反論の論理パターン解析: データセット構築と実現性検証

内藤昭一*[1,2,7] 王文質*[1,2] Paul Reisert[3] 井之上直也[4,2] Camélia Guerraoui[1,2,5]
山口健史[1] Jungmin Choi[2] Irfan Robbani 乾健太郎[6,1,2]
[1] 東北大学 [2] 理化学研究所 [3]Beyond Reason [4]JAIST [5]INSA Lyon [6]MBZUAI [7] リコー
{naito.shoichi.t1, wang.wenzhi.r7, guerraoui.camelia.kenza.q4}@dc.tohoku.ac.jp, beyond.reason.sp@gmail.com
naoya-i@jaist.ac.jp, jungmin.choi@riken.jp, robbaniirfan@jaist.ac.jp, kentaro.inui@mbzuai.ac.ae

概要

反論は批判的思考力を育成する有効な手段として教育の現場で用いられている．立論と反論の論理構造を精緻に解析する技術は，反論の評価やフィードバックの提供といった応用につながる可能性があるが，反論を対象とした取り組みは少ない．そこで，本研究では新たなタスク「反論の論理パターン解析」を提案する．反論が立論をどのように攻撃しているか，その要点を表した論理パターンを定義し，778 件の反論に対して論理パターンのアノテーションを行った．また，大規模言語モデルを用いてタスクの実現性を検証した結果，現行の言語モデルにとっても挑戦的なタスクであることが分かった．構築したデータセット及びアノテーションガイドラインは公開する予定である．

1 はじめに

反論は批判的思考力を向上させるための有効な手段であり，教育の現場で広く利用されている [1, 2, 3, 4]．反論をするためには，相手の論理構造を把握し，どこに反論をするのが効果的か見極めた上で，説得力のある論を組み立てる能力が要求される．これらは批判的思考力において重要な要素であり，反論の実践はそれらを育成するための有用な手段となる．

効果的な学習のためには適切なフィードバックが不可欠だが，個別にフィードバックを与える作業には多くの時間と労力を要し，現実的な運用には限界がある [5]．本研究では，学習者が作成した反論に自動でフィードバックを提示する下流タスクを想定し，**反論の論理パターン解析**を提案する．

論述の論理構造を分析する研究は多岐にわたる．中でも Argumentation Schemes [6] は日常の談話に見られる論証を約 60 種類の類型（パターン）として整理したもので，論述に対してフィードバックを提供する枠組みとして広く採用されている [7, 8, 9]．Argumentation Schemes のそれぞれのパターンには，その妥当性を検証するための批判的な質問（Critical Question）が紐づいている．例えば，Argumentation Schemes のひとつである *Argument from Expert Opinion* は，専門家 E の証言をもとに命題 A を真と結論付ける論証である．このパターンの論証には「E は A が属する分野の専門家か？」「他の専門家も A について一貫した評価をしているか？」といった Critical Question を促すことができる．

Argumentation Schemes から明らかなように，論述におけるパターンの導入は教育において有用な性質を備える．類似の論理構造を持つ論述をパターンとして抽象化することで，認知的負荷を減らすことができる [10]．また，論述を個別に理解するのではなく，一段抽象化した視点から見ることで，過去の経験を新しい状況に適用しやすくなる．しかし，反論に特化した類型はまだ存在しない．反論の類型が確立されれば Argumentation Schemes と同様の枠組みで，反論に対して効果的なフィードバックを提供するため基盤となることが期待される．

図 1 の例では，立論が宿題により自由時間が減っていると主張し，これに対し反論は宿題も一因ではあるが，クラブ活動や塾も自由時間を減らす要因であり，そちらを減らすべきと主張している．このような代替案を提示する論理パターンは，議題によらず用いられる．論理パターンが特定できれば，「なぜ反論側が提案する代替案を優先すべきか？」といった具体的にフィードバックを提示できる．

本研究では，反論に対してフィードバック提示することを見据え，反論が立論をどのように攻撃しているのかのパターンを特定する論理パターン解析を提案する．タスク提案にあたり，次の研究課題を設

