Probabilistic interpretations of argumentative attacks: Logical and experimental results

Niki Pfeifer[1] , **Christian G. Fermüller**[2]

[1]Department of Philosophy, University of Regensburg, Germany
[2]Institute of Logic and Computation, TU Wien, Austria
niki.pfeifer@ur.de, christian.fermueller@tuwien.ac.at

Argumentation, reasoning, and uncertainty are key aspects of knowledge representation. We present an interdisciplinary approach to argumentation combining logical, probabilistic, and psychological perspectives. We investigate logical attack principles which relate attacks among claims with logical form. For example, we consider the principle that an argument that attacks another argument claiming A triggers the existence of an attack on an argument featuring the stronger claim $A \wedge B$. More precisely, let $F \longrightarrow G$ denote that, in a given Dung-style argumentation frame, there exists an argument with claim F that attacks some other argument with claim G; for short we will say "F attacks G". The mentioned principle about attacking a conjunction, can thus be formulated as follows:

(C.∧) If $F \longrightarrow A$ or $F \longrightarrow B$ then $F \longrightarrow A \wedge B$.

Similarly, one might also express the following inverse principle for conjunctive claims:

(C.∧)' If $F \longrightarrow A \wedge B$ then $F \longrightarrow A$ or $F \longrightarrow B$.

In the present paper we also consider a number of similar principles pertaining to disjunctive, negated, and implicational claims, which were originally presented in (Corsi and Fermüller 2017). Some of these attack principles seem to be *prima facie* more plausible than others. To support this intuition, we suggest an interpretation of these principles in terms of coherent conditional probabilities (see, e.g., (Coletti and Scozzafava 2002; Gilio 2002; Pfeifer and Sanfilippo 2017)). The basic intuition of coherence is usually explained in betting terms, specifically in terms of avoiding Dutch books. Accepting a Dutch book implies sure loss, thus making sure to avoid such bets is the basic rationality requirement. A *conditional event* $C|A$ is the (conditional, trivalent) object which is measured by the corresponding conditional probability $p(C|A)$.

Definition 1. *A* conditional event $C|A$ *is* true *if* $A \wedge C$ *is true,* false *if* $A \wedge \neg C$ *is true, and* void *(or undetermined) if* $\neg A$ *is true.*

In betting terms, Definition 1 can be read such that you *win* the bet on $C|A$ when $A \wedge C$ is true, you *lose* when $A \wedge \neg C$ is true, and you *get your money back* when $\neg A$ is true. Because of its trivalence, $C|A$ cannot be expressed by any Boolean function. Within the coherence approach, conditional probability is primitive (and not defined by the fraction, $p(A \wedge C)/p(A)$, which—in order to avoid fractions over zero—requires positive-probability antecedents, $p(A) > 0$) and allows for properly managing zero-probability antecedents.

Concretely, we suggest to read "F attacks A" as the assertion that it is likely that A does not hold, given that F holds. More precisely, we suggest an interpretation of $F \longrightarrow A$ as $p(\neg A|F) \geq t$, which is *parameterized* for some threshold $0.5 < t \leq 1$. We note that $p(\neg A|F) \geq t$ is equivalent to $p(A|F) < t$.

Thus, **(C.∧)** and **(C.∧)'** turn into

(C.∧)^t_p If $p(\neg A|F) \geq t$ or $p(\neg B|F) \geq t$, then we also have $p(\neg(A \wedge B)|F) \geq t$;

and

(C.∧)'^t_p If $p(\neg(A \wedge B)|F) \geq t$ then $p(\neg A|F) \geq t$ or $p(\neg B|F) \geq t$;

respectively.

This interpretation is naturally generalized from qualitative to quantitative principles. Rather than just considering whether $F \longrightarrow A$ holds or not, we will use $F \xrightarrow{w} A$ to denote that F attacks A with the weight (or to the degree) w and interpret this probabilistically by $p(\neg A|F) = w$. Let us stress again that "attack", here, is a relation between propositions and not between arguments. In the literature, there are various suggestions for generalizing ordinary AFs to weighted AFs (or systems), where real numbers attached to attacks between arguments are intended to represent degrees of strength of such attacks (see, in particular, (Dunne et al. 2011)). The weights are understood to be normalized, with 1 being the maximal weight of any attack, whereas $F \xrightarrow{0} A$ means that F in fact does not attack the claim A at all.

Concerning conjunctive claims, the following probabilistic inference principle is proven to be coherent (Gilio 2002):

(And)$_p$ From $p(A|F) = x$ and $p(B|F) = y$ infer $\max(0, x + y - 1) \leq p(A \wedge B|F) \leq \min(x, y)$.

