ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240919

EthiX: A Dataset for Argument Scheme Classification in Ethical Debates

Elfia Bezou-Vrakatseli[a,*], **Oana Cocarascu**[a] **and Sanjay Modgil**[a]

[a]King's College London

Abstract. Argument schemes represent stereotypical patterns of reasoning that capture the inferences from premise(s) to conclusion. Despite their usefulness in argument mining, argument scheme classification remains a largely understudied task in NLP. In this paper, we present *EthiX*, a novel dataset for classifying argument schemes, comprising arguments spanning 22 ethical topics which are manually annotated with argument schemes following Walton's taxonomy. We evaluate pre-trained models fine-tuned on our dataset and propose a baseline to the community.

1 Introduction

Argument mining has become an established area within the natural language processing (NLP) field. Despite considerable advances in identifying argument components and the relations between them (see surveys [20, 22] for an overview), argument mining remains one of the most challenging areas in NLP. One task that has so far been understudied is argument scheme classification. Classifying argument schemes is an important step towards understanding the reasoning process behind arguments and can help strengthen their quality and scope.

Argument schemes and critical questions have traditionally been used in formal argumentation to support individual agent reasoning and joint deliberation (dialogue) [28]. Historically, their origin traces back to Aristotle's topics but they have also become pivotal in modern-day argumentation theory [12, 30, 44], leading to various taxonomies, e.g. the pragma-dialectical classification of Van Eemeren et al. [43], the argument schemes proposed by Walton et al. [51], and the periodic table of arguments [47]. Argument schemes provide structured templates for stereotypical forms of arguments that capture the inferences from premise(s) to conclusion, and can thus constitute a means for identifying arguments. While using argument schemes can be beneficial for classifying arguments and determining the validity of arguments and whether they are fallacious, the literature has seen limited focus on argument scheme classification.

Annotating argument schemes is particularly challenging [50]. This can be attributed to *1)* a lack of a universally accepted taxonomy in argumentation theory, and *2)* the need for highly trained expert annotators. Indeed, the cognitive load for annotating argument schemes is higher compared to other tasks in argument mining such as identifying argument components (e.g. claims, premises) and determining argument relations (e.g. support, attack) [29]. There are few datasets annotated using argument schemes (e.g. [8, 21, 11, 39]), and these tend to be based on Walton's taxonomy [51] as Walton's proposed schemes were introduced as a bottom-up approach to encapsulate human arguments. Some works simplify the task by using a small number of argument schemes (e.g. [16]) or defining types related to schemes (e.g. [15, 33]), whilst the larger corpora are automatically generated in order to reduce the annotation effort (e.g. [37, 38]). However, automatically generated datasets are a result of instantiations of logical formulas and pre-defined rules, and can thus dilute the complexity and contextual richness of human-generated arguments.

In this paper we introduce *EthiX*[1], a manually curated dataset for argument scheme classification that captures human reasoning in ethical debates. We extract arguments from Kialo[2], a platform for rational debate, with a moderation policy where arguments are backed by reasons and make constructive points, while also capturing the dialogical and interactive element of human argumentation. We annotate arguments using Walton's taxonomy of argument schemes [51], resulting in 686 ethical arguments categorised into eight classes (*argument from example*, *argument from values*, *argument from positive consequences*, *argument from cause to effect*, *argument from expert opinion*, *argument from negative consequences*, *argument from alternatives*, *argument from analogy*), spanning 22 topics. Moreover, we evaluate pre-trained models fine-tuned on our dataset and conduct experiments to assess how well models generalise to unseen topics as well as adapt to new topics. We propose a baseline to the community to stimulate further progress on the argument scheme classification task.

2 Background

We first provide core background on argument schemes, the Argument Scheme Key methodology, as well as enthymemes, which we use to annotate arguments.

2.1 Argument schemes

Argument schemes (AS) represent stereotypical patterns of reasoning, which, together with their critical questions (CQs), provide generic templates that can be instantiated by computational and natural languages to support human-human and human-AI debate and dialogue [28]. In particular, Walton et al. [51] proposed over 60 argument schemes with corresponding sets of CQs. For example, the

* Corresponding author. Email: elfia.bezou_vrakatseli@kcl.ac.uk

[1] Dataset available at: https://github.com/ElfiaBv/EthiX
[2] https://www.kialo.com

scheme *argument from example* is defined by a premise, conclusion, and several CQs, as follows:

Premise: *In this particular case, the individual a has property F and also property G.*

Conclusion: *Therefore, generally, if x has property F, then it also has property G.*

CQ1: Is the proposition claimed in the premise in fact true?

CQ2: Does the example cited support the generalisation it is supposed to be an instance of?

CQ3: Is the example typical of the kinds of cases the generalisation covers?

CQ4: How strong is the generalisation?

CQ5: Do special circumstances of the example impair its generalisability?

In this paper we use the argument schemes proposed by Walton et al. [51], whilst acknowledging that other schemes have been developed in the literature, e.g. the pragma-dialectical classification of Van Eemeren et al. [43] and the periodic table of arguments [47].

2.2 Argument Scheme Key (ASK)

The Argument Scheme Key (ASK) was proposed by Visser et al. [46] as an effective methodology for choosing the appropriate scheme from Walton's taxonomy [51]. It is an annotation heuristic consisting of a 'dichotomous identification key' where the steps/choices are a product of grouping together certain scheme types, based on shared characteristics. Table 1 shows the first ten steps from ASK.