＊ equal contribution

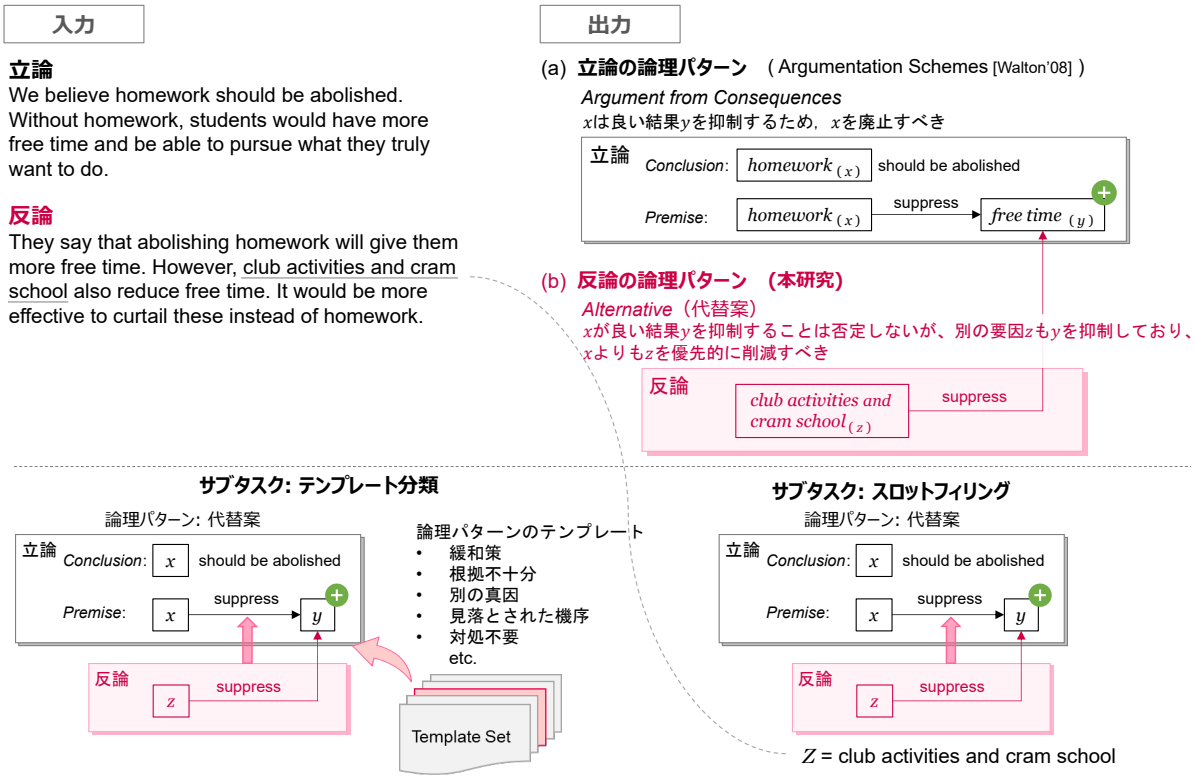


図 1 提案タスク「反論の論理パターン解析」の概略図．立論と反論を入力として，1) 事前に定義されたテンプレートセットから適切なものを選択するテンプレート分類，2) テンプレートのスロット z に相当するフレーズを反論から抽出するスロットフィリング の 2 つのサブタスクで構成される．

定する：(i) 十分なカバレッジを有する反論の論理パターンを定義できるか？ (ii) 大規模言語モデルはこれらの論理パターンをどの程度識別できるか？ 本研究の主要な貢献は以下のとおりである：

- 反論が立論をどのように攻撃をしているかを表した論理パターンを定義し，新たなタスクである「反論の論理パターン解析」を提案する．
- 778 件の反論に論理パターンをアノテートしたデータセットを構築する．データセットとアノテーションガイドラインは公開予定である．
- 構築したデータセット上で大規模言語モデルの性能を評価し，本タスクの実現性を検証する．

2 関連研究

議論における攻撃関係の表現 議論の構造，特に反論の表現に焦点を当てたものとして，*abstract argumentation frameworks* [11] は論証をノード，論証間の攻撃関係をエッジとした有向グラフにより議論を表現している．*ASPIC*$^+$ [12] は Pollock [13] の影響を受け，攻撃関係を rebuttal（結論への攻撃），undercutting（結論と前提の間の推論ステップへの攻撃），undermining（結論を支持する前提への攻撃）の三つに区別している．これらのフレームワークでは抽象的なレベルで攻撃関係を表現するもので，内容に深く踏み込むものではない．

