cohen's Kappa (κ) (Cohen, 1960), which accounts for chance agreement. A κ value above 0.6 is generally considered as substantial agreement (Landis and Koch, 1977). As observed, all topics, except for the *LawAI* in study 2, exhibit substantial agreement, indicating that the annotators largely concurred on the labels.

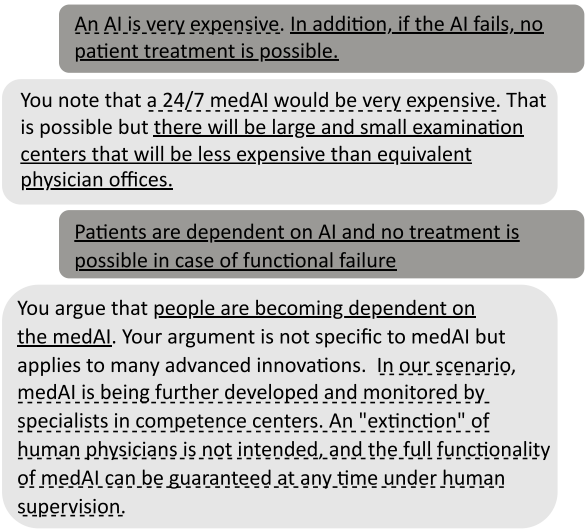

Figure 3: Translated excerpt from a medAI user study dialog. User utterances are right-aligned bubbles, and bot utterances are left-aligned. Arguments within each utterance are underlined, with solid and dashed lines distinguishing different arguments. Bot responses are templated and auto-annotated. Not underlined units are labeled in the corpus as well.

Study	Topic	Kappa κ	Obs. Agreement
1	MedAI	**0.62**	58.66
	LawAI	**0.69**	66.84
	CarAI	**0.62**	60.07
2	MedAI	**0.61**	55.64
	LawAI	0.58	56.63
	CarAI	**0.71**	65.19
	RefAI	**0.66**	63.7

Table 2: Inter Annotator Agreement for labels assigned to textual units in user utterances by two different Annotators. We report Cohens Kappa κ (Cohen, 1960) and observed agreement (%). The κ values indicate substantial agreement for all but the LawAI 2 dataset. (Landis and Koch, 1977)

However, the observed agreement scores, suggest there's still room for improvement.

The second aspect of IAA, presented in Table 3, focuses on textual unit boundaries. Here, we considered two scenarios: exact match and split. An exact match means that both annotators have chosen to include the same words in a textual unit. Split indicates scenarios where annotators disagreed on the boundaries by at least one word. Across all topics, exact match percentages exceeded 92%, demonstrating strong agreement on textual unit boundaries. However, the presence of splits highlights the subtle complexities of defining textual units within the dynamic context of user dialogues.

Study	Topic	Exact Match	Split
1	MedAI	95.86	4.13
	LawAI	95.41	4.59
	CarAI	96.04	**3.96**
2	MedAI	92.97	7.02
	LawAI	96.48	3.52
	CarAI	**96.62**	5.38
	RefAI	97.03	**2.96**

Table 3: IAA as observed agreement (%) for boundaries of textual units in user utterances. Split means no exact match with different word boundaries.

Study	Topic	Labels only	Labels & Bounds
1	MedAI	**87.25**	12.75
	LawAI	**84.62**	15.38
	CarAI	**86.19**	13.51
2	MedAI	**80.33**	19.25
	LawAI	**79.61**	20.39
	CarAI	**83.11**	16.89
	RefAI	**91.84**	8.16

Table 4: Distribution by topic (%) of two types of annotation disagreement.

To improve the quality of the dataset, we initiated a second annotation stage, focusing on resolving disagreements between annotators. An experienced third annotator, familiar with the first two annotations, played a key role at this stage. Relying on deep domain knowledge, this annotator effectively resolved disagreements and ensured that the annotations were consistent and accurate. This process not only improved the consistency of the dataset, but also highlighted different types of disagreement, either caused by the nuances of the guidelines or by the annotators' interpretations. They provide valuable insights into the challenges of the annotation process. The analysis of these disagreements is discussed in more detail in the following section.