1. Argument relies on a source's opinion or character ..... 2.
- Argument does not depend on a source's opinion or character 17.
2. Argument is about the source's character ..... 3.
- Argument is about the source's opinion ..... 9.
3. Argument establishes the source's character .....
..... *Argumentation from interaction of act and person*
-Argument refers to the source's existing character ..... 4.
4. Argument relies on the source's good character .....
..... *Ethotic argument*
- Argument relies on bad character ..... 5.
5. Source is biased ..... 6.
- Argument is not related to bias ..... 7.
6. Source does not take both sides into account .....
..... *Argument from bias*
- Source's opinion is not acceptable .....
..... *Bias ad hominem*
7 (5). Source is of bad overall character .....
..... *Generic ad hominem*
- Source's actions are not compatible with their commitments .. 8.
8. Source's actions contradict the advocated position .....
..... *Pragmatic inconsistency*
- Source is not credible due to inconsistent commitments .....
..... *Circumstantial ad hominem*
9 (2). Argument establishes a source's opinion ..... 10.
- Argument is based on an existing opinion ..... 11.
10. Commitment at issue is consistent with existing commitments
..... *Argument from commitment*
- Commitment at issue not consistent with existing commitments .
..... *Argument from inconsistent commitment*

Table 1. The first 10 steps from ASK.

2.3 Enthymemes

Natural language arguments rarely explicitly mention all the premises considered and the conclusion drawn from the claims expressed; in the majority of cases, they leave some premises, or often the conclusion, implicit. An argument that excludes missing components is called an *enthymeme* [48]. The following is an example of an enthymeme in the debate *Should the death penalty be abolished?*:

> A: "The death penalty negatively affects both the families of the victims as well as the defendant(s)."

The enthymeme A expresses only some justification (i.e. a premise) for the implicit unrepresented claim supporting abolition of the death penalty. The implicit "therefore the death penalty should be abolished" is elicited from the *context* (i.e. the debate topic) and left for the reader to infer.

3 Related work

There are few works in the literature that focus on classifying the argument schemes defined by Walton et al. [51]. Some use machine learning to classify schemes [8, 21], while others focus only on annotating texts using argument schemes [11, 39]. Feng and Hirst [8] used 393 arguments from Araucaria [32] annotated with five argument schemes (i.e. *argument from example*, *argument from cause to effect*, *practical reasoning*, *argument from consequences*, and *argument from verbal classification*), while Lawrence and Reed [21] used 78 arguments from AIFdb [23] annotated with two argument schemes (*argument from expert opinion* and *argument from positive consequences*). In both works, the authors relied on feature engineering to deploy supervised machine learning algorithms to classify argument schemes. Focusing solely on annotation, Hansen and Walton [11] used 14 argument schemes, in addition to a special class, *can't classify*, to annotate 256 political arguments whereas Schneider et al. [39] annotated Wikipedia deletion discussions, resulting in 555 arguments labelled with 17 argument schemes.

Some works have opted to address the complexity of numerous argument schemes by concentrating on a considerably small subset of schemes or defining types related to schemes. Jo et al. [16] developed an annotation protocol focusing on two schemes (*argument from consequences* and *practical reasoning*) to annotate 1000 arguments with normative claims randomly selected from Kialo, and used BERT [5] to classify the relations. Jo et al. [15] used four proposition types "related to argument schemes" (i.e. *normative*, *desire*, *future possibility*, *reported speech*) to annotate U.S. presidential debates. Kondo et al. [18] introduced a dataset comprising 2,370 pairs of argumentative text and corresponding Bayesian networks representing the reasoning structure of arguments, spanning six topics from ProCon (procon.org). They identified 17 'idioms' via crowdsourcing, covering 25 argument schemes. Their experiments show that the choice of a suitable idiom is a difficult task.

Some recent studies have tried to address the lack of annotated corpora via argument generation [38, 37]. Saha and Srihari [38] proposed a neural argument generator that uses scheme-based control codes derived from six of Walton's argument schemes (i.e. *means for goal*, *goal for means*, *from consequence*, *source knowledge*, *source authority*, *rule or principle*) to generate factual arguments, resulting in almost 70,000 arguments spanning six topics. However they show that there are cases in which the generated argument template is modified, and hence the meaning of the argument is changed. Ruiz-Dolz et al. [37] used GPT-3.5-TURBO and GPT-4 to gener-

ate arguments, instantiating 20 different argument schemes with two stances and spanning 50 topics, resulting in 1,893 arguments in English and 1,917 arguments in Spanish. The human validation process did not consider the soundness or strength of arguments — both important features of human argumentation — which are captured on the Kialo platform from which we collect our dataset. Moreover, the corpus solely consists of complete arguments and does not include enthymemes. This is indicative of the more general issue of artificially constructed arguments and their shortcomings regarding incorporating features typical of human argumentation, including common structures such as enthymemes.

To address the challenging task of manual annotation with argument schemes, several works developed effective annotation guidelines [24, 29, 42, 46, 49, 50]. Building on the *Argumentum Model of Topics* [34, 35], where argument schemes are organised in hierarchical clusters, Musi et al. [29] proposed a set of guidelines for annotating schemes and conducted their annotation study on 10 essays, achieving fair agreement. Song et al. [42] defined three general argument schemes (i.e. *policy*, *causal*, *sample*) to annotate 600 essays from graduate school admissions, whereas Reisert et al. [33] used argument templates inspired by Walton's argument schemes to annotate argumentative texts on policy topics. Visser et al. [46] annotated 491 arguments from televised election debates along with related Reddit posts following the Inference Anchoring Theory (IAT) [31]. Based on the evaluation of inter-annotator agreement, they proposed the Argument Scheme Key (ASK) (see Section 2.2) for annotating with Walton's schemes, which makes use of the indicative properties of particular argument schemes. Lawrence et al. [24] developed an annotation tool that combines OVA [14], an online software for analysing schemes in argumentative discourse, with the ASK, aiming to enrich datasets for studying argument schemes by improving the annotation process.