攻撃関係の修辞的な側面を分析する研究として，Afantenos ら [14] は分節談話表示理論（SDRT） [15] を用いて，相手の論の一部を認めつつ，別の論を攻撃するといった修辞の表現方法を示している．LPAttack [16] は，特定の論証のタイプに焦点を当て，立論が述べる因果関係や価値観をどのように攻撃しているかを表現するアノテーションスキームを提案している．しかし，これらの研究は反論の類型の定義はされていない．Argumentation Schemes [6] は日常の談話に現れる論証の類型を整理しており，それぞれの型についてテンプレートを定義している．しかし，これらのスキームは独話的な論述に焦点を置いており，立論と反論の関係については考慮されていない．

攻撃関係の識別タスク Argument Mining では論述構造の自動解析に関するタスクが提案されている．独話的な議論を扱った研究として，Stab ら [17] は論証内の主張や前提といった要素間の支持・攻撃関係を同定するタスクを提案している．Peldszus

ら [18] は攻撃関係を rebuttal と undercut の二つに区別したデータセットを作成している．

対話的な議論を対象にした研究として，2 つの議論の間の支持・反対関係を識別するタスクが，オンラインフォーラム [19, 20] やピアレビュー [21]，ディベートのスピーチ [22, 23] を対象に存在する．議論全体ではなく，より詳細な単位（Argumentative Discourse Unit）で攻撃関係を識別するタスクも存在する [24, 25]．しかし，これらの研究では攻撃が行われている箇所を特定できるが，どのように攻撃されているかまでは分からない．

3 反論の論理パターン解析

本節では，タスクに対する要件とタスク設定，テンプレートセットについて説明する．

3.1 タスクに対する要件

反論の類型（論理パターン）の特定 一つ目の要件として，本タスクでは事前に定義した論理パターンの集合から適切なものを選択する分類問題として定式化する．より自由度の高い木構造やグラフ構造ではなく，パターンを採用する理由は類型の利点を活かすためである．共通のエッセンスを持つ反論をパターンとして抽象化することで，認知的負荷を減らすことができる [10]．また，反論を一段抽象化した視点で理解することで，以前の経験を新しい状況に適用しやすくなる．教育への応用を想定したときに類型が持つこれらの性質は有効だと考える．

複数の論理パターンの混在 二つ目の要件は，ひとつの反論内の複数の論理パターンを表現できることである．反論はしばしば複数の論理パターンを組み合わせられることがある．例えば，相手の論を否定した後に「仮に〜だったとしても」というフレーズを用いて，相手の論を一部認めつつも限定するようなケースがある．このようなケースに対応するため，複数の論理パターンの混在を許容する．

反論対象となる立論の種類 三つ目の要件として，本研究で扱う立論の種類は *Argument from Consequences* とする．*Argument from Consequences* とはある事柄が良い（悪い）をもたらすことを理由に命題の真偽を結論付ける論証である．図 1 の立論は *Argument from Consequences* に該当する例であり，宿題が自由時間という良い結果が抑制することを理由に宿題の廃止を主張している．

反論の論理パターンは立論の種類に応じて変わるため，本研究が定義する論理パターンは立論が *Argument from Consequences* の場合のみの適用となる．しかし，実際の利用シーンを考えると，それでも十分に価値があると考える．学習者に反論を作成してもらう際，立論は出題をする側でコントロールでき，はじめからあらゆる立論に対応する必要はないためである．

3.2 タスク設定

本研究では，反論の論理パターン解析を「テンプレート分類」と「スロットフィリング」の二つのサブタスクとして定義する（図 1）．

テンプレート分類 テンプレート分類では，立論と反論を入力として，反論の各文について論理パターンのテンプレートを選択する．各文に対して異なるテンプレートが選択可能なタスク設定としているのは，タスクに対する要件（§3.1）の「複数の論理パターンの混在」に対応するためである．文ごとにテンプレートを変えることもできるため，反論内の複数の論理パターンを表現することができる．テンプレート分類の性能は，文単位の F1 スコアにより評価する．