6. Disagreement Analysis

When annotating complex tasks, disagreements can arise due to the complexity of the guidelines, nuances of content, or individual interpretations. In our dataset, disagreements can be categorized based on labels, boundaries or both. Table 4 shows the distribution of different types of annotation disagreement. As there are only two cases of pure boundary disagreements due to mid-sentence splits, we focus on label and combined disagreements. In the following, we describe different types of disagreements and how they were handled by the third annotator.

6.1. Label-only Disagreements

In the second labeling stage, the third annotator, typically having to decide between two suggestions, turned to the argument graph only when neither seemed fitting. Such decisions, especially in ambiguous cases, benefited significantly from the expertise of this third annotator. Most disagreements (ranging from 80% to 92% across topics) were resolved by the third annotator choosing one of the two initial annotations. This method, being simpler than annotating from scratch, likely enhanced dataset quality (cf. figure 4).

Certain user utterances had plausible annotations from both initial annotators. These cases required extra discernment: Some utterances matched multiple arguments in the argument graph. To navigate such ambiguities, distinguishing features of similar arguments were identified to guide decisions. When both annotations appeared appropriate, subsequent dialogues provided clarity, particularly in terms of coherence and alignment of counterarguments. Occasionally ($< 5\%$ per topic), user nuances made it challenging to pinpoint the exact argument. In such complex situations, a model predicting either of the two labels would be adequate in dialogue contexts.

Moreover, instances arose where the FAQ labels were found to be imprecise. Strict adherence to the guidelines rectified this. The third annotator, apart from mediating between initial annotations, documented cases where they had to derive a correct solution from the argument graph (observed in 9% to 22% of disagreements). Even though this approach is prone to error, these cases were commented on with special attention because the annotator initially had to evaluate both annotations as not fitting. The fact that in most cases a decision between two labels was possible underscores the effectiveness of the annotation process and further improves the quality of the dataset.

6.2. Labels and Bounds Disagreement

Disagreements in this category were primarily similar to those in the label-only section. The majority were resolved by aligning with one of the annotators' boundaries and labels. However, there were occasional ambiguities, similar to those previously discussed, that caused disagreements, especially when an utterance resonated with multiple arguments.

Non-compliance with guidelines, specifically setting boundaries within sentences, led to other disagreements. Boundaries must cover complete sentences; failing this can change not only boundaries but labels, as additional labels, such as consent or dissent, may be incorporated within the utterance.

In some rare instances, an annotator missed an argument or overlooked an utterance entirely. The third annotator's expertise proved invaluable here, determining whether to accept one annotator's label or seek an alternative.

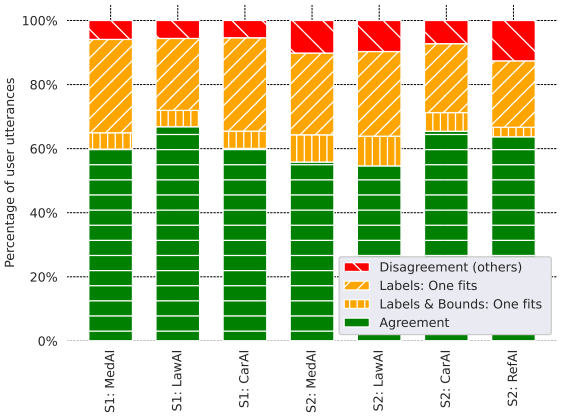

Figure 4: Inter Annotator agreements by topic. Disagreements resolved by the third annotator, aligning with one prior annotation, are denoted as *One fits*. Disagreements (other) are the cases where the third annotator did not choose one of the two pre-existing annotations.

While our first annotation process witnessed substantial agreement between annotators, we recognized the importance of a second annotation stage to further refine and enhance the dataset's quality. The majority of disagreements were straightforward to address, which can also be seen in Figure 4. For labels, in most instances, one of the initial annotations was accurate, simplifying the resolution process. When it came to boundaries, the initial agreement was commendably high.

The task of the third annotator was inherently less error-prone than the initial annotation from scratch, since this primarily involved deciding between two pre-existing annotations instead of performing a search in the whole graph. This approach not only ensured the dataset's robustness but also likely reduced the potential for errors.