Some studies focus specifically on detecting fallacies, which have some overlap with argument schemes. Goffredo et al. [9, 10] used six categories, four of which are from Walton [52], to detect fallacies in U.S. presidential debates, whereas Ruiz-Dolz and Lawrence [36] focused on identifying four fallacies (i.e. *ad hominem*, *appeal to majority*, *appeal to authority*, *slippery slope*) and created a small dataset comprising 14 arguments that included seven different types of argument schemes. Other uses of argument schemes include their application in educational frameworks [27, 2] and discourse analysis [3]. Macagno and Konstantinidou [27] used two argument schemes (i.e. *argument from cause to effect*, *argument from analogy*) to reconstruct students' arguments, whilst Anthony and Kim [2] revealed difficulties in annotating dialogues with argument schemes in classroom deliberation, emphasising the ambiguity of the scheme definitions. Cabrio et al. [3] mapped five schemes (i.e. *argument from example*, *argument from cause to effect*, *argument from effect to cause*, *practical reasoning*, and *argument from inconsistency*) to discourse relation categories in the Penn Discourse TreeBank.

Lastly, Lindahl et al. [25] conducted an annotation exercise and used 30 argument schemes to annotate arguments in political text from Swedish newspaper editorials. They found that annotators differ in argument annotation (e.g. one annotator identifies double the number of argument schemes, the most prominent scheme identified by one annotator is the one least identified by the other, etc.) as well as in the identification of arguments and their components (i.e. premise and conclusion), concluding that there is a need for strict and detailed instructions to annotate using argument schemes.

4 Dataset

In this section, we describe the process for constructing *EthiX* and provide an overview of the dataset.

4.1 Dataset construction

Figure 1 gives an overview of the creation process of *EthiX*. Starting from Kialo, a platform designed to support rational debate, we select ethical debates and extract arguments based on their relevance and originality (see Section 4.1.1). Then, we identify the enthymemes and reconstruct them for annotating the arguments with Walton's argument schemes [51] using the ASK algorithm [46] (see Section 4.2), keeping the most prominent to be included in *EthiX*.

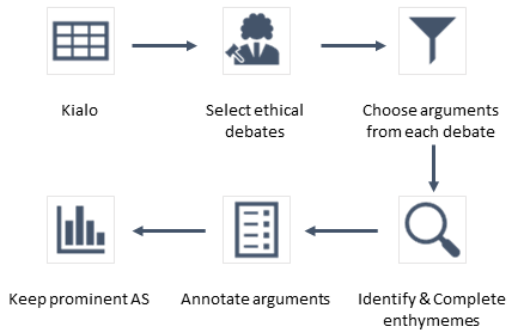

Figure 1. The creation process of *EthiX*.

Figure 2. Part of the discussion tree for the debate *Pro-life vs Pro-choice: Should abortion be legal?* from Kialo. The thesis statement is depicted at the top, in blue. Supporting/attacking arguments are depicted in green/red, branching out left/right. The impact score can be found in the bar on the top left corner of each argument.

4.1.1 Extracting arguments from ethical debates

We extracted arguments from Kialo, a collaborative platform consisting of debates, with concise arguments presented in a structured way. Figure 2 shows a debate from Kialo. Each debate addresses a certain question and has a thesis statement which is the original claim of the debate directly answering the debate's question. Users contribute to the debate by providing statements/arguments that support or attack an existing statement, with the discussion being represented as

Scheme	Representation	Sample
Argument from Example	P: In this particular case, a has properties F and G. C: So, generally, if x has property F, then it also has G.	*A*: In some places in the US, healthcare workers are required to get vaccinated against the seasonal flu. *D*: Should Covid-19 vaccines be mandatory?
Argument from Values	P1: Value V is positive/negative as judged by agent a. P2: If V is positive it is a reason to commit to goal G. P2': If V is negative it is a reason to not commit to goal G. C: V is a reason for agent a to (not) commit to goal G.	*A*: Even if the five people on the tracks have done heinous things, many believe it is illegitimate to kill them. *D*: What's the right solution to the trolley problem?
Argument from Pos. Conseq.	P: If A is brought about, good consequences will occur. C: Therefore, A should be brought about.	*A*: Access to legal abortion improves the health and safety of pregnant people. *D:* Pro-life vs Pro-choice: Should abortion be legal?
Argument from Cause to Effect	P1: Generally, if A occurs, then B will (might) occur. P2: In this case, A occurs (might occur). C: Therefore, in this case, B will (might) occur.	*A*: The cost of palliative care is extremely high, thus many people who opt for euthanasia are more likely to be poor. *D:* Should euthanasia be legalized?
Argument from Expert Opinion	P1: Source E is an expert in subject S. P2: E asserts that proposition A of subject S is true/false. C: A is true/false.	*A*: Masks reduce Covid transmission according to WHO. *D:* Do people have a right to not wear a mask in public spaces during the COVID-19 pandemic?
Argument from Neg. Conseq.	P: If A is brought about, good consequences will occur. C: Therefore, A should not be brought about.	*A*: Capitalism inevitably leads to various forms of exploitation. *D:* Is capitalism the most moral system?
Argument from Alternatives	P1: Either X or Y can be the case. P2: X is plausibly not the case. C: Y is plausibly the case.	*A*: Putting the child up for adoption is an alternative to abortion. *D:* Pro-life vs Pro-choice: Should abortion be legal?
Argument from Analogy	P1: Generally, case $C1$ is similar to case $C2$. P2: A is true (false) in case $C1$. C: A is true (false) in case $C2$	*A*: The right to reproductive freedom already includes assisted reproductive technologies such as in-vitro fertilization. Cloning humans could be seen as another assisted reproductive technology. *D:* Is cloning humans ethical?