スロットフィリング スロットフィリングでは，テンプレート分類で選択されたテンプレートのスロットを埋める．テンプレートには，反論の主要な概念を表すスロットを含むものがある．このサブタスクでは，スロットに相当するフレーズを反論文から抽出する．スロットフィリングの性能は，トークン単位の F1 スコアにより評価する．

3.3 テンプレートセット

本研究では，*Argument from Consequences* の立論に対する反論のテンプレートセットを定義する．テンプレートセットの詳細については付録 A の表 2 に示す．

4 アノテーションスタディ

4.1 データ

アノテートを付与するデータとして，TYPIC [26] を用いる．TYPIC は 10 件の立論と約 1000 件程度の反論を含むコーパスである．上記コーパスから，*Argument from Consequences* の立論を対象とした反論 500 件 を選択する．さらに「学生はアルバイトをす

べきか」の論題について新たに 3 件の立論と 288 件の反論を追加で収集し，合計で 788 の反論を用いてアノテーションスタディを実施する．

4.2 アノテーションプロセス

アノテーションは，立論と反論のペアに対して，テンプレートセットから該当する論理パターンのテンプレートを選択し，テンプレート内のスロットに対応するフレーズを反論本文から抽出する作業を行う．適切なテンプレートがない場合は「該当なし」を選択する．アノテータは Amazon Mechanical Turk の選抜された七人のワーカーが担当する．アノテータ間の解釈の相違を抑えるため，事前に少数のサンプルを用いたキャリブレーションを三回実施する．778 件の反論に対して三人分のアノテーションを収集し，合計で 2,334 件のアノテーションを行う．

4.3 アノテーション結果

アノテーションの一貫性を評価するため，テンプレート分類の一致率を示す．778 件の反論に対するテンプレート分類の一致率は，Krippendorff's α [27] の値が 0.35 という結果となった．ある程度一貫性をもったアノテーションにはなっているが，改善の余地は多い．不一致事例の分析を通じて，ルール決めにより一致にもっていけるケース，主観性を考慮すべきケースを今後見極めたい．

定義したテンプレートセットの適用可能性を評価するため，テンプレートセットのカバレッジを示す．カバレッジは，テンプレート分類において「該当なし」以外の選択肢が選ばれた割合を算出する．カバレッジの値は 86.5% となり，定義した論理パターンは大部分に適用できると言える．

5 実現性検証

反論の論理パターン解析の自動化に向けて，現行の大規模言語モデルが本タスクをどの程度解けるのかの検証を行う．

5.1 実験設定

性能検証は in-context learning と fine-tuning の二つの設定で行う．in-context learning では，モデルとして GPT-4 と GPT-3.5 を用いる [1]．few-shot 設定を採用し，GPT-4 は 3 shot，GPT-3.5 は 8 shot の事例をランダムに与える [2]．fine-tuning では，モデルとして Llama2-7b を用いる．データ全体の 3 分の 2 である 518 件を訓練データとして用いる．テンプレート分類の正解ラベルは各アノテータの結果を MACE [28] により集約したものを使用する．スロットフィリングについては，テンプレート分類の結果が採用されたアノテータの抽出結果をすべて正解スパンとする．

表 1 テンプレート分類とスロットフィリングの Macro F1（3 分割交差検証の平均値）

設定	モデル	F1 (template)	F1 (slot)
few-shot	GPT-4	0.20	0.07
	GPT-3.5	0.15	0.13
fine-tuning	Llama2-7b	0.31	0.14

5.2 実験結果

テンプレート分類とスロットフィリングの結果を表 1 に示す．どちらのタスクにおいても，fine-tuning が few-shot の結果を上回った．few-shot においては，GPT-4 は与えている事例数が少ないにも関わらず，テンプレート分類において GPT-3.5 を超えている．一方で，スロットフィリングについては．GPT-3.5 が上回り，一貫した結果は得られなかった．全体として性能は低く，現行の大規模言語モデルにとっても挑戦的なタスクと言える．

6 おわりに

本研究は反論の論理パターン解析という新たなタスクを提案した．*Argument from Consequences* の立論に対する反論の論理パターンを整理し，778 の反論に対して論理パターンのアノテーションを施したデータセットを構築した．タスクの実現性を検証した結果，現行の大規模言語モデルを用いても大きな改善の余地があるタスクであることが分かった．

