Fallacy Attribution in Two-Party Disputes: Extension, Agreement, and Depth

E. Cavill*, Q. Barratt
Institute for Terminal Discourse Studies

Coding manual and the move set in Appendix A.

Abstract

Participants in these channels name fallacies. We report three measurements of the practice across 1,540 attributions drawn from 47 channels. Extension: of 612 attributions of one named fallacy, the personal remark at issue was offered in support of a conclusion in 71 cases; in the remainder it accompanied an argument without bearing on it. We treat this as a change in what the term denotes rather than as an error in applying it. Agreement: four coders shown the same 200 moves and asked which fallacy, if any, each commits agreed at $\kappa = 0.31$, and on 44 moves returned four different answers. Depth: an attribution is itself attributable, and the resulting stack reaches a median depth of 2 and a maximum of 6, at which point the exchange concerned the propriety of naming fallacies and no longer the matter in dispute. In 1,504 of the 1,540 attributions the participant named made no subsequent change to the argument named.

1 Introduction

To name the fallacy an opponent has committed is to make a claim about the form of their argument rather than its content, and it is available at any point in an exchange without reference to the matter in dispute. It is, in the archive's own vocabulary, a move.

This paper does not assess whether the attributions we collected are correct. Whether a given remark commits a given fallacy is a question in logic, and the disputes we study are not conducted by logicians; an assessment would tell us about our own coders. We measure three properties of the practice that can be established without adjudicating any instance of it: what one heavily used term now denotes, how far trained coders agree on applying the taxonomy, and how deep the attributions stack.

2 Corpus and coding

The corpus is 1,540 attributions drawn from 47 channels over 22 months. An attribution was coded where a participant named a fallacy, in any of the forms in Appendix A, in reference to a turn by the other participant. Naming a fallacy in the abstract, without reference to a turn, was not coded.

3 Extension

Six hundred and twelve of the 1,540 attributions concern one term, which we do not print in full here for the reason given in Section 4. In its standard treatment the term applies where a remark about the person is offered *as* the argument, so that rejecting the person is doing the work of rejecting the claim.

*Corresponding author.

We coded whether the remark at issue was offered in support of a conclusion. It was in 71 cases. In the remaining 541 the remark accompanied an argument without bearing on it: the participant insulted the other and also argued, and the two were separable.

We do not record this as an error. A term used by a large population in a consistent way has that use, and the use we observe is consistent: it names an insult delivered during a dispute. It is a broader class than the standard treatment names, it is easier to identify, and it has evidently displaced the narrower one. What follows is only that the participant making the attribution and the participant receiving it may not be discussing the same property, and that neither has any way to discover this from the exchange.

4 Agreement

Four coders were shown the same 200 moves, drawn at random from the corpus, and asked which fallacy each committed, if any, choosing from the list in Appendix A or answering NONE.

Table 1: Agreement among four coders on 200 moves.

Measure	Value
Fleiss' κ, all categories	0.31
Moves with unanimous agreement	38
Moves with three of four agreeing	61
Moves returning four different answers	44
Moves all four coded NONE	12

We take the figure to characterise the taxonomy rather than the coders. The coders were trained together, worked from one list, and disagreed most on the moves that occur most often. A taxonomy on which trained coders return four different answers to one move in five is not an instrument that a participant can apply to another participant's turn and expect to be understood.

5 Depth

An attribution is a turn, and a turn can be the subject of an attribution. The stack that results has a median depth of 2 and a maximum, in this corpus, of 6.

Table 2: The deepest stack observed, one turn per level.

Level	The turn is
1	an argument
2	an attribution against the argument
3	an attribution against the attribution
4	an objection to attribution as a practice
5	an attribution against the objection
6	an observation that the matter in dispute has not been raised since level 1

The observation at level 6 is correct, and was made by the participant who opened the stack at level 2. It did not conclude the exchange, which continued for a further 311 messages.

6 Discussion

In 1,504 of the 1,540 attributions the participant named made no subsequent change to the argument named. Of the 36 who did, 29 changed the wording and retained the claim.

We note that the practice is defended on the ground that it improves the quality of argument, and that we are not in a position to evaluate that defence: it concerns a counterfactual population of exchanges in which fallacies go unnamed, and we have not been able to locate one.

7 Limitations

The extension result rests on our own coding of whether a remark was offered in support of a conclusion, which is the judgement the disputants are themselves making and disagreeing about; our coders agreed on it at $\kappa = 0.68$, which is adequate and is not high. The depth measure counts turns and not participants, and a stack sustained by two participants alternating is not distinguished here from one sustained by four. Appendix A's list was assembled from the lists in circulation (see arghXiv:2603.09550) and is therefore not a standard instrument, which is the substance of Section 4 and a limitation of it.