Table 2. The 8 argument schemes, their representation in the form of premise (P) and conclusion (C), with example arguments (A) and debate (D) question.

a directed tree. Users can vote on the impact an argument has on the tree's parent claim, where impact is measured by the argument's veracity and relevance.

Several works have used Kialo for collecting and analysing arguments (e.g. [1, 6, 7, 16, 17, 40, 41, 45]), with the majority concentrating on all topics available on the platform. Our focus is on ethical reasoning. To this end, we analyse debates that address contemporary ethical questions of societal importance. We manually selected 22 debates addressing a variety of ethical topics, ensuring coverage of commonly discussed issues such as the death penalty and abortion, while also ensuring no overlap between the debates' topics (see Table 3).[3] Indeed, some Kialo debates are slight reformulations of others (e.g. *Should Aborting a Disabled Child Be Legal?* and *Should abortion of disabled fetuses be allowed for the reason that the fetus is disabled?* cover a similar topic). Similarly, some debates cover one argument that represents the central question of another debate (e.g. arguments from *Should Aborting a Disabled Child Be Legal?* can be found in the debate *Pro-life vs Pro-choice: Should abortion be legal?*).

A frequent phenomenon in discussions is that some statements tend to be over-debated. Thus, we manually analysed the arguments in the selected debates and chose a subset of them as follows. We include all arguments directly linked to the thesis statement, along with their sub-arguments (i.e. their supporting and attacking arguments). Then, we filter sub-sub-arguments based on the impact score: if it was higher than zero, the argument was included; if the impact score was zero, the annotator's estimation of the originality of the argument was used, mainly to deal with the common phenomenon of (almost) identical arguments.

4.2 Annotating argument schemes

We use the Argument Scheme Key (ASK) [46] (see Section 2.2) to annotate arguments with argument schemes. Enthymemes (see Section 2.3) play a significant part in the annotation process as annotators often have to fill in the missing parts in the arguments under analysis in order to identify the corresponding argument scheme. Often, the missing claim of the argument can be inferred from the *context* of the debate. Thus, we first determine whether an argument is an enthymeme, in which case we reconstruct the implicit premise/conclusion from the parent argument or the thesis statement from Kialo.[4] For example, consider the following argument from the debate *Pro-life vs Pro-choice: Should abortion be legal?*.

> A: "Access to legal abortion improves the health and safety of pregnant people."

A is an enthymeme whose implicit conclusion is assumed to be the thesis statement "pregnant people should have the right to choose abortion", which directly responds to the debate's central question "Should abortion be legal?". Thus, we consider the complete argument to be the following.

> A': "Access to legal abortion improves the health and safety of pregnant people, so pregnant people should have the right to choose abortion."

[3] We acknowledge that the debates may predominantly reflect concerns specific to certain parts of the (Western) world.

[4] Whilst we take this approach to reconstruct arguments, various interpretations may be suitable.

Although we use A' to annotate A, A is the argument we include in the dataset to avoid introducing biases. The ASK path corresponding to the steps taken in classifying argument A' is shown below, resulting in the identification of the scheme *argument from positive consequences*.

1. Argument relies on a source's opinion or character: *No*
2. Conclusion is about a course of action: *Yes*
3. Argument focuses on the outcome of the action: *Yes*
4. Conclusion promotes a positive outcome: *Yes*
5. Course of action assists someone else: *No*
6. Course of action promotes a goal: *Yes*

We identified 45 argument schemes out of the 60 defined by Walton [51]. However, nearly half of these had a frequency of less than 10. Our dataset contains the most prevalent argument schemes (i.e. with a frequency higher than 50), and includes 8 argument schemes for 686 pairs of arguments and the central question of the debates from which they were extracted.[5] Table 2 shows the 8 argument schemes and examples from our dataset.

4.3 Dataset statistics

Table 3 shows the number of arguments for each debate in our dataset. We computed the frequency, distribution, and average token length of arguments for each argument scheme in our dataset, which are summarised in Table 4.

Debates	Args
1. Should all drugs be legalized?	54
2. Pro-life vs Pro-choice: Should abortion be legal?	42
3. Should schools close during the Covid-19 pandemic?	40
4. Would the world be a better place without humans?	37
5. Should euthanasia be legalized?	35
6. Has social media been good for humanity?	34
7. Should all humans be vegans?	34
8. Should individuals sentenced to life in prison be allowed to choose death instead?	33
9. AI: Should an Artificial General Intelligence (AGI) be created?	32
10. Do people have a right to not wear a mask in public spaces during the COVID-19 pandemic?	32
11. Is it OK to incentivise moral behavior?	31
12. Should unpaid internships be banned?	31
13. Should Covid-19 vaccines be mandatory?	31
14. Is cloning humans ethical?	30
15. Is it ethically wrong to watch pornography?	30
16. What's the right solution to the trolley problem?	30
17. Is cannibalism ethically permissible?	27
18. Should the death penalty be abolished?	27
19. Do we have a moral duty to intervene in nature to limit animal suffering?	26
20. Is capitalism the most moral system?	18
21. Free Will or Determinism: Do we have free will?	17
22. Are moral properties natural properties?	15

Table 3. The ethical debates and the number of arguments in *EthiX*.

