Probabilistic interpretations of argumentative attacks: logical and experimental foundations[*]

Niki Pfeifer
Munich Center for Math. Phil.
LMU Munich
niki.pfeifer@lmu.de

Christian G. Fermüller
Dept. for Logic & Computation
TU Vienna
chrisf@logic.at

Abstract

We present an interdisciplinary approach to study systematic relations between logical form and attacks between claims in an argumentative framework. We propose to generalize qualitative attack principles by quantitative ones. Specifically, we use coherent conditional probabilities to evaluate the rationality of principles which govern the strength of argumentative attacks. Finally, we present an experiment which explores the psychological plausibility of selected attack principles.

1 Introduction

Various disciplines study argumentation, including computer science (e.g., [6, 1]), philosophy (e.g., [21]), and psychology (e.g., [11, 13]). Our approach is an interdisciplinary one, as we combine elements of Dung-style abstract argumentation [6], logical argument forms, coherent conditional probability, and also present an experimental assessing the descriptive validity of selected formal principles.

We investigate systematic relations between logical form and attacks between claims in an argumentative framework. Argumentation is a highly complex and dynamic process. Usually, arguments are conceived as premise ("support") and conclusion ("claim") pairs. We focus on static argumentation and are only interested in claims formalized by classical propositional formulæ.

The outline of the paper is as follows: Section 2 gives a brief survey of qualitative attack principles which were investigated in a modal logical framework [3]. We argue, that the modal logical framework appears to be too coarse, especially for

[*]We thank Gernot Salzer for making the experiment possible during his class: thanks also to his students for their participation. Niki Pfeifer is supported by his DFG project PF 740/2-2 (part of the SPP1516).

modelling the quantitative dimension of attack principles. We therefore propose to generalize these principles by adopting a probabilistic framework. Specifically, we use coherent conditional probabilities to evaluate systematically the rationality of attack principles: coherence provides a criterion for selecting attach principles (i.e., "good" principles should be coherent). In Section 3 we show how to model the qualitative attack principles in probabilistic terms. Section 4 presents our probabilistic analysis of the quantitative attack principles and their semantics. Section 5 presents an experiment which aims to explore the psychological plausibility of selected quantitative attack principles. Section 6 concludes the paper by some remarks on future research.

2 Qualitative attack principles

In what follows we write "$A \longrightarrow B$" to denote that there is an argument claiming A that attacks an argument with claim B. Thus, strictly speaking, attack relations are between arguments. However, we simply say "A attacks B".

It seems intuitively obvious that given attacks on claims implicitly entail attacks on further claims which logically imply the original, explicitly attacked claims. A corresponding 'general attack principle' has been formulated in [3]:

(A.gen) If $F \longrightarrow A$ and $B \models A$ then $F \longrightarrow B$.

While it may be problematic to consider *all* classical logical implicants as inducing implicit attacks in this manner, at least the following instances of **(A.gen)** seem reasonable, since they are immediate and hold even if the consequence relation ($\models$) is constrained to minimal logic [12].

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

(A.$\vee$) If $F \longrightarrow A \vee B$ then $F \longrightarrow A$ and $F \longrightarrow B$.

(A.$\supset$) If $F \longrightarrow A \supset B$ then $F \longrightarrow B$.

Actually, **(A.$\supset$)** may raise concerns, since $A \supset B$ does not *relevantly* follow from B, cf. [7]. Hence, one may prefer the following weaker rationality postulate, instead.

(B.$\supset$) If $F \longrightarrow B$ and $F \not\longrightarrow A$[1] then $F \longrightarrow A \supset B$.

Concerning negation, the following principle is intuitively plausible.

(A.$\neg$) If $F \longrightarrow A$ then $F \not\longrightarrow \neg A$.

On the other hand, one can formulate inverse forms of the above principles:

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

(C.$\vee$) If $F \longrightarrow A$ and $F \longrightarrow B$ then $F \longrightarrow A \vee B$.

[1] $F \not\longrightarrow A$ denotes that A is *not* attacked by F.

(C.$\supset$) If $F \longrightarrow A \supset B$ then $F \longrightarrow B$ and $F \not\longrightarrow A$.

(C.$\neg$) If $F \not\longrightarrow A$ then $F \longrightarrow \neg A$.

