Templates for Fallacious Arguments Towards Deeper Logical Error Comprehension

Irfan Robbani[1] Paul Reisert[2] Naoya Inoue[1,3] Surawat Pothong[1]
Camélia Guerraoui[4,3,5] Wenzhi Wang[4,3] Shoichi Naito[6,4,3] Jungmin Choi[3] Kentaro Inui[7,4,3]
[1]JAIST [2]Beyond Reason [3]RIKEN [4]Tohoku University
[5]INSA Lyon [6]Ricoh Company, Ltd. [7]MBZUAI
{robbaniirfan,naoya-i,spothong}@jaist.ac.jp beyond.reason.sp@gmail.com
{guerraoui.camelia.kenza.q4, wang.wenzhi.r7, naito.shoichi.t1}@dc.tohoku.ac.jp
jungmin.choi@riken.jp kentaro.inui@mbzuai.ac.ae

Abstract

Fallacious arguments often lead to misinformation. Previous studies have primarily focused on creating benchmarks for fallacy detection, neglecting the need to logically explain why a fallacy is committed. To address this issue, we propose 20 templates for annotating the reasons behind a fallacy for 5 types of informal logical fallacies. Our templates are designed to capture an underlying structure of fallacies by making implicit assumptions explicit. Our preliminary annotation study using the LOGIC dataset [1] shows substantial inter-annotator agreement, and obtaining a coverage score more than 74%, indicating the feasibility of our templates.

1 Introduction

An argument is deemed invalid if its conclusion does not necessarily follow from its premises. This study concentrates on *informal fallacies*, where such invalidity arises from the content of the argument rather than its structure. Suppose the following argument:

- *I took an NLP class, an advanced course in Stanford. I suggest not taking further advanced courses because they will hurt your GPA.*

This argument is classified as *faulty generalization* fallacy because it generalizes into not taking further advanced courses based on taking an NLP class. Exercising such fallacious arguments leads to manipulation and the obscuration of truth, making it necessary to detect and explain why these arguments are fallacious [2].

Previous studies have focused on creating benchmark datasets for studying fallacious arguments, including notable works like ElecDeb60To16-fallacy [3], Argotario [4], and LOGIC [1]. The LOGIC dataset, compiled from various student quiz websites, serves an educational purpose, aiding students in understanding fallacies. Its accessibility and simplicity make it an effective tool in educational settings that emphasize fallacy learning. However, a significant limitation of these datasets is their lack of detailed explanations for the fallacies they contain, which impedes the development of critical thinking skills, particularly in the realm of education.

Furthermore, the use of argument templates, commonly applied for mapping annotation schemes, has not been extensively explored in the context of fallacious arguments. Various argumentation corpus, such as [5], which annotated the Araucaria corpus using [6] argumentation scheme, and [7], which developed a framework for instantiating argumentation schemes through natural language templates, have not calculated inter-annotator agreement. Additionally, [8] employed templates inspired by argument from consequence to annotate the arg-microtexts corpus [9], highlighting the underlying reasoning in arguments. Despite these advancements, the application of templates in the fallacious domain remains unexplored.

To bridge this gap, we have developed a new template for explaining fallacious arguments, drawing inspiration from [8]. This template-based approach is designed to provide a deeper understanding of the reasoning behind fallacious arguments. Our goal is to enhance critical thinking skills by offering more nuanced and comprehensive explanations of fallacious arguments.