To ensure annotation quality, a second annotator — a PhD student with expertise in explainable AI — labelled arguments from three debates (debates 10, 13, and 17 from Table 3), amounting to approximately 13% of the total number of arguments. One of this paper's authors, who was involved in data annotation, provided the second annotator with a comprehensive introduction to Walton's taxonomy as well as examples of natural language arguments for each argument scheme, followed by a thorough explanation of the classification process (i.e. ASK) with further examples and a collaborative exercise involving the annotation of ten arguments. To evaluate the agreement, we calculated Cohen's Kappa [19] and obtained κ=0.523, which indicates moderate agreement, similar to other tasks in argument mining [22].

[5] We chose to include the debate question as opposed to the thesis statement. While the question is neutral, the thesis statment adopts a stance.

Scheme	Fr	Distr	Avg Tok
Arg from Example	120	0.17	25.2
Arg from Values	118	0.17	22.4
Arg from Pos. Conseq.	96	0.14	19.5
Arg from Cause to Effect	87	0.13	25.8
Arg from Expert Opinion	81	0.12	23.8
Arg from Neg. Conseq.	78	0.11	18.9
Arg from Alternatives	56	0.08	22.9
Arg from Analogy	50	0.07	22.5

Table 4. Frequency (Fr), distribution (Distr), and average token length (Avg Tok) for each class.

Two argument schemes had low agreement: *argument from cause to effect* and *argument from alternatives*. The low score for the former can be attributed to its corresponding path in ASK [46], which attempts to distinguish it based on the distinctive property of 'causality' (see Table 5), a feature that can be easily mistaken with the general structure of reasoning present in argument schemes (if-then statements), making it less detectable. We believe the low score for *argument from alternatives* stems from the fact that often, in day-to-day arguments, the alternatives compared are not necessarily mutually exclusive. Overall, we believe the disagreements are due to the inherent complexity of the annotation task, exacerbated by the interpretative skills required for identifying enthymemes.

47 (45). Argument relies on a causal relation .....48.
- Argument does not specifically rely on causality .....49.
48. Argument relies on a known causal relation .....
..... *Argument from cause to effect*
- Argument establishes a new causal relation .....
..... *Argument from correlation to cause*

Table 5. The step in ASK leading to *argument from cause to effect*.

5 Argument scheme classification in ethical debates

Determining argument schemes in ethical debates is a multi-class classification task with two inputs: the argument and the central question of the debate for which the argument was put forward. We evaluate pre-trained language models fine-tuned on our dataset.

5.1 Experiments

We experiment with six different models, BERT [5], RoBERTa [26], DeBERTa [13], ELECTRA [4], XLNet [53], ERNIE [54], as well as a random baseline where each test instance is randomly assigned from the eight classes with equal probability. We split the data into train (70%), test (20%) and validation (10%) sets, ensuring that each set

contains examples from each debate. We performed hyper-parameter search on the validation set and selected the best-performing combination from the following values: {8, 16, 32} for batch size, {3,4,5} × 10^{-5} for learning rate, and {1, ..., 50} for training epochs. We report the precision, recall and average macro F_1.

Furthermore, we conduct cross-topic experiments to evaluate how well the models generalise to an unseen topic debate. The experiment was conducted 22 times, each time training on 21 debates and testing on the remaining, in order to evaluate the models' robustness. Moreover, we also evaluate the performance of the models on a topic when limited examples for that topic are present in the training data.

5.2 Results & Discussion

Which model performs the best? Table 6 shows the best performance of each model on the test set, with the best scores per metric highlighted in bold. ERNIE yields the highest macro F_1 of 0.63, while the other models achieve a similar performance, with F_1 equal to or below 0.55.

Model	Precision	Recall	$\mathbf{F_1}$
Random baseline	0.09	0.08	0.08
BERT	0.54	0.52	0.53
RoBERTa	0.56	0.54	0.55
DeBERTa	0.57	0.54	0.54
ELECTRA	0.55	0.52	0.53
XLNet	0.60	0.56	0.55
ERNIE	**0.65**	**0.62**	**0.63**

Table 6. Results for the multi-class classification task on the test set.

Which argumentation schemes pose a challenge to the models? The most misclassified scheme in the first experiment was *argument from expert opinion*, which was mostly classified as *argument from example* and *argument from positive consequences*. The fact that the *argument from expert opinion* was found to be the most misclassified can serve as an indication of the distinctive properties of source-based arguments, i.e. the arguments that correspond to step 2 from ASK (see Table 1), also called 'second-order' [47]. They do not relate *directly* to the original standpoint (they are epistemological in nature) and they can be seen as having 'first-order' arguments embedded. An *argument from expert opinion* incorporates an unexpressed premise that an argument *has been uttered from* a person with expertise and that is generally indication of it *being true and acceptable*. Consider the following example:

Example 1. *Argument taken from Do people have a right to not wear a mask in public spaces during the COVID-19 pandemic?*

"We should wear masks in public places because WHO says so."

This example can be interpreted as having the format 'We should do something because it is good', indirectly communicated by a source of expertise. Thus, the more complex, second-order *argument from expert opinion* can be perceived to embed the format of a first-order scheme, *argument from positive consequences*.

How well do models generalise to an unseen topic? We report the best performance of each model for the cross-topic experiments in Table 7. ERNIE yields the highest macro F_1, while BERT and DeBERTa achieve the lowest recall.

The most misclassified scheme was *argument from cause to effect*, which was classified as an *argument from example* or an *argument from analogy*. Indeed, *argument from cause to effect* was also the scheme that scored the lowest in inter-annotator agreement. As explained in Section 4.3, we believe this is because of the nature of the scheme since its distinctive feature (i.e. causality) can be identified in other schemes as well, while its corresponding path in ASK does not succeed in efficiently distinguishing it. The following is an example of a *cause to effect* argument misclassified as an *argument from example*. The two argument schemes share many common features, also reflected in the ASK.