7. Dataset Statistics

The dataset captures interactions from two user studies, split across several AI ethics topics. In total, these studies produced 378 dialogs with 2,880 user utterances. Table 5 provides a breakdown for each topic.

Dialogs averaged 7.8 user turns, and user utterances contained about 12.3 words on average. The length of dialogs varied, but they were generally reflective of in-depth discussions on the respective topics.

One metric of interest is the number of distinct arguments per dialogue. On average, users pre-

Study	Topic	Dialogue	User Utterances			Distinct Args. per Dia.			Arguments		Non-Arguments	
		Count	Count	Avg. Words	Avg. Count	User	Bot	Union	WF	UF	Q	Misc
1	MedAI	62	519	12	8.4	3.6	6.6	8.9	58.96	10.79	4.43	26.20
	LawAI	26	203	13.2	7.8	4.1	9.4	12.1	67.00	5.91	1.48	26.11
	CarAI	90	834	11.1	9.3	4.2	5.3	8.9	70.02	6.24	0.48	23.38
2	MedAI	82	534	14	6.5	4.1	10.3	11.9	61.05	13.86	5.99	20.41
	LawAI	33	227	13.7	6.9	4.6	8.2	9.6	78.41	3.96	2.20	15.42
	CarAI	58	428	12.6	7.4	5	9.6	11.3	76.17	3.97	0.70	19.16
	RefAI	27	135	9.2	5.0	3.1	4.8	6.3	64.44	2.22	3.70	30.37
1 + 2	Total	378	2880	12.3	7.8	4.2	7.6	10	-	-	-	-

Table 5: Statistics on dialogs and user utterances across different topics in two studies. Arguments and non-arguments show the distribution of user utterance types across topics. Categories include well-founded (WF) and unfounded (UF) arguments, as well as questions (Q) and miscellaneous non-argumentative utterances (Misc). Values indicate the percentage occurrence of each type.

sented 4.2 unique arguments per dialogue. The number of distinct bot arguments per dialogue is with a value of 7.6 substantially higher than that of user arguments. This is because when the system successfully classifies a user's argument, it paraphrases the user's statement and presents its counterargument, effectively presenting two arguments for every user argument.

We also observed instances where user arguments didn't map to the argument graph. As outlined in section 5, we categorize textual units into *argumentative*, *non-argumentative*, *questions*, and *miscellaneous*. The right side of table 5 displays the distribution of these types. Most user utterances include well-founded arguments, indicating users engaged in substantial discussions and that our graph captures the majority of user arguments. This dataset's richness makes it ideal for studying structured argumentative dialogs in AI ethics. The variety of unfounded arguments and misc. phrases across topics indicate users brought diverse perspectives, enhancing the dataset's value for training models to recognize a wide spectrum of arguments and other types of user input.

Moreover, the dataset isn't merely a set of statements. The inclusion of questions and other non-argumentative expressions highlights the dialogues' interactive nature, making the dataset suitable for tasks beyond just argument recognition (further discussed in section 8).

The proportion of argumentative and non-argumentative phrases differs between the two studies, reflecting the evolving performance of our dialogue system. For the first study, we used preliminary data, but we enhanced our system using data from the first study for the second.

In conclusion, this dataset provides a detailed view of real-world discussions on AI ethics topics. Its structured format, combined with the diverse arguments and interactive elements, makes it a versatile tool for various NLP tasks.

8. Applications on the Dataset

This section demonstrates the potential applications of our dataset. We start by setting up a benchmark using the *OpenAI API*[1] and continue by describing other possible tasks that can make use of this dataset.

8.1. User Utterance Classification: A Preliminary Benchmark

Recognition of user arguments is a cornerstone of any argumentative dialogue system. Our benchmark focuses on this capability, aiming to map user utterances to specific nodes in the argument graph. Essentially, the task is to tag each utterance with an argument label or to label it as miscellaneous (misc). Accurate prediction requires

- Matching the predicted label to any argument within the utterance, or
- Predicting *misc* for utterances without graph arguments.