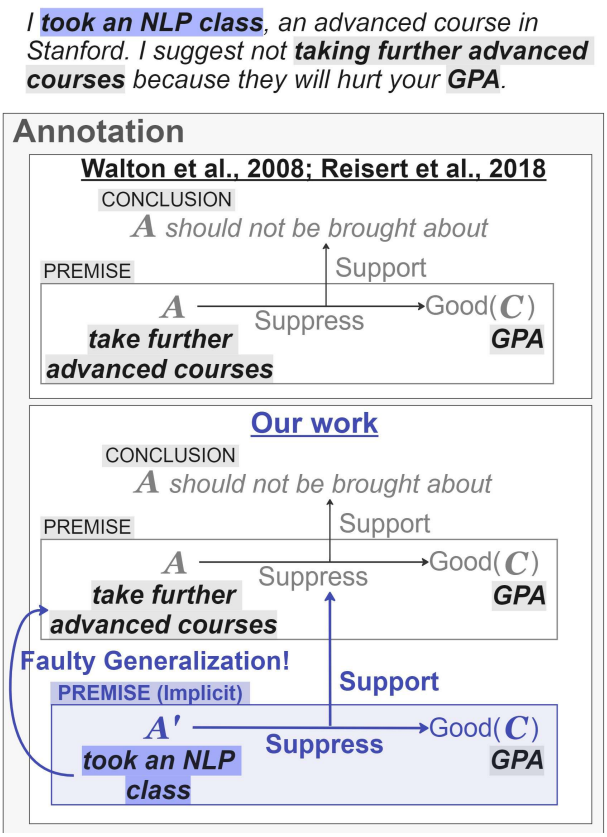

Figure 1 An overview of our fallacious argument instantiation task. Our template elucidates the example as *faulty generalization* fallacy because taking an NLP class is used to support the premise of taking further advanced courses. It explicates the implicit part of taking an NLP class because it contains a hidden message of hurting a GPA. Thus, the explicit part, taking further advanced courses, directly suppresses GPA which is simple to be noticed.

Template-based approaches have proven effective in explicating implicit knowledge within arguments, offering ease of control and consistency in annotation [10, 11, 12]. Since implicit knowledge often contributes to vagueness and assumption-making in arguments, making it explicit through templates enhances understanding because explicit knowledge is straightforward. This is particularly beneficial in identifying fallacies, which can be challenging due to their implicit components [13].

An overview of our task is shown in Fig. 1. Inspired by argumentation schemes [6] and argument templates [8], we create an inventory of fallacy templates consisting of slot-fillers and argumentative components based on the most common argumentation scheme type (i.e., Argument from Consequences). As many fallacy types exist in the wild [14], we select 5 informal logical fallacy (i.e., fallacy of defective induction [15]) as a start for exploring creating a fallacy template.

To evaluate the feasibility of our newly proposed fallacy templates, we first conduct a small trial annotation using our templates on an existing dataset of labeled fallacious arguments (i.e., LOGIC dataset [1]). We employ two experts in argumentation, both authors of this paper, to conduct the trial annotation on 250 instances from LOGIC dataset spanned across 5 different fallacy types (i.e., *faulty generalization*, *false dilemma*, *fallacy of credibility*, *fallacy of logic*, *false causality*). We find that our fallacy templates have a notable coverage and our annotation task has significant results on coverage and Cohen's kappa [16].

2 Creating Fallacy Templates

2.1 Desired Criteria

As an initial step in instantiating a fallacious argument into a template, we begin by defining the desired criteria:

Informative. Relevant information regarding fallacy within the fallacious argument is occasionally concealed and not presented concisely. Having sufficient information extracted from the template reveals high-quality results. Therefore, the templates offer a promising approach for constructing a dataset that is both simple and concise while maintaining high quality.

Easy-to-annotate. As for the annotating fallacy, the task is challenging, particularly given prior research [3, 1] indicating that a fallacious argument may be assigned multiple labels, highlighting the inherent ambiguity in identifying fallacies. The template designed by [8] is considered an easy yet simple predefined template by utilizing the slot-filler approach. We desire the template by slot-filler approach for instantiating the fallacious argument to ensure ease in the annotation process.

Critical Questions. The Argumentation scheme consists of various critical questions to define a strong argument or fallacious argument [17]. Thus, the template needs to contain a structured pattern of critical questions to identify the fallacious argument. The long-term utilization of this critical question is described in §4.