These last mentioned principles seem, at least partly, to be intuitively much more demanding than those following from **(A.gen)**. Indeed, the results of [3] imply that imposing *all* of the above (connective specific) attack principles amounts to an alternative characterization of classical logic, while proper subsets of the full set of these principles lead to weaker logics that result from discarding some of the logical inference rules of Gentzen's classical sequent calculus **LK**.

The indicated situation calls for a robust interpretation of the attack relation that is capable of formally supporting (or questioning, as appropriate) informal intuitions about the varying strength of the attack principles. To this aim the authors of [3] suggest to translate $F \longrightarrow A$ into the modal formula $\Box(F \wedge \neg G)$, where in the underlying Kripke frame $\langle W, R \rangle$, W models the set of possible states of affairs and wRv is read as "v is a possible alternative from the viewpoint of w". In other words, this setup suggests that a given attack refers to all possible states of affairs in which the attacking claim holds and asserts that the attacked claim does not hold in any of those states. If one stipulates that R is reflexive (or at least serial) than this interpretation of $F \longrightarrow A$ renders the principles **(A.$\wedge$)**, **(A.$\vee$)**, **(C.$\vee$)**, **(C.$\supset$)**, and **(A.$\neg$)** formally sound while one may construct counter examples for the translations of the principles **(C.$\wedge$)**, **(C.$\neg$)**, **(B.$\supset$)**, and therefore also of **(A.$\supset$)**. Since this result is unsatisfying, in particular with respect to the arguably counter-intuitive classification of attack principles for implication, three alternative modal interpretations where briefly discussed in [3] as well. However, each of the suggested translations of $F \longrightarrow A$ into modal logic is too coarse, since there seems be no principled way to disentangle strong and weak attack principles. Moreover, modal logic does not support quantitative refinements of the qualitative attack principles.

3 Probabilistic semantics

In light of the results of [3], as sketched in Section 2, the challenge to come up with an intuitively convincing and formally sound interpretation of the attack relation between claims of arguments remains open. This motivates us to explore to which extent one may employ coherence-based *conditional probability* (see, e.g., [2, 9]) for this purpose. 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 interpret $F \longrightarrow A$ by $p(\neg A|F) \geq t$ for some threshold $0.5 < t \leq 1$. Throughout the paper, we assume that F is not a logical contradiction (i.e., F is not equivalent to $\bot$), since otherwise the corresponding conditional probability is undefined.

Translating the attack principles that refer to conjunction, disjunction, and negation according to the suggested interpretation is straightforward. The following claims correspond to the 'weak' principles **(A.$\wedge$)**, **(A.$\vee$)**, and **(A.$\neg$)**:

$(\mathbf{A.}\wedge)_p$ If $p(\neg A|F) \geq t$ or $p(\neg B|F) \geq t$, then $p(\neg(A \wedge B)|F) \geq t$.

$(\mathbf{A.}\vee)_p$ If $p(\neg(A \vee B)|F) \geq t$, then $p(\neg A|F) \geq t$ and $p(\neg B|F) \geq t$.

$(\mathbf{B.}\neg)_p$ If $p(\neg A|F) \geq t$, then $p(\neg\neg A|F) = p(A|F) < t$.

Analogously, the inverse ('strong') principles translate as follows:

$(\mathbf{C.}\wedge)_p$ If $p(\neg(A \wedge B)|F) \geq t$ then $p(\neg A|F) \geq t$ or $p(\neg B|F) \geq t$.

$(\mathbf{C.}\vee)_p$ If $p(\neg A|F) \geq t$ and $p(\neg B|F) \geq t$ then $p(\neg(A \vee B)|F) \geq t$.

$(\mathbf{C.}\neg)_p$ If $p(\neg A|F) < t$ then $p(\neg\neg A|F) = p(A|F) \geq t$.

It is straightforward to check the following.