Model	Precision	Recall	$\mathbf{F_1}$
Random baseline	0.12 ± 0.18	0.11 ± 0.19	0.10 ± 0.14
BERT	0.46 ± 0.10	0.40 ± 0.07	0.47 ± 0.08
RoBERTa	0.58 ± 0.09	0.51 ± 0.07	0.51 ± 0.08
DeBERTa	$\mathbf{0.64 \pm 0.22}$	0.41 ± 0.20	0.44 ± 0.19
ELECTRA	0.59 ± 0.09	0.52 ± 0.07	0.51 ± 0.07
XLNet	0.62 ± 0.13	0.50 ± 0.11	0.50 ± 0.09
ERNIE	0.59 ± 0.12	$\mathbf{0.52 \pm 0.08}$	$\mathbf{0.52 \pm 0.09}$

Table 7. Results for cross-topic experiments.

Example 2. *Argument from Should individuals sentenced to life in prison be allowed to choose death instead?*

"Since many prisoners suffer from mental health conditions, a prisoner who desires euthanasia as a product of poor mental health cannot be considered to be making a voluntary decision."

A similar phenomenon was observed for *argument from analogy* and *argument from example*, which are closely related in the ASK algorithm (the differentiation between the two relies on only one step). Often one would be misclassified (by both models and annotators) as the other, and vice versa. An example can be found below.

Example 3. *Argument from Should individuals sentenced to life in prison be allowed to choose death instead?*

"Medical practitioners follow scientific standards to determine whether patients possess autonomy and the capacity to make an informed choice. Such requirements could also be applied to prisoners in these circumstances."

How do models adapt to new topics? In addition to the cross-topic experiments, we evaluated the performance of the models on a topic when limited examples for that topic are present in the training data. In our train-test splits for the first experiment, we ensured that each set contains examples from each debate. In this experiment, we used different percentages of randomly sampled data from the training data of Debate 1 *Should all drugs be legalized?* as it has the largest number of examples (see Table 3). Figure 3 shows the models' performance on the test set according to the amount of data from Debate 1 in the training set. Whilst the precision and recall of most models improves when adding debate topic examples to the training data, the recall for XLNet decreases with the addition of further examples.

What are the challenges in annotating argument schemes? One of the main challenges arises from interpretation and processing of enthymemes. During annotation, more than half (52%) of claims consisted of a single statement with the conclusion left implicit (as can also be observed in Table 2). This often meant that the dialectical impact of an argument was less evident, given that it is typically an argument's conclusion that identifies the target being attacked. The CQs associated with the schemes were crucial for enthymeme completion.

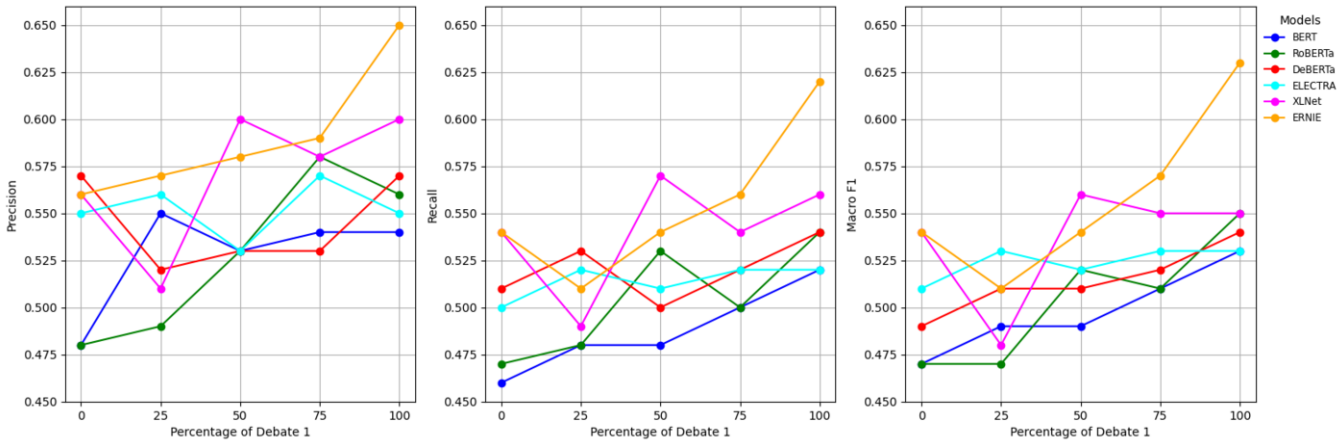

Figure 3. Performance of models on the test set according to the amount of data from Debate 1 (*Should all drugs be legalized?*) in the train set.

Additionally, a major hurdle in argument classification is the fact that an argument can often (seem to) fit multiple argument schemes. Consider the following argument.

Example 4. *Argument taken from Pro-life vs Pro-choice: Should abortion be legal?*

A: "The war on drugs is a good example of why prohibition does not work. Criminalizing behavior may not deter that behavior as long as incentives exist and may only compound the problem without addressing the underlying cause."

Argument A can be seen as both an *argument from example* and an *argument from analogy*. In A, a clear analogy is made between drugs and abortion and how prohibiting does not stop either. However, there is also an inductive generalisation drawn from a specific historical fact which can be interpreted as an *argument from example*. This issue emerges also in cases where certain schemes are sub-schemes of others, such as in the case of *argument from negative consequences* and *argument from danger*, since danger is a special case of negative consequences.