2.2 Our Fallacy Template Inventory

We created 20 fallacy templates for 5 fallacy types, 4 templates for each fallacy type. The fallacy types are related to the fallacy of defective induction group due to the focus on logical structure, including *faulty generalization*,

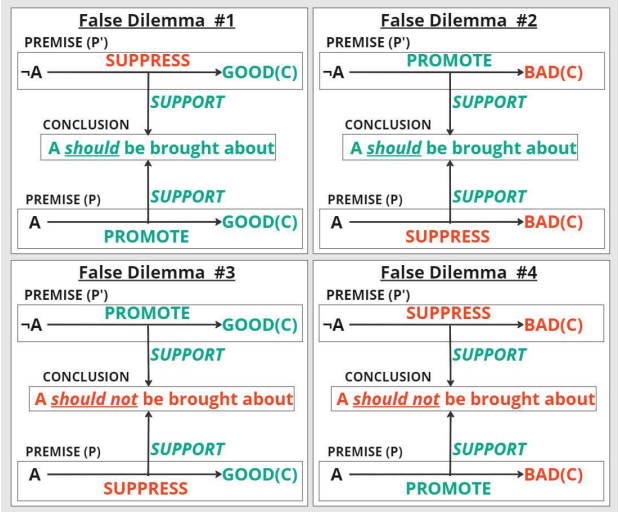

Figure 2 The 4 fallacy templates for *false dilemma* fallacious arguments.

false dilemma, *false causality*, *fallacy of credibility*, and *fallacy of logic*.

Following [8], we focus on argument from consequences [6]. The scheme is as follows:

- **Premise**: If A is brought about, good (bad) consequences C will plausibly occur.
- **Conclusion**: Therefore, A should (not) be brought about.

Where **A** and **C** are both slot-fillers which should be filled with either event or entity from the argument. The terms **good** and **bad** refer to the intention of the argument. **Promote** and **Suppress** refer to the relation between slot-filler. **Promote** refers to trigger the consequence and **Suppress** refers to prevent the consequence [18].

Fig. 2 shows our 4 templates for the *false dilemma* fallacy template. Two templates utilize argument from positive consequences while the other two templates utilize argument from negative consequences, where both schemes are the form of argument from consequences scheme.

Argument from Consequences has a total of 3 critical questions. However, we only adopted 2 critical questions since our focus is detecting fallacies rather than affirming the argument (details in §A.1).

We proposed a new set of supporting evidence to answer the critical questions. The usage of supporting evidence is to support the premise and every template proposes a different set of supporting evidence. The pattern of the supporting evidence is created based on the logical form of each fallacy type.

3 Annotation Study

The fallacious argument template instantiation requires a fallacious argument dataset consisting of 5 selected fallacy types. Moreover, to evaluate the feasibility of the proposed template, we utilize the LOGIC dataset, which encompasses 13 fallacy types, including those we have selected.

3.1 Template Instantiation

The LOGIC dataset is utilized, comprising 2,449 arguments. Upon reviewing the distribution of datasets, it becomes evident that the dataset is not equally distributed, reflecting a challenge in which the statistical instances are dominated by *faulty generalization* class with 18.01%.

We selected 250 samples across 5 fallacy types from the development set and random sampling from the training set for guideline construction and development set annotation.

3.2 Annotation Guidelines

Towards constructing guidelines, we first had two annotators, both authors of this paper, independently annotate the development set without guidelines, discussing results with each other and taking notes along the way. The annotation was spread into multiple rounds, and the annotator agreement (Cohen's kappa) was calculated for each round to ensure coherence and consistency were improving along the way. After all instances were annotated, annotators aggregated all notes and devised a new set of guidelines for carrying out the development stage. The guidelines and annotation of our experiment can be publicly accessed at https://github.com/irfanrob/fallacy-template.

3.3 Annotating Development Set

After a revised set of guidelines was created, both annotators independently annotated all 250 instances. Due to labelled instances, we annotate without identifying the fallacy type. The identification of the fallacy type is out of scope in this study. We report the results and discuss interesting findings and challenges along the way.