Proposition 1. $(\mathbf{A.}\wedge)_p$, $(\mathbf{A.}\vee)_p$, $(\mathbf{B.}\neg)_p$, *and* $(\mathbf{C.}\neg)_p$ *hold in the sense of coherence-based probability logic. However,* $(\mathbf{C.}\wedge)_p$ *and* $(\mathbf{C.}\vee)_p$ *do not hold in this sense.*

Note that our probabilistic interpretation of the attack relation, justifies not only $(\mathbf{A.}\neg)$, but also the intuitively more demanding principle $(\mathbf{C.}\neg)_p$. This is a consequence of the fact that we insist on classical negation here and hence have $p(\neg\neg A) = p(A) = 1 - p(A)$. It might be worth mentioning that actually both $(\mathbf{B.}\neg)_p$ and $(\mathbf{C.}\neg)_p$ cease to hold if one admits .5 as a threshold value. Another interesting observation is that for $t = 1$ $(\mathbf{C.}\vee)$ is justified, since: if $p(\neg A|F) = 1$ and $p(\neg B|F) = 1$, then $p(\neg(A \vee B)|F) = p(\neg A \wedge \neg B|F) = 1$ is coherent (cf. the probabilistic version of the And rule of System P, [9]).

Interpreting attack principles involving the implication connective is more delicate, since it is widely agreed that the natural language conditional ('if ... then ...') should not be identified with classical (truth-functional) implication. Actually, as argued, e.g., in [10, 15], coherence-based conditional probability itself provides a sound and robust semantics for the conditional. Following this insight would force us to use degrees of beliefs in nested conditionals (e.g., in terms of previsions in conditional random quantities; see, e.g., [19, 20]) to interpret principles like $(\mathbf{A.}\supset)$. While this is an interesting topic for future research, here we only want to check how our probability-based interpretation of the attack relation classifies $(\mathbf{B.}\supset)$, $(\mathbf{A.}\supset)$, and $(\mathbf{C.}\supset)$, if we replace $A \supset B$ by $\neg A \vee B$. The corresponding translations are as follows:

$(\mathbf{A.}\supset)_p$ If $p(\neg B|F) \geq t$ then $p(\neg(A \supset B)|F) \geq t$.

$(\mathbf{B.}\supset)_p$ If $p(\neg B|F) \geq t$ and $p(\neg A|F) < t$, then $p(\neg(A \supset B)|F) \geq t$.

$(\mathbf{C.}\supset)_p$ If $p(\neg(A \supset B)|F) \geq t$ then $p(\neg B|F) \geq t$.

$A \supset B = \neg A \vee B$ turns $(\mathbf{A.}\supset)_p$ and $(\mathbf{C.}\supset)_p$ into instances of $(\mathbf{A.}\vee)_p$ and $(\mathbf{C.}\vee)_p$, respectively. Moreover, $(\mathbf{B.}\supset)_p$ follows from $(\mathbf{A.}\supset)_p$. Consequently we obtain:

Proposition 2. $(\mathbf{A.}\supset)_p$ *and* $(\mathbf{B.}\supset)_p$ *both hold in the sense of coherence-based probability logic, but* $(\mathbf{C.}\supset)_p$ *does not hold in this sense.*

In [3] also logically contradictory claims are considered by formulating the following corresponding attack principle:

(A.$\bot$) For every F: $F \longrightarrow \bot$.

In other words, it is stipulated that every argument (implicitly or explicitly) attacks contradictory claims. We may observe that this assumption is in line with our interpretation of the attack relation, since $p(\neg\bot|F) = 1$. However, note that we cannot interpret any principles that involve contradictory claims of attacking arguments, since the corresponding conditional probability must remain undefined.

4 Quantitative attack principles & their semantics

So far, we have only discussed qualitative attack principles, i.e., principles that only care for the presence or absence of an attack between (claims of) given arguments. However it is natural to refine such an analysis by considering *weights* or *varying strength* of attacks. Various suggestions regarding so-called weighted argumentation frames can be found in the literature on argumentation in AI, see, e.g., [8, 5]. But, similarly to the qualitative scenario, there is as yet hardly any analysis of rationality postulates that systematically relates weights of explicit and implicit attacks to the *logical form* of involved claims of arguments. A first step in that direction has been attempted in [4], where the principles introduced in [3] are generalized to the context of weighted argumentation frames. The aim of [4] is to explore under which assumptions one can characterize various t-norm based fuzzy logics in terms of 'weighted attack principles'. As expected, it turns out that some of the principles that are needed to recover a truth-functional (fuzzy) semantics are implausible from an intuitive, argumentation based point of view. In any case, the situation, once more, calls for a systematic interpretation of the relevant principles, that enables one to formally judge their respective plausibility.