Using the *gpt-4-0613* model via the *OpenAI API*, we structured our text classification prompt. We included all nodes of the argument graph in the prompt, each entry combining a label with a summary text (see figure 2). The task of the model was to find the most matching summary for each user utterance and predict its label. If no match is found, the model is instructed to predict *misc*. We used the following prompt with the model:

> You are a text classifier. Your job is to predict for different user inputs to which of the following classes (format: class label; example) it is most similar. If you are convinced it is not similar to anyone

[1] https://platform.openai.com/docs/introduction

predict 'OTHER':
"""
Z.P1; MedAI will improve medical care.
Z.P2; A MedAI is standardized and reproducible.
...
"""
User messages will have the following format:
"""
1. <message1>
2. <message2>
"""

Make sure to only output labels and not the examples as well. Your output has to look like this:
"""
1. <Label>
2. <Label>
"""

Since GPT-4 does not have any knowledge of our dataset, this essentially becomes a one-shot classification challenge. Performance was measured using accuracy scores across the datasets, and GPT-4's results were compared to a majority baseline. Additionally, we measured accuracy over different subsets of the data, as well as precision and recall for mere recognition of argumentative phrases. The results are presented in table 6.

GPT-4 consistently outperforms the majority baseline across all topics. Nevertheless, an accuracy of around 0.5 suggests that there is much room for improvement, which we will explore below by addressing potential classification errors.

Multi-label Utterances. The performance on multi-label accuracy is either at or below average, which is to be expected given the complexity of categorizing multifaceted utterances under a single label.

Utterance Length. A clear trend emerges when examining the relationship between utterance length and model performance: Accuracy tends to decrease as utterance length increases. Short utterances, especially those labeled *misc*, are easier because they often consist of simple phrases of agreement, disagreement, or non-argumentative content. On the other hand, longer utterances, especially those with multiple labels, are more complex. In addition, the data prompted to the model could influence the performance. The distribution of utterance lengths differs between the data sets and the summary texts used for classification. The user utterances contain both longer and more lengthy texts compared to the summary texts.

Performance on Miscellaneous Phrases. While the accuracy of *misc* exceeds the overall rate, the model struggles with argumentative phrases that are not in the graph. Such phrases may be mistaken for existing arguments due to their similarity.

Lack of Contextual Awareness. The model evaluates utterances individually, without considering the context of the dialogue. Incorporating context could improve performance by allowing the model to exclude certain arguments based on previous utterances.

Benchmarking using the OpenAI API reveals the challenges of classifying user utterances in argumentative dialogues. Despite outperforming the majority baseline, the model has areas for improvement. Utterance length, multi-label complexity, and the absence of dialogue context affect the effectiveness of the model. Subsequent sections examine other potential research applications for this dataset concerning argumentative dialogue systems.

8.2. Potential Use Cases

Our dataset highlights the challenges of linking user comments to the argument graph. Its main strength lies in the dialogues between our system and users, making it a valuable resource for developing chatbots tailored to human discussions. Beyond argument classification, this dataset could facilitate further research on various aspects of German argument mining:

Identifying Argumentative Content. It's important to distinguish argumentative content from other types of content to enhance system responses and reduce errors. According to recall and precision scores in table 6, GPT-4 can distinguish between argumentative and non-argumentative content in most cases. However, there is still room for improvement which should be investigated in future studies.

Stance Classification. Knowing the stance of an argument can make classification more efficient. This can also help in understanding the main point of view of longer texts.

Segmentation of User Utterances. Given the multifaceted nature of user utterances, segmentation is essential. Segmenting utterances allows systems to provide more contextual responses by capturing the full range of user intentions.

Using Conversational Context. The dialogue-based format of the dataset emphasizes the importance of context. As AI models continue to improve, our dataset can help train them to better capture the full conversation.

In summary, our dataset has multiple uses that aim to advance both argument analysis and conversational AI.