3.4 Results and Analysis

We report the Cohen's kappa between both annotators for the template selection for all 250 instances. Even with an insignificant result for *faulty generalization* type, our

Table 1 Agreement scores calculated across all fallacy types for 250 instances (All) and 209 instances (Filtered). We observe in the case of *faulty generalization*, annotators have a low agreement unless filtering out low confident instances. Furthermore, in the case of a *false dilemma*, filtering 3 instances reduces the kappa score enormously

Fallacy Type	Template Exists (Yes/No) All (κ)	Template Exists (Yes/No) Filtered (κ)	Template Selection (1-4) All (κ)	Template Selection (1-4) Filtered (κ)
False Dilemma	0.634	0.484	0.501	0.462
Faulty Generalization	0.276	0.530	0.383	0.548
False Causality	0.790	0.788	0.740	0.831
Fallacy of Credibility	0.675	0.760	0.817	0.944
Fallacy of Logic	0.688	0.769	0.839	0.944
Total	0.730	0.852	0.645	0.731

Table 2 Coverage results for instances that can be instantiated into the template from both annotators for 250 instances (All) and 209 instances (Filtered) from LOGIC dev set.

Annotator	All Coverage (%)	Filtered Coverage (%)
Annotator 1	74.8	76.0
Annotator 2	76.6	77.0

result is 0.645 which indicates a good agreement level.

During annotation, there are various instance in which annotators has uncertainty in the annotation process. So, both annotators report the confidence score for 23 instances where 18 instances are different between annotators. To improve the agreement score, we excluded uncertain instances, resulting in a slight increased in Cohen's kappa score to 0.731.

We also report the Cohen's kappa between both annotators for template coverage for both all 250 instances and excluded instances. Our result is 0.730 for all instances, and 0.852 for excluded instances which indicates an excellent agreement level. Furthermore, The coverage percentage from both annotators is considerable and some fallacy types have significant improvement results namely, *faulty generalization* by 0.254, *fallacy of credibility* by 0.085, and *fallacy of logic* by 0.081. This indicates that the templates can represent several fallacy types for the LOGIC dataset.

3.5 Discussion

Benefits of annotation We found the benefit of annotation is the ability to explicate the implicit premise behind the fallacious argument. In Fig. 1, the other premise that we propose in our template can capture the implicit premise inside the argument to define fallacy.

Potential error There are multiple templates, and each template consists of several slot-fillers, all of which are in free-text form which potentially causes the error. To minimize errors during annotation, the guidelines took into account various scenarios, such as giving preference to entity slot-fillers over events, among other considerations.

Implicit components The annotators encountered instanced with implicit components during the annotation process. In some cases, implicitness was primarily at the level of argumentative component (i.e., claim or premise). An example of the component is as follows:

- We should abolish the death penalty. Many respected people, such as actor, Chewbacca, have publicly stated their opposition to it.

On the other hand, for fallacies such as *false dilemma*, sentiment and other ingredients were left implicit. This was especially the case, as annotators were instructed to think of fallacious arguments in terms of being paraphrased/rewritten as an argument from consequences argument. An example is as follows:

- You can either support our police or Black Lives Matter.

Other schemes Following [8], we adopted templates in the form of argument from consequences, but both annotators felt that other argumentation schemes could be considered. In the case of the *fallacy of credibility*, argumentation schemes such as *argument from position to Know* [6] and *argument from expert opinion* [6] could be a better fit. Such schemes will be considered in future work.

4 Conclusion and Future Work

This work proposed fallacy templates as a novel tool for explaining the fallacious argument. Using the LOGIC dataset, we evaluate the templates across 5 fallacy types by conducting small trial annotation. We achieved a significant Cohen's kappa and coverage score for both all and filtered instances.

In the future, we aim to conduct a large-scale annotation of a fallacy template on larger and more natural arguments. We plan to create a dataset for Large Language Model (LLM) experiments and expand the template selection which addresses critical questions and the challenges we encountered.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22H00524. Additionally, we thank Yufeng Zhao and Tien Dang Huu, members of RebelsNLU lab for their generous support.

References

[1] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Logical fallacy detection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 7180–7198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[2] Martin Hinton. **Evaluating the Language of Argument**, Vol. 37. Springer Cham, 1 edition, 11 2020.

[3] Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. Fallacious argument classification in political debates. In **Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI**, pp. 4143–4149, 2022.

[4] Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. Argotario: Computational argumentation meets serious games. In Lucia Specia, Matt Post, and Michael Paul, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 7–12, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[5] Chris Reed. Preliminary results from an argument corpus. **Linguistics in the twenty-first century**, pp. 185–196, 2006.