Rather than just distinguishing between $F \longrightarrow A$ and $F \not\longrightarrow A$ ("F attacks / does not attack A"), we will use $F \xrightarrow{w} A$ to denote that F attacks A with weight (or degree) w. The corresponding 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. Note that this stipulation entails that the qualitative scenario discussed in sections 2 and 3 amounts to an instance of the weighted case, where the only possible weights are 0 and 1.

An attractive feature of the probabilistic approach taken here is the fact that it immediately leads to a quantitative refinement of the qualitative case: interpreting attacks in terms of coherent conditional probabilities suggests to directly attach weights, instead of using thresholds to judge whether a given statement attacks another one. As pointed out in [4], there are several non-equivalent ways in which the the qualitative attack principles reviewed in Section 2 can be generalized to 'weighted attack principles'. The most straightforward generalization of principle **(A.$\wedge$)** to weighted attacks is arguably the following:

$(\mathbf{A}^w.\wedge)$ If $F \xrightarrow{x} A$ and $F \xrightarrow{y} B$, then $F \xrightarrow{z} A \wedge B$, where $z \geq \max(x, y)$.

Actually, since we also consider attacks of weight 0 (interpreted as 'no attack'), we may assume without loss of generality that there is a weighted attack between any pair of formulæ. This means that $(\mathbf{A}^w.\wedge)$ can be reformulated as a constraint on the corresponding weights, s.t.:

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

Alternative weighted attack principles for conjunction, formulated in the same manner, are:

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

$(\mathbf{P}^w_\geq.\wedge)$ If $F \xrightarrow{x} A$, $F \xrightarrow{y} B$, and $F \xrightarrow{z} A \wedge B$, then $z \geq x + y - xy$.

As the labels indicate, these principles are essential for obtaining an argumentation based semantics for Gödel logic G, Łukasiewicz logic $\mathsf{Ł}$ and Product logic P, respectively. Moreover the subscript '$\geq$' attached to these letters indicate that upper bounds for the weight of attacks of conjunctive claims (in terms of weights of attacks on conjuncts) are formulated here. In fact, also principles expressing matching lower bounds are needed to characterize the three mentioned t-norm based fuzzy logics. Correspondingly, we use $(\mathbf{G}^w_\leq.\wedge)$, $(\mathbf{Ł}^w_\leq.\wedge)$, and $(\mathbf{P}^w_\leq.\wedge)$ to refer to the principles that arise by just replacing '$\geq$' by '$\leq$' in the respective constraint.

As already indicated, in contrast to the qualitative case of Section 3, we do not have to involve threshold values in interpreting a weighted attack relation, but simply identify the weight with which F attacks A with the conditional probability that A does not hold, given that F holds. More formally, our probabilistic semantics interprets $F \xrightarrow{w} A$ by $p(\neg A|F) = w$. (Remember that this is only viable if we exclude the possibility that F is a logical contradiction.) Accordingly, the above versions of weighted attack principles translate into the following statements.

$(\mathbf{G}^w_\geq.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \geq \max(x, y)$.

$(\mathbf{Ł}^w_\geq.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \geq \min(1, x + y)$.

$(\mathbf{P}^w_\geq.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \geq x + y - xy$.

$(\mathbf{G}^w_\leq.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \leq \max(x, y)$.

$(\mathbf{Ł}^w_\leq.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \leq \min(1, x + y)$.

$(\mathbf{P}^w_\leq.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \leq x + y - xy$.

According to our probability based interpretation we obtain the following classification of these principles.

Proposition 3. *The principles $(\mathbf{G}^w_\geq.\wedge)_p$ and $(\mathbf{Ł}^w_\leq.\wedge)_p$ hold in the sense of coherence-based probability logic. However, $(\mathbf{Ł}^w_\geq.\wedge)_p$, $(\mathbf{P}^w_\geq.\wedge)_p$, $(\mathbf{G}^w_\leq.\wedge)_p$, and $(\mathbf{P}^w_\leq.\wedge)_p$ do not hold for all coherent probability assessments.*

Proof. Remember that we assume that all involved propositions are classical. Therefore $\neg(A \wedge B)$ is equivalent to $\neg A \vee \neg B$, and hence the well known Fréchet inequalities (generalized to conditional probabilities) for logical disjunction yield $(\mathbf{G}^w_{\geq}\mathbf{.}\wedge)_p$ and $(\mathbf{Ł}^w_{\leq}\mathbf{.}\wedge)_p$.