今後の課題として，より洗練されたアプローチの探求が挙げられる．論理パターンは多くの情報を内包するラベルである．一度の質問で回答を出力するのではなく，質問を分解し Step by Step に回答を出力するアプローチの検証を進めている [29]．また，論理パターン解析を取り入れたフィードバック提示システムの実装，およびユーザスタディを検討中である [30]．

1 OpenAI API を利用しており，GPT-4 は gpt-4，GPT-3.5 は gpt-3.5-turbo-16k を利用している

2 理想的な状態での性能を評価するため，few-shot における事例数は文脈として渡せる最大限の数としている．

謝辞

本研究は JSPS 科研費 22H00524 の助成，JST CREST JPMJCR20D2 の支援を受けたものである．

参考文献

[1] Abhijit Roy and Bart Macchiette. Debating the issues: A tool for augmenting critical thinking skills of marketing students. **Journal of Marketing Education**, Vol. 27, No. 3, pp. 264–276, 2005.

[2] Fulan Liu and Paul Stapleton. Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test. **System**, Vol. 45, pp. 117–128, 2014.

[3] David W. Johnson and Roger T. Johnson. Creative and critical thinking through academic controversy. **American Behavioral Scientist**, Vol. 37, No. 1, pp. 40–53, 1993.

[4] Michael Nussbaum. Using argumentation vee diagrams (avds) for promoting argument-counterargument integration in reflective writing. **Journal Of Educational Psychology**, Vol. 100, pp. 549–565, 08 2008.

[5] Rosalind Driver, Paul Newton, and Jonathan Osborne. Establishing the norms of scientific argumentation in classrooms. **Science Education**, Vol. 84, pp. 1–312, 05 2000.

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

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

[8] Yi Song and Ralph P. Ferretti. Teaching critical questions about argumentation through the revising process: effects of strategy instruction on college students' argumentative essays. **Reading and Writing**, Vol. 26, pp. 67–90, 2013.

[9] Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. Applying argumentation schemes for essay scoring. In **Proceedings of the First Workshop on Argumentation Mining**, pp. 69–78, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[10] Eleanor Rosch. Principles of categorization. In Eleanor Rosch and B. B. Lloyd, editors, **Cognition and Categorization**, pp. 27–48. Erlbaum, Hillsdale, NJ, 1978.

[11] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. **Artificial Intelligence**, Vol. 77, No. 2, pp. 321–357, 1995.

[12] Sanjay Modgil and Henry Prakken. The aspic + framework for structured argumentation: A tutorial. **Argument & Computation**, Vol. 5, , 02 2014.

[13] John L. Pollock. Defeasible reasoning. **Cognitive Science**, Vol. 11, No. 4, pp. 481–518, 1987.

[14] Stergos Afantenos and Nicholas Asher. Counter-argumentation and discourse: A case study. **CEUR Workshop Proceedings**, Vol. 1341, , 01 2014.

[15] N. Asher and A. Lascarides. **Logics of Conversation**. Studies in Natural Language Processing. Cambridge University Press, 2003.

[16] 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 **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2446–2459, Marseille, France, June 2022. European Language Resources Association.

[17] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. **Computational Linguistics**, Vol. 43, No. 3, pp. 619–659, September 2017.

[18] 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.

[19] Akiko Murakami and Rudy Raymond. Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In **Coling 2010: Posters**, pp. 869–875, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[20] Sara Rosenthal and Kathy McKeown. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In **Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 168–177, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

[21] Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 7000–7011, Online, November 2020. Association for Computational Linguistics.

[22] Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. Never retreat, never retract: Argumentation analysis for political speeches. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 32, No. 1, Apr. 2018.

[23] Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Out of the echo chamber: Detecting countering debate speeches. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7073–7086, Online, July 2020. Association for Computational Linguistics.

[24] Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. Analyzing argumentative discourse units in online interactions. In **Proceedings of the First Workshop on Argumentation Mining**, pp. 39–48, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[25] Jacobus H Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. **Language Resources and Evaluation**, Vol. 54, pp. 123 – 154, 2019.

[26] 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, 2022.