6 Conclusion and future work

We presented *EthiX*, a novel dataset of 686 arguments labelled with eight schemes defined by Walton et al. [51], and spanning 22 topics, specifically designed for the classification of argument schemes in ethical debates. By incorporating user-generated arguments from a debate platform, *EthiX* offers a structured approach to understanding and classifying forms of reasoning that are commonly employed in argumentative discourse. Specifically, our work aims to provide a resource that captures the complexity of human reasoning across a diverse range of ethical topics. Additionally, we proposed a baseline to stimulate future research in this area. Despite the inherent challenges of argument scheme classification, our results demonstrate that the task, while complex, is feasible with carefully curated data and appropriate computational models.

There are several avenues for future work. We plan to integrate critical questions so as to semantically enrich argument classification; in particular the semantic relationships amongst arguments. CQs identify the ways in which premises and inferential steps in argument schemes can be challenged or supported. They can then potentially be used to annotate attacking and supporting relationships between arguments, therefore enabling mining not just of arguments per.se., but also the dialectical relationships between arguments.

Furthermore, we are currently developing schemes and critical questions (AS&CQs) specialised for ethical reasoning, based on well studied philosophical ethical theories. Our long-term objective is to use AS&CQs to scaffold both individual agent ethical reasoning, and reasoning as conducted through dialogue and debate amongst human agents, and human and AI agents [28]. This entails accounting not only for how humans *descriptively* engage in ethical reasoning (the primary focus of this paper), but providing *prescriptive* guidance for how agents should ideally reason about ethical issues. Note that the schemes identified in our dataset do not all necessarily correspond to one ethical theory. However, there are alignments: arguments *from positive/negative consequences* align with consequentialism, arguments *from values* align more with virtue and duty/deontological ethics (schemes justifying actions can also be interpreted as deontological rules), whilst arguments *from example* could be used by both theories. We aim to develop AS&CQs that directly draw on these various ethical theories and their nuanced refinements; for instance, AS&CQs that accommodate varieties of consequentialism (with a focus on utilitarianism), deontology, and virtue ethics, and the nuanced ways in which arguments drawing on these theories are challenged. For example, *ad hominem* arguments[6] could be framed as virtue ethics based challenges to consequentialist arguments (e.g. individuals advocating actions on consequentialist grounds are sometimes challenged given that in so doing, they reveal some deficit in a virtuous character trait).

In respect of the above long term objectives, we are also investigating how AS&CQs can be utilised by large language models (LLMs) to structure their outputs. Indeed, initial results show that LLMs can generate both concrete individual arguments, and, when appropriately prompted, dialogues constituted by multiple arguments, and moreover recognise the schemes these arguments instantiate. The anticipated use of LLMs to support decision making will require human input if decisions are to be aligned with human values and ethically salient preferences. The use of AS&CQs to scaffold dialogues for joint human-AI decision making is a potentially promising approach to support value alignment.

[6] *Ad hominem* arguments (based on criticisms of character) are commonly encountered in everyday discourse and debate. They are underrepresented in our dataset, most likely because the Kialo platform emphasises the contents of claims and arguments over individuals.

Acknowledgements

We thank Hana Kopecká for annotating a subset of the data. This work was supported by the UK Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [grant number EP/S023356/1].

References

[1] K. Al Khatib, L. Trautner, H. Wachsmuth, Y. Hou, and B. Stein. Employing argumentation knowledge graphs for neural argument generation. In *ACL/IJCNLP*, pages 4744–4754, 2021.

[2] R. Anthony and M. Kim. Challenges and remedies for identifying and classifying argumentation schemes. *Argumentation*, 29:81–113, 2015.

[3] E. Cabrio, S. Tonelli, and S. Villata. From discourse analysis to argumentation schemes and back: Relations and differences. In *Computational Logic in Multi-Agent Systems*, pages 1–17, 2013.

[4] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[6] E. Durmus, F. Ladhak, and C. Cardie. Determining relative argument specificity and stance for complex argumentative structures. In *ACL*, pages 4630–4641, 2019.

[7] E. Durmus, F. Ladhak, and C. Cardie. The role of pragmatic and discourse context in determining argument impact. In *EMNLP-IJCNLP*, pages 5668–5678, 2019.

[8] V. W. Feng and G. Hirst. Classifying arguments by scheme. In *ACL*, pages 987–996, 2011.

[9] P. Goffredo, S. Haddadan, V. Vorakitphan, E. Cabrio, and S. Villata. Fallacious argument classification in political debates. In *IJCAI*, pages 4143–4149, 2022.

[10] P. Goffredo, M. Espinoza, S. Villata, and E. Cabrio. Argument-based detection and classification of fallacies in political debates. In *EMNLP*, pages 11101–11112, 2023.

[11] H. V. Hansen and D. N. Walton. Argument kinds and argument roles in the ontario provincial election, 2011. *Journal of Argumentation in Context*, 2(2):226–258, 2013.

[12] A. C. Hastings. *A Reformulation of the Modes of Reasoning in Argumentation*. Northwestern University, 1962.

[13] P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*, 2020.

[14] M. Janier, J. Lawrence, and C. Reed. OVA+: an argument analysis interface. In *COMMA*, volume 266, pages 463–464, 2014.

[15] Y. Jo, E. Mayfield, C. Reed, and E. Hovy. Machine-aided annotation for fine-grained proposition types in argumentation. In *LREC*, pages 1008–1018, 2020.

[16] Y. Jo, S. Bang, C. Reed, and E. Hovy. Classifying argumentative relations using logical mechanisms and argumentation schemes. *TACL*, 9: 721–739, 2021.