The four other principles can all be violated:

$(\mathbf{Ł}^w_{\geq}\mathbf{.}\wedge)_p$, $(\mathbf{P}^w_{\geq}\mathbf{.}\wedge)_p$: Let $A = B$ and $p(\neg A|F) = p(\neg B|F) = 0.5$. Then $p(\neg(A \wedge B)|F) = p(\neg(A \wedge A)|F) = p(\neg A|F) = 0.5$, which is strictly smaller than $\min(1, 0.5 + 0.5) = 1$, but also strictly smaller than $0.5 + 0.5 - 0.5^2 = 0.75$.

$(\mathbf{G}^w_{\leq}\mathbf{.}\wedge)_p$, $(\mathbf{P}^w_{\leq}\mathbf{.}\wedge)_p$: Let $A = \neg B$ and $p(\neg A|F) = p(\neg B|F) = 0.5$. Then $p(\neg(A \wedge B)|F) = p(\neg(A \wedge \neg A)|F) = p(\neg\bot|F) = p(\top|F) = 1$, which is strictly larger than $\max(0.5, 0.5) = 0.5$ and strictly larger than $0.5 + 0.5 - 0.5^2 = 0.75$. □

Note that $(\mathbf{G}^w_{\geq}\mathbf{.}\wedge)_p$ and $(\mathbf{Ł}^w_{\leq}\mathbf{.}\wedge)_p$ define the best possible coherent lower and upper bounds, respectively. The principles $(\mathbf{Ł}^w_{\geq}\mathbf{.}\wedge)_p$, $(\mathbf{P}^w_{\geq}\mathbf{.}\wedge)_p$, $(\mathbf{G}^w_{\leq}\mathbf{.}\wedge)_p$, and $(\mathbf{P}^w_{\leq}\mathbf{.}\wedge)_p$, which do not hold under coherence, are not simply unjustifiable from a probabilistic point of view. They rather apply only to specific cases. The following corresponding propositions are straightforward.

Proposition 4. *Under the assumption that $p(A|F)$ and $p(B|F)$ are independent, $(\mathbf{P}^w_{\geq}\mathbf{.}\wedge)_p$ and $(\mathbf{P}^w_{\leq}\mathbf{.}\wedge)_p$ hold.*

Proposition 5. *Under the assumption that $A \models B$ or $B \models A$ $(\mathbf{G}^w_{\leq}\mathbf{.}\wedge)_p$ holds.*

Proposition 6. *Under the assumption that $A \models \neg B$ or $B \models \neg A$ $(\mathbf{Ł}^w_{\leq}\mathbf{.}\wedge)_p$ holds.*

The picture obtained for attack principles involving disjunction is, of course, dual to that just outlined for conjunction. The Fréchet inequalities justify the following two principles:

$(\mathbf{G}^w_{\leq}\mathbf{.}\vee)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \vee B)|F) \leq \min(x, y)$.

$(\mathbf{Ł}^w_{\geq}\mathbf{.}\vee)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \vee B)|F) \geq \max(0, x+y-1)$.

Other principles are justified according to the probabilistic semantics of argument attack only under additional assumptions about the (in)dependence of involved propositions.

For negation the probability semantics directly justifies the following attack principle, that combines and generalizes the qualitative principles $(\mathbf{A.}\neg)$ and $(\mathbf{C.}\neg)$.

$(\mathbf{AC}^w\mathbf{.}\neg)$ $F \xrightarrow{x} A$ if and only $F \xrightarrow{1-x} \neg A$.

Regarding implication, one may of course extract corresponding principles from the above mentioned ones, under the stipulation that $A \supset B$ is understood, classically, as equivalent to $\neg A \vee B$. But, once more, let us emphasize that it were actually more adequate to model (informal) implication as a conditional. This leads to the tricky and, as yet, only partially explored terrain of iterated conditional probabilities; thus providing a challenging topic for future research.