Metric	Subset	Study 1			Study 2			
		MedAI	LawAI	CarAI	MedAI	LawAI	CarAI	RefAI
Accuracy	All (Majority Baseline)	0.32	0.28	0.3	0.27	0.12	0.11	0.3
	All (GPT-4)	**0.54**	0.46	0.52	0.50	0.51	0.51	0.51
	Arguments	0.50	0.42	0.46	**0.51**	0.46	0.49	0.46
	Misc.	0.62	0.57	0.65	0.48	**0.74**	0.56	0.63
	Multi-label utterances	0.45	0.14	**0.53**	0.51	0.40	0.32	0.40
	Shortest 25% of utterances	0.70	0.65	0.70	0.64	0.73	0.69	**0.76**
	Medium length utterances	**0.51**	0.47	0.47	0.43	0.45	0.46	0.46
	Longest 25% of utterances.	0.42	0.27	0.44	**0.48**	0.40	0.41	0.35
Recall	Arguments	0.91	0.92	0.95	0.95	0.94	0.94	**0.99**
	Misc.	0.62	0.57	0.65	0.48	**0.74**	0.56	0.63
Precision	Arguments	0.83	0.85	0.87	0.83	**0.94**	0.88	0.86
	Misc.	0.76	0.73	0.84	0.76	0.72	0.74	**0.96**

Table 6: All rows but the first one show classification results produced by GPT-4. For precision and recall, "arguments" indicates that the model successfully determined that it was an argument from the graph, regardless of its correctness.

9. Conclusion and Future Work

In this paper, we have presented a German dataset on AI ethics discussions, created through dialogues between users and a conversational agent. Our two-stage annotation process, using our created argument graphs, ensures the quality of the dataset. We discussed the challenges of annotation, in particular label inconsistencies and subjective interpretations. The applications of the dataset go beyond argument recognition. A first benchmark with the OpenAI API shows the complexity of mapping user utterances to an argument graph. The interactive and argumentative nature of the dataset positions it as a valuable resource for German argument mining and conversational AI research.

Acknowledging the limitations of the dataset, including potential biases and the limitations of the argument graph, we're planning to expand to more topics. We also aim to improve our annotation strategy and further explore contextual understanding for classification. With advances in large-scale models, we're optimistic about using these models to increase the utility of the dataset and provide more nuanced insights into argumentative dialogues on ethical issues.

10. Limitations

Annotating argumentative dialogues about AI ethics is inherently challenging. While we have detailed guidelines, the subjective interpretation of user utterances can lead to inconsistencies. The multifaceted nature of some dialogues makes it difficult to assign a single label or boundary, potentially missing nuances of the user's intent. In addition, our comprehensive guidelines can't anticipate every unique scenario, creating potential ambiguity.

Annotators, being human, bring their biases and are prone to error, especially when dealing with complex dialogues. While this human element is essential, it can introduce bias and error.

Beyond individual interpretations, the basic structure of our annotations, the argument graph, presents its own challenges. The completeness and accuracy of the graph directly affect the quality of the annotation. Omissions in the graph are reflected in the annotations. Furthermore, because the graph is constructed by a selected group of experts, it may inadvertently carry certain biases or perspectives that shape the annotation process. We chose the topics for the dialogues on the assumption that they would be of wider interest to the public debate.

It's important to acknowledge these limitations to truly appreciate the scope of the dataset and identify potential improvements.

11. Ethical Considerations

All study participants were asked at the beginning of the study to agree to the storage of dialogue data upon logging into the website. The study website and instructions make it explicitly clear that this is a conversation with a chatbot. The page explains to users that the bot's strategy is to always argue against their point of view. Users could stop the study at any time and request that their data be deleted after the study. No personal information was collected.

12. Acknowledgements

I would like to express my sincere gratitude to my professor, Frank Puppe, for his invaluable guidance and mentorship throughout this research project. I am deeply appreciative of the significant contributions made by my co-authors, Adrian Krenzer and Antonia Krause. I also extend my thanks to our Hiwis, Erik Spiegel and Marlon Berten, for their dedicated support. Finally, a special thank you to Julia Kiesel for her unwavering encouragement. This research would not have been possible without the generous funding provided by the Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research)

13. Bibliographical References

Annalena Aicher, Nadine Gerstenlauer, Isabel Feustel, Wolfgang Minker, and Stefan Ultes. 2022. Towards building a spoken dialogue system for argument exploration. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1234–1241.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2015a. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, Denver, CO. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2015b. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Yuri Cath. 2016. Reflective equilibrium. *The Oxford handbook of philosophical methodology*, pages 213–230.

