Towards a Computational Ontology of Fallacies

Luca Biccheri
University of Geneva
luca.biccheri@unige.ch

Roberta Padlina
University of Geneva
roberta.padlina@unige.ch

Frederic Goubier
University of Geneva
frederic.goubier@unige.ch

1 Reversing fallacious reasoning into sound argumentation

Recent generative AI developments, preceded by the massive employment of Information Technology during the last decades, have revolutionized the way people access, consume, and produce information. While such technologies positively empower human capacity to communicate and exchange information, they also yield side effects such as the spreading of fake news and the use of manipulative tools, with harmful implications for society that are hard to estimate. Alongside fake-news production, deceptive forms of argumentation, aka fallacies[1], traditionally analyzed in humanistic studies, are being introduced into everyday public dimensions.

The development of technologies able to detect and break down fallacious forms of reasoning represents a very welcome *desideratum*. This is precisely one of the main objectives of the REVLOG project [2], which, by investigating the rich corpus of medieval theories of argumentation across the Arabic, Greek East, Hebrew, and Latin West traditions, aims to build a computational infrastructure (including digital scholarly editions, ontologies, and AI-driven systems for information retrieval and classification) able to enhance our understanding of fallacies.

Despite having been studied for centuries, we still lack a standard definition of 'fallacy'. Very roughly, *fallacious arguments* are those that, while appealing or seemingly sound at first sight, actually break some relevant *reasoning principles* upon further analysis. It follows that, in defining fallacies, much depends on the *kind* of reasoning one is considering: in deductive reasoning, invalid arguments count as fallacies (however, 'begging the question' is not invalid and yet is typically considered fallacious); in inductive reasoning, a weak strength between premises and the conclusion is a sign that a fallacy is at stake (Hansen, 2024). Moreover, many types of fallacies overlap one another, and there is no complete list of their types, as much depends on the taxonomical system adopted (Copi et al., 2016; Hansen, 2024).

Along these lines, REVLOG takes this complexity at face value, based on two research hypotheses: i) studying fallacious reasoning means, at the same time, setting the principles according to which an argument can be considered *sound*; ii) embracing a pluralistic view on fallacies does not rule out the possibility of finding some common *ontological patterns* helpful to shed light on how arguments are built. Ontologies are indeed at the core of REVLOG. In what sense and to what extent this represents a novelty in the study of fallacies is introduced in the following.

2 A DOLCE-based account of arguments

In the fields of Computer Science and AI, 'computational ontology' generally denotes artifacts used to represent, through formal languages, information about some domain of knowledge and to implement automated reasoning (Guarino et al., 2009). Within REVLOG, the ontological development will consist of two phases: a) conceptual modeling in First-Order Logic (FOL); b) implementation in the Web Ontology Language (OWL).

[1] Note that 'fake news' and 'misinformation' are broader concepts than 'fallacy'. The treatment of the relation between these notions is outside the scope of this work. For more information on the topic, see e.g. (Beisecker et al., 2024; Hruschka and Appel, 2023).

[2] As in reverse engineering, engineers investigate the faults that cause machine failures and therefore improve their design, in the same way, it is possible to break down fallacious reasoning to reconstruct what it means for an argument to be sound. Hence the name of the project, REVLOG, meaning 'logic in reverse'.

Since Revlog is still in its early stages[3], we are currently dealing with conceptual modeling, by 'negotiating' a philosophical agreement between the different traditions studied within the project, each bringing their own perspective on fallacies. However different these perspectives may be, we argue that it is possible to identify a common schema on the notion of fallacy, provided that we look at very general **ontological and logical patterns**. These will help to identify the root node of the taxonomy of fallacies, which will be specialized into nodes representing their types. Through a feedback loop mechanism, this process will be iterated to further sharpen our ontological model.

To achieve this aim, the so-called 'foundational ontologies' are the best candidates, as these deal with broad categories including *objects, events, qualities*, and relations such as *constitution, participation, dependence, being part-of.* An example of a foundational ontology that shall be employed is Descriptive Ontology for Linguistic and Cognitive Engineering (**DOLCE**) (Borgo et al., 2022).

To the best of our knowledge, no foundational ontologies have been used so far to model fallacies and implement automated reasoning on them. Recent attempts rely instead on machine learning algorithms or LLMs trained on shallow, e.g. web-scraped data with annotations by non-expert labelers (Jin et al., 2022). REVLOG will also use AI algorithms, but the dataset will be built from high-quality data provided by proficient scholars, and algorithm accuracy will be further augmented through systems embedding the patterns of our ontology (Vsevolodovna and Monti, 2025).

Back to our model, as a concrete example of ontological patterns, let us outline some hypotheses we are developing within the Latin tradition, starting from Aristotelian studies. Scholars are interested in documenting both the *material* corpus of fallacies, e.g. a specific argument (comprising premises and a conclusion) found in a manuscript, and the abstract properties (such as 'Valid', 'Invalid', etc.) predicated of such arguments. This raises an ontological tension: stating that an abstract property such as 'Invalid' inheres in a physically located argument sounds at least bizarre. **DOLCE** helps us resolve this, since it axiomatically constrains abstract properties to inhere only in non-physical entities. This motivates classifying arguments as *social entities* (after all, language and manuscripts are social enterprises) and, crucially, social entities in **DOLCE** are themselves non-physical, which is consistent with treating properties such as 'Invalid' as abstract and inherent to the latter. Furthermore, since both Aristotle and modern scholars treat fallacies as a kind of argument, **DOLCE** axioms on social entities allow us to constrain the meaning of *fallacy* accordingly: a fallacy exists at some time interval but is not located in space. We believe these kinds of ontological patterns generalize well beyond the Latin tradition and the Aristotelian framework, capturing something fundamental about the nature of fallacies across different contexts. The strength of our conceptual model, as we will show, is therefore that it allows multiple perspectives to be present in the same knowledge base without, on the one hand, triggering inconsistencies, while at the same time highlighting what, if anything, is common to such views.

[3]The project started in May 2025 and will end in April 2031.

References

Sven Beisecker, Christian Schlereth, and Sebastian Hein. 2024. Shades of fake news: How fallacies influence consumers' perception. *European Journal of Information Systems* 33(1):41–60.

Stefano Borgo, Roberta Ferrario, Aldo Gangemi, Nicola Guarino, Claudio Masolo, Daniele Porello, Emilio M Sanfilippo, and Laure Vieu. 2022. Dolce: A descriptive ontology for linguistic and cognitive engineering. *Applied ontology* 17(1):45–69.

Irving M Copi, Carl Cohen, and Kenneth McMahon. 2016. *Introduction to logic*. Routledge.

Nicola Guarino, Daniel Oberle, and Steffen Staab. 2009. What is an ontology? In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, Springer, Berlin, Heidelberg.

Hans Hansen. 2024. Fallacies. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University. Fall 2024 edition.

Timon MJ Hruschka and Markus Appel. 2023. Learning about informal fallacies and the detection of fake news: An experimental intervention. *PLoS One* 18(3):e0283238.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. pages 7180–7198.

Ruslan Idelfonso Magana Vsevolodovna and Marco Monti. 2025. Enhancing large language models through neuro-symbolic integration and ontological reasoning. *arXiv preprint arXiv:2504.07640* .