Draft names	Task/argument form	Task
$(\mathbf{A.}\wedge)$	if $A \xrightarrow{x} B$, then $A \xrightarrow{[x,1]} (B \wedge C)$	B2,C4
$(\mathbf{C.}\wedge)$	if $A \xrightarrow{x} (B \wedge C)$, then $A \xrightarrow{[0,x]} B$	A1,C1
$(\mathbf{A.}\vee)$	if $A \xrightarrow{x} (B \vee C)$, then $A \xrightarrow{[x,1]} B$	A2,C3
$(\mathbf{C.}\vee)$	if $A \xrightarrow{x} B$, then $A \xrightarrow{[0,x]} (B \vee C)$	B3,C6
Irrelevant premise	if $A \xrightarrow{x} B$ and $C \models B$ then $A \xrightarrow{x} B$	A3,C5
$(\mathbf{B.}\neg')$	if $A \xrightarrow{x} B$, then $A \xrightarrow{1-x} \neg B$	B1,C2
Complement	if $A \xrightarrow{x} \neg B$, then $A \xrightarrow{1-x} B$	A4,C7
$(\mathbf{B.}\neg)$	"if $A \longrightarrow B$, then $\neg(A \longrightarrow \neg B)$" is true	B11,C18
$(\mathbf{B.}\neg'')$	"if $A \longrightarrow \neg B$, then $\neg(A \longrightarrow B)$" is true	B12,C19
Narrow negation	if $A \xrightarrow{x} B$, then $A \xrightarrow{1-x} \neg B$	A7, B5,C11
$(\mathbf{A.\bot})$	$A \xrightarrow{1} (B \wedge \neg B)$	B4,C9
$(\mathbf{A.\top})$	$A \xrightarrow{0} (B \vee \neg B)$	B8,C15
Aristotle's thesis 1	$\neg(\neg A \longrightarrow A)$ is false	B6,C12
Aristotle's thesis 2	$\neg(A \longrightarrow \neg A)$ is false	A5,C8
Abelard's thesis	$\neg((A \longrightarrow B) \wedge (A \longrightarrow \neg B))$ is true	B7,C14
Reflexivity	$A \xrightarrow{0} A$	A6,C10
Contingent attack	$A \xrightarrow{[0,1]} B$	A8,C13
ProbToAttack	if $P(B\|A) = x$, then $A \xrightarrow{x} \neg B$	A10,B9,C17
AttackToProb	if $A \xrightarrow{x} B$, then $P(\neg B\|A) = x$	A9,B10
AttackToProb'	if $A \xrightarrow{x} B$, then $P(B\|A) = 1 - x$	C16
ProbToAttack'	if $P(B\|A) = x$, then $A \xrightarrow{1-x} B$	C17

Table 1: Task names/argument forms of the task sets with closed (i.e., conditions A and B) and open (i.e., C) response format. "$A \xrightarrow{x} B$" denotes "*A attacks by strength x the assertion B*", where x can be point- or interval-valued.

5 Experiment

In this section we explore the psychological plausibility of the proposed approach. Coherence-based probability logic received empirical support in recent years (e.g., [14, 16, 17, 18]). However, principles governing the strength of attacks have not yet been investigated empirically (neither within nor outside the coherence framework).

Participants The sample consists of 139 students of the Technical University of Vienna (18 females, 116 males, and 5 who chose not to reveal their gender) with a mean age of 21.1 years ($SD = 3.2$). Only German native speakers were included in the data analysis. Seven participants were excluded from the analysis because of missing data in the target tasks. Most students were in their second semester and did not receive a thorough training in logic yet.

On the average, the participants rated the overall task clearness and difficulty on an intermediate level ($M = 4.9$ and $M = 4.3$, respectively, on a rating scale out of 10). This reflects the fact that since our study aims to explore the interpretation of attack principles, the participants had first to reason towards how to interpret the tasks and then, after fixing their interpretation, to draw conclusions based on their interpretation. This can also explain why the participants were not highly confident in the correctness of their solutions ($M = 4.1$ out of 10) even if in general they tend to like solving mathematical puzzles ($M = 7.5$ out of 10).