[27] Klaus Krippendorff. **Content Analysis: An Introduction to Methodology**. Sage Publications, Inc., Beverly Hills, CA, 1980.

[28] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1120–1130, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[29] Wenzhi Wang, Shoichi Naito, Paul Reisert, Naoya Inoue, Camélia Guerraoui, Jungmin Choi, Irfan Robbani, and Kentaro Inui. Exploring task decomposition for assisting large language models in counter-argument logical structure analysis. **30th Annual Meeting of the Natural Language Processing (NLP2024)**, 2024.

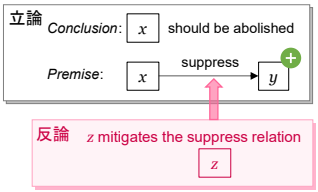
[30] Camélia Guerraoui, Paul Reisert, Naoya Inoue, Wenzhi Wang, Shoichi Naito, Jungmin Choi, Irfan Robbani, and Kentaro Inui. Argvantage: the new pedagogical system to learn argumentation. **30th Annual Meeting of the Natural Language Processing (NLP2024)**, 2024.

A テンプレートセット

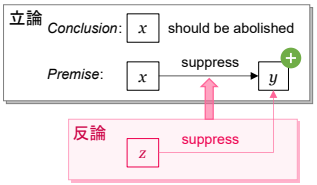
Argument from Consequences 型の立論に対する反論のテンプレートセットを表 2 に示す．

表 2 反論のテンプレートセット．反論対象の立論は *Argument from Consequences* に該当し，x が y という良い結果を抑制することを理由に x を廃止すべきという主張をしている．各テンプレートは異なる反論のパターンを表現している．

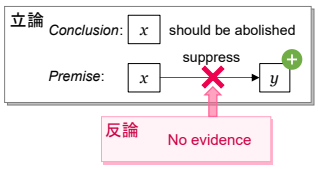
緩和策
x が y という良い結果を抑制することは否定しないが，その因果関係は z という手段で緩和できる



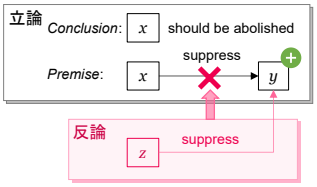
代替案
x が y という良い結果を抑制することは否定しないが，別の要因 z も y を抑制しており，x よりも z を優先的に削減すべき



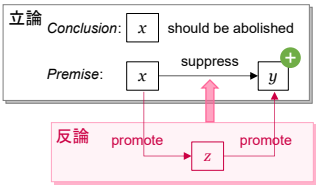
根拠不十分
x が y を抑制する十分な根拠が述べられていない



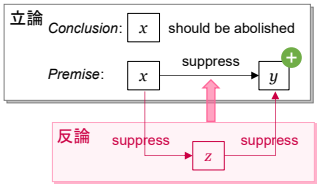
別の真因
y を抑制する真の原因は x ではなく z である。よって，x を廃止しても問題は解決しない



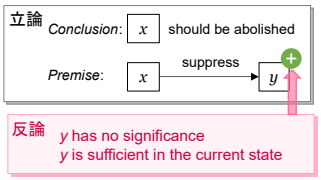
見落とされた機序 #1
y を促進する別の要因 z が存在し，x は z を促進する。よって，x は y を抑制するのではなく，むしろ y を促進する



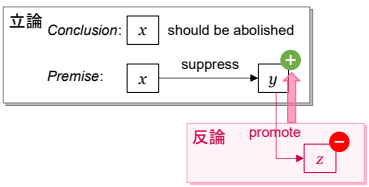
見落とされた機序 #2
y を抑制している別の要因 z が存在し，x は z を抑制する。よって，x は y を抑制するのではなく，むしろ y を促進する



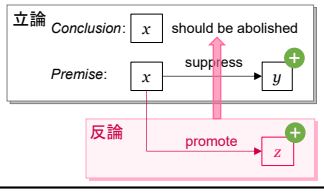
対処不要
y の意義は限定的である，y は現状でも十分である



y による好影響
y により z という悪い結果が生まれるため，y は良い結果ではない



y とは別観点のメリット #1
x は y とは別観点の良い結果 z を促進する



y とは別観点のメリット #2
x は y とは別観点の悪い結果 z を抑制する