Applying the upper bound to a negated claim, thus turning minimum into maximum, yields the following quantitative version of the respective qualitative attack principle **(C.∧)**:

(G$^w_{\geq}$.∧) If $F \xrightarrow{x} A$, $F \xrightarrow{y} B$, and $F \xrightarrow{z} A \wedge B$, then $z \geq \max(x, y)$.

Likewise, the lower bound in **(And)**$_p$ yields

(Ł$^w_{\geq}$.∧) If $F \xrightarrow{x} A$, $F \xrightarrow{y} B$, and $F \xrightarrow{z} A \wedge B$, then $z \geq \min(1, x + y)$.

A key feature of our approach is that we use our probabilistic semantics to evaluate the rationality of principles which govern the strength of argumentative attacks: we show how the coherence approach to probability can serve to guide the rational selection of qualitative and quantitative principles regarding the existence of attacks on logically compound claims.

For example, concerning the qualitative principles for conjunctive claims we show

Proposition 1. **(C.∧)**^t_p *holds for every threshold* $t > .5$. *However,* **(C.∧)'**^t_p *does not hold for any* $t > .5$.

This result confirms that **(C.∧)** is intuitively plausible, while **(C.∧)'** appears to be too strong. The paper also features a running example referring to arguments about various weather conditions. This example instantiates our abstract principles about attacks involving claims of a particular logical form to the level of concrete statements. It also serves to support our judgments on the intuitive (im)plausibility of the various attack principles.

Concerning the corresponding quantitative principles governing conjunction, **(G$^w_{\geq}$.∧)** and **(Ł$^w_{\leq}$.∧)** are justified analogously, whereas corresponding inverse principles are rejected. The labels **G** and **Ł** refer to the Gödel and Łukasiewicz logic, respectively, since the truth functions for conjunction match the exhibited bounds. In the paper we also investigate Product logic **P**, where we show that the corresponding attack principle holds under independence assumptions.

In order to complement our theoretical analysis with an empirical perspective, we report on an experiment with students of the TU Vienna ($n = 139$) which explores the psychological plausibility of selected attack principles. While we are convinced that our approach is intuitive and plausible from a theoretical point of view, we were surprised by the relatively heterogeneous experimental results. We observed some evidence in favor of our hypotheses under the experimental condition where participants generated strengths of attacks. Interestingly, the majority of the participants hit some of the optimal coherent bounds as predicted. Violations most frequently concerned the lower bounds. When the participants merely judged the correctness of attack strength candidates, however, most responses did not confirm our hypotheses. The heterogeneous agreement between the predictions and the responses could be caused by various factors including (i) lower data quality in a lecture hall experiment compared to individual testing, (ii) different response formats (the open response format (strength generation) appeared to be more appropriate compared to the forced choice response format (correctness judgments) to investigate quantitative attack principles), and (iii) possible confusions caused by the negations involved in the probabilistic semantics of the attack relations (i.e., $p(\neg B|A)$ should be high in order that $A \longrightarrow B$ holds). Although attack relations are intuitive and plausible from theoretical points of views, maybe support relations are *psychologically* more intuitive, as they can be represented positively by the human mind without requiring implicit negations. Future experimental work is needed to further explore the psychological plausibility of formal attack principles.

As acknowledged by the list of topics in the call of KR2024, argumentation theory emerged as an important part of knowledge representation in the past decades. Likewise, reasoning and uncertainty are important aspects in the study of knowledge bases. Although referring to the mainstream paradigm of Dung-style argumentation frames, we emphasize several often neglected aspects of logical argumentation on what we call a "semi-abstract level" of analysis: the interplay between the logical form of claims arguments and the attack relation between arguments, the relation to probabilistic reasoning, and the experimental assessment of theoretical findings. Hence our approach is interdisciplinary, combining research in logics, probabilistic reasoning, and experimental psychology. We conceive our paper *programmatically*, as first steps towards new research directions on the interface between knowledge representation research and formal argumentation theory. In particular we plan to investigate logical principles for the support relation between arguments in a similar vain.

References

Coletti, G., and Scozzafava, R. 2002. *Probabilistic logic in a coherent setting*. Kluwer.

Corsi, E. A., and Fermüller, C. G. 2017. Logical argumentation principles, sequents, and nondeterministic matrices. In Baltag, A.; Seligman, J.; and Yamada, T., eds., *LORI 2017*, volume 10455 of *LNCS*, 422–437. Berlin: Springer.

Dunne, P. E.; Hunter, A.; McBurney, P.; Parsons, S.; and Wooldridge, M. 2011. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence* 175(2):457–486.

Gilio, A. 2002. Probabilistic reasoning under coherence in System P. *Annals of Mathematics and Artificial Intelligence* 34:5–34.

Pfeifer, N., and Sanfilippo, G. 2017. Probabilistic squares and hexagons of opposition under coherence. *International Journal of Approximate Reasoning* 88:282–294.