Lisa Chalaguine and Anthony Hunter. 2021. Addressing popular concerns regarding covid-19 vaccination with natural language argumentation dialogues. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21–24, 2021, Proceedings 16*, pages 59–73. Springer.

Lisa Andreevna Chalaguine and Anthony Hunter. 2019. Knowledge acquisition and corpus for argumentation-based chatbots. In *CEUR Workshop Proceedings*, volume 2528, pages 1–14. CEUR Workshop Proceedings.

Lisa Andreevna Chalaguine, Anthony Hunter, Henry Potts, and Fiona Hamilton. 2019. Impact of argument type and concerns in argumentation with a chatbot. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1557–1562. IEEE.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Lorik Dumani, Manuel Biertz, Alex Witry, Anna-Katharina Ludwig, Mirko Lenz, Stefan Ollinger, Ralph Bergmann, and Ralf Schenkel. 2021. The recap corpus: A corpus of complex argument graphs on german education politics. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 248–255. IEEE.

Bettina Fazzinga, Andrea Galassi, and Paolo Torroni. 2021. An argumentative dialogue system for covid-19 vaccine information. In *International Conference on Logic and Argumentation*, pages 477–485. Springer.

Emmanuel Hadoux and Anthony Hunter. 2019. Comfort or safety? gathering and using the concerns of a participant for better persuasion. *Argument & Computation*, 10(2):113–147.

Emmanuel Hadoux, Anthony Hunter, and Jean-Baptiste Corrégé. 2018. Strategic dialogical argumentation using multi-criteria decision making with application to epistemic and emotional aspects of arguments. In *Foundations of Information and Knowledge Systems: 10th International Symposium, FoIKS 2018, Budapest, Hungary, May 14–18, 2018, Proceedings 10*, pages 207–224. Springer.

Emmanuel Hadoux, Anthony Hunter, and Sylwia Polberg. 2023. Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. *Argument & Computation*, 14(2):109–161.

Christian Hauptmann, Adrian Krenzer, Justin Völkel, and Frank Puppe. 2024. Argumentation effect of a chatbot for ethical discussions about autonomous ai scenarios. *Knowledge and Information Systems*, pages 1–31.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press.

LENZ Mirko, Premtim Sahitaj, Sean Kallenberg, Christopher Coors, Lorik Dumani, Ralf Schenkel, and Ralph Bergmann. 2020. Towards an argument mining pipeline transforming texts to argument graphs. In *Computational Models of Argument: Proceedings of COMMA*, volume 326, page 263.

Hugo Gonçalo Oliveira, Patrícia Ferreira, Daniel Martins, Catarina Silva, and Ana Alves. 2022. A brief survey of textual dialogue corpora. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1264–1274.

H Prakken et al. 2020. A persuasive chatbot using a crowd-sourced argument graph and concerns. *Computational Models of Argument*, 326:9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9048–9056.

14. Language Resource References

Rob Abbott and Brian Ecker and Pranav Anand and Marilyn Walker. 2016. *Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it*. European Language Resources Association (ELRA).

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. Dart: a dataset of arguments and their relations on twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Katri Leino, Juho Leinonen, Mittul Singh, Sami Virpioja, and Mikko Kurimo. 2020. Finchat: Corpus and evaluation setup for finnish chat conversations on everyday topics. *arXiv preprint arXiv:2008.08315*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Julia Romberg and Stefan Conrad. 2021. Citizen involvement in urban planning-how can municipalities be supported in evaluating public participation processes for mobility transitions? In *Proceedings of the 8th Workshop on Argument Mining*, pages 89–99.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2022. Argscichat: A dataset for argumentative dialogues on scientific papers. *arXiv preprint arXiv:2202.06690*.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A corpus for argumentative writing support in german. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 856–869.

Ruijian Xu, Chongyang Tao, Jiazhan Feng, Wei Wu, Rui Yan, and Dongyan Zhao. 2021. Response ranking with multi-types of deep interactive representations in retrieval-based dialogues. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–28.