[6] Douglas Walton, Christopher Reed, and Fabrizio Macagno. **Argumentation schemes**. Cambridge University Press, 2008.

[7] John Lawrence and Chris Reed. Argument mining using argumentation scheme structures. In **COMMA**, pp. 379–390, 2016.

[8] Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In **Proceedings of the 5th Workshop on Argument Mining**, pp. 79–89, 2018.

[9] Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In **Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon**, Vol. 2, pp. 801–815, 2015.

[10] Keshav Singh, Naoya Inoue, Farjana Sultana Mim, Shoichi Naito, and Kentaro Inui. IRAC: A domain-specific annotated corpus of implicit reasoning in arguments. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 4674–4683, Marseille, France, June 2022. European Language Resources Association.

[11] Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, Keshav Singh, and Kentaro Inui. LPAttack: A feasible annotation scheme for capturing logic pattern of attacks in arguments. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2446–2459, Marseille, France, June 2022. European Language Resources Association.

[12] Shoichi Naito, Shintaro Sawada, Chihiro Nakagawa, Naoya Inoue, Kenshi Yamaguchi, Iori Shimizu, Farjana Sultana Mim, Keshav Singh, and Kentaro Inui. TYPIC: A corpus of template-based diagnostic comments on argumentation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 5916–5928, Marseille, France, June 2022. European Language Resources Association.

[13] Catherine E. Hundleby. 238The Status Quo Fallacy: Implicit Bias and Fallacies of Argumentation. In **Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology**. Oxford University Press, 03 2016.

[14] Christopher W Tindale. **Fallacies and argument appraisal**. Cambridge University Press, 2007.

[15] Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. Robust and explainable identification of logical fallacies in natural language arguments. **Knowledge-Based Systems**, Vol. 266, p. 110418, 2023.

[16] Jacob Cohen. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, Vol. 20, No. 1, pp. 37–46, 1960.

[17] Douglas Walton. **Informal logic: A pragmatic approach**. Cambridge University Press, 2008.

[18] Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, et al. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In **Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**, pp. 619–630, 2012.

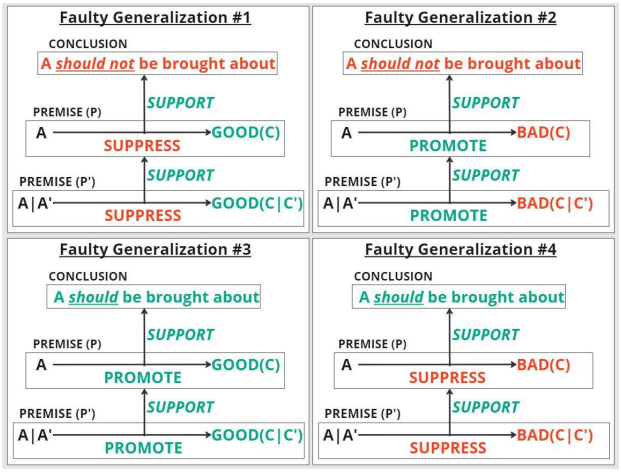

Figure 3 The 4 fallacy templates for *faulty generalization* fallacious arguments.

A Appendix

A.1 Critical Questions

Argument from consequences consists of 3 critical questions [6], including:

- **CQ1:** How strong is the likelihood that the cited consequences will (may, must) occur?
- **CQ2:** What evidence supports the claim that the cited consequences will (may, must) occur, and is it sufficient to support the strength of the claim adequately?
- **CQ3:** Are there other opposite consequences (bad as opposed to good, for example) that should be taken into account?

CQ1 and **CQ2** are the critical questions that we employ in our templates.

A.2 Templates

Fig. 2 is a list of *false dilemma* templates. The fallacy occurs due to restrictions on the available choices without considering any potential options. For example, "We either have to cut taxes or leave a huge debt for our children."

Fig. 3 is a list of *faulty generalization* templates. The fallacy occurs because applying a belief to a large population without having sufficient sample and non-biased. For example, "I know five people from Kentucky. They are all racists. Therefore, Kentuckians are racist."