[17] I. Jundi, N. Falk, E. M. Vecchi, and G. Lapesa. Node placement in argument maps: Modeling unidirectional relations in high & low-resource scenarios. In *ACL*, pages 5854–5876, 2023.

[18] T. Kondo, K. Washio, K. Hayashi, and Y. Miyao. Bayesian argumentation-scheme networks: A probabilistic model of argument validity facilitated by argumentation schemes. In *The 8th Workshop on Argument Mining*, pages 112–124, 2021.

[19] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[20] A. Lauscher, H. Wachsmuth, I. Gurevych, and G. Glavaš. Scientia potentia Est—On the role of knowledge in computational argumentation. *TACL*, 10:1392–1422, 2022.

[21] J. Lawrence and C. Reed. Combining argument mining techniques. In *The 2nd Workshop on Argumentation Mining*, pages 127–136, 2015.

[22] J. Lawrence and C. Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, 2020.

[23] J. Lawrence, F. Bex, C. Reed, and M. Snaith. AIFdb: Infrastructure for the argument web. In *COMMA*, pages 515–516. IOS Press, 2012.

[24] J. Lawrence, J. Visser, and C. Reed. An online annotation assistant for argument schemes. In *The 13th Linguistic Annotation Workshop*, pages 100–107, 2019.

[25] A. Lindahl, L. Borin, and J. Rouces. Towards assessing argumentation annotation-a first step. In *The 6th Workshop on Argument Mining*, pages 177–186, 2019.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692, 2019.

[27] F. Macagno and A. Konstantinidou. What students' arguments can tell us: Using argumentation schemes in science education. *Argumentation*, 27:225–243, 2013.

[28] S. Modgil. Dialogical scaffolding for human and artificial agent reasoning. In *AIC*, pages 58–71, 2017.

[29] E. Musi, D. Ghosh, and S. Muresan. Towards feasible guidelines for the annotation of argument schemes. In *The 3rd Workshop on Argument Mining*, pages 82–93, 2016.

[30] C. Perelman and L. Olbrechts-Tyteca. *The new rhetoric: a treatise on argumentation*. 1969.

[31] C. Reed and K. Budzynska. How dialogues create arguments. In *The 7th Conference of the International Society for the Study of Argumentation (ISSA)*, pages 1633–1645, 2011.

[32] C. Reed and G. Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979, 2004.

[33] P. Reisert, N. Inoue, T. Kuribayashi, and K. Inui. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *The 5th Workshop on Argument Mining*, pages 79–89, 2018.

[34] E. Rigotti. Whether and how classical topics can be revived within contemporary argumentation theory. In *Pondering on problems of argumentation*, pages 157–178. Springer, 2009.

[35] E. Rigotti and S. Greco Morasso. Comparing the argumentum model of topics to other contemporary approaches to argument schemes: The procedural and material components. *Argumentation*, 24:489–512, 2010.

[36] R. Ruiz-Dolz and J. Lawrence. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *The 10th Workshop on Argument Mining*, 2023.

[37] R. Ruiz-Dolz, J. Taverner, J. Lawrence, and C. Reed. NLAS-multi: A multilingual corpus of automatically generated natural language argumentation schemes. *CoRR*, abs/2402.14458, 2024.

[38] S. Saha and R. K. Srihari. ArgU: A controllable factual argument generator. In *ACL*, pages 8373–8388, 2023.

[39] J. Schneider, K. Samp, A. Passant, and S. Decker. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Computer Supported Cooperative Work (CSCW)*, pages 1069–1080, 2013.

[40] G. Skitalinskaya, J. Klaff, and H. Wachsmuth. Learning from revisions: Quality assessment of claims in argumentation at scale. In *EACL*, pages 1718–1729, 2021.

[41] G. Skitalinskaya, M. Spliethöver, and H. Wachsmuth. Claim optimization in computational argumentation. In *INLG*, pages 134–152, 2023.

[42] Y. Song, M. Heilman, B. B. Klebanov, and P. Deane. Applying argumentation schemes for essay scoring. In *The 1st Workshop on Argumentation Mining*, pages 69–78, 2014.

[43] F. Van Eemeren, R. Grootendorst, and F. H. van Eemeren. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press, 2004.

[44] F. H. van Eemeren, R. Grootendorst, and T. Kruiger. Argumentatietheorie [argumentation theory]. *Utrecht: Het Spectrum.*, 1978.

[45] V. Varadarajan, N. Soni, W. Wang, C. Luhmann, H. A. Schwartz, and N. Inoue. Detecting dissonant stance in social media: The role of topic exposure. In *The 5th Workshop on Natural Language Processing and Computational Social Science*, pages 151–156, 2022.

[46] J. Visser, J. Lawrence, C. Reed, J. Wagemans, and D. Walton. Annotating argument schemes. *Argumentation*, 35(1):101–139, 2021.

[47] J. Wagemans. Constructing a periodic table of arguments. In *Argumentation, objectivity, and bias: The 11th international conference of the Ontario Society for the Study of Argumentation*, pages 1–12, 2016.

[48] D. Walton. Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*, pages 1–22. Springer, 2009.

[49] D. Walton. Using argumentation schemes for argument extraction: A bottom-up method. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 6(3):33–61, 2012.

[50] D. Walton and F. Macagno. A classification system for argumentation schemes. *Argument & Computation*, 6(3):219–245, 2015.

[51] D. Walton, C. Reed, and F. Macagno. *Argumentation schemes*. Cambridge University Press, 2008.

[52] D. N. Walton. *Informal fallacies*. 1987.

[53] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. XLNet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019.

[54] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. Ernie: Enhanced language representation with informative entities. In *ACL*, pages 1441–1451, 2019.