Fig. 4 is a list of *false causality* templates. The fallacy occurs when assuming two events are correlated, they must also have a cause-and-effect. For example, "I drank bottled water and now I am sick, so the water must have made me sick."

Fig. 5 is a list of *fallacy of credibility* templates. The fallacy occurs when an appeal is made to some form of ethics, authority, or credibility. For example, "We are going to protest and not get in trouble because Mr. Iglesias said it is okay."

Fig. 6 is a list of *fallacy of logic* templates. The fallacy occurs when there is a logical flaw in the reasoning behind the argument, such as a propositional logic flaw. For example, "Allowing people to possess guns is like giving a bomb to a bunch of kids."

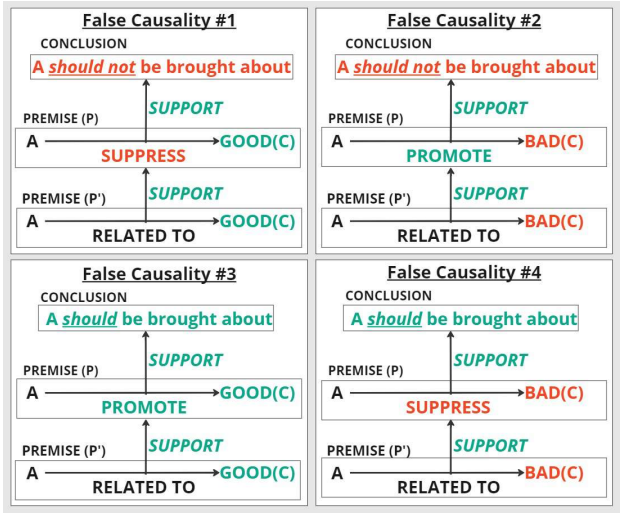

Figure 4 The 4 fallacy templates for *false causality* fallacious arguments.

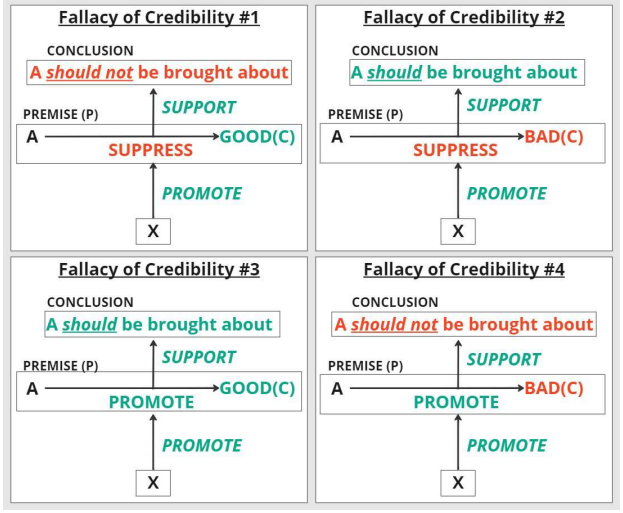

Figure 5 The 4 fallacy templates for *fallacy of credibility* fallacious arguments.

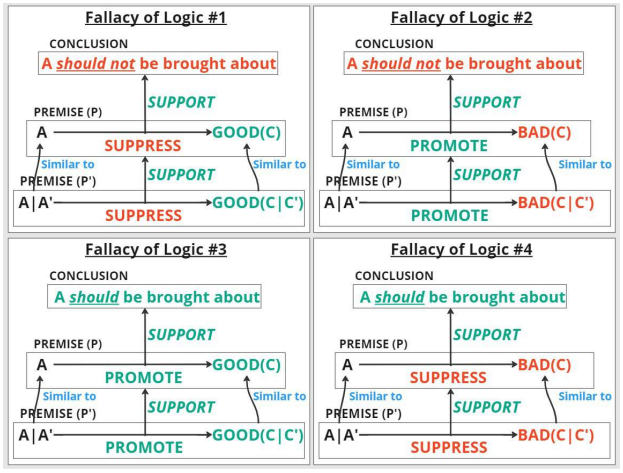

Figure 6 The 4 fallacy templates for *fallacy of logic* fallacious arguments.