Method and materials Each participant was administered a DIN-A4 page, containing an introduction on the first page and the target tasks on both pages. There were three between-participant conditions, two with multiple-choice (A: $n_1 = 44$ and B: $n_2 = 48$) and one with an open choice response format (C: $n_3 = 47$). After showing how to express the degree of attack from a scale form 0 to 10 and that claims can also be compounded (like [**A and B**]), the participants were presented with those tasks which are described in Table 1. For example, Task A1 presents the antecedent of a conditional: "If **A** attacks with **exactly** the strength **7** the claim **B**, then ...". Then seven consequent candidates were presented, which completed the conditional. Eight consequents were of the form "...attacks **A** with [M] with the strength [S] the claim **B**", where "[M]" indicates a precise value ("exactly"), a lower ("at least"), or an upper bound ("at most") on the strength [S]. [S] was either 0, 3, 7, or 10. All possible point and interval options were formulated in ascending order (see Table 2 for the attack strength options we used). Except for the interval $[0, 10]$ we used "nothing follows about how strong ...attacks ...", as the ninth response option within each task. The participants were asked to tick for each of nine items whether the according sentence is correct ("*richtig*") or false ("*falsch*"). In the open response format condition C, the participants were instructed to fill in "exactly", "at least", or "at most", the value of the strength, and additionally had to mark the strength of attack (either as a point value or an interval) on a scale as introduced in the introduction. In all conditions, those tasks which were not formulated directly in terms of a conditionals, the instruction required to choose among "true", "false", or "undetermined" by ticking one corresponding box (e.g, A6, B4, B6, or B11 ; see Table 1).

The experiment took place during the last part of the first lecture on "formal modeling". The three conditions were administered in a systematically alternated way to reduce the chance of plagiarized responses.

Results and discussion The main results are presented in tables 2–6. First we observe that most people are unaware of the best possible coherent bounds (marked in **bold**). Responses which are within the optimal coherent bounds are of course also coherent, like in task A1 where 45% of the participants responded that "precisely 7" is correct. In this task, 43% responded that the interval "at most 7" is correct, which corresponds to the coherent interval. Second, we observe that compared to direct tests of coherence-based probability logic (e.g., [14, 16, 17, 18]), the agreement between the predictions concerning the quantitative attack principles and the participant's responses are modest, especially for the conditions with closed response formats (A and B). For the condition C, more than half of the participants responded by at least a coherent lower or a coherent upper bound as predicted (see

Task	0	[0,3]	3	[0,7]	[3,10]	7	[7,10]	10	nf
A1	0.00	0.00	0.00	**43.18**	18.18	45.45	18.18	0.00	31.82
A2	0.00	0.00	0.00	63.64	6.82	25.00	**9.09**	0.00	34.09
A3	0.00	2.27	0.00	25.00	18.18	**93.18**	27.27	0.00	4.55
A4	20.45	18.18	**18.18**	11.36	2.27	2.27	0.00	0.00	59.09
A7	15.91	22.73	**20.45**	13.64	6.82	9.09	0.00	0.00	52.27
A8	6.82	4.55	4.55	6.82	4.55	4.55	4.55	4.55	**88.64**
A9	2.27	13.64	22.73	2.27	9.09	**13.64**	6.82	4.55	56.82
A10	4.55	4.55	**13.64**	2.27	9.09	11.36	11.36	2.27	63.64

Table 2: Percentages of "correct" responses concerning the point valued/interval attack strength options in condition A ($n_1 = 44$). The response options of A9 were normalized to probability values. "nf" denotes "nothing follows". Best possible coherent response options are in **bold** (for predictions see Table 1).

Task	0	[0,3]	3	[0,7]	[3,10]	7	[7,10]	10	nf
B1	8.33	31.25	**29.17**	2.08	4.17	2.08	0.00	0.00	43.75
B2	2.08	4.17	2.08	22.92	18.75	16.67	**20.83**	0.00	39.58
B3	2.08	4.17	2.08	**27.08**	18.75	25.00	33.33	4.17	27.08
B5	8.33	31.25	**29.17**	0.00	4.17	0.00	2.08	4.17	45.83
B9	4.17	14.58	16.67	8.33	0.00	**2.08**	4.17	0.00	62.50
B10	2.08	12.50	14.58	8.33	4.17	**20.83**	4.17	0.00	47.92

Table 3: Percentages of "correct" responses in condition B ($n_2 = 48$). The response options of B10 were normalized to probability values. See also caption of Table 2.

median values in Table 5). Concerning the seven forced choice tasks in condition C, the most frequent responses were consistent with our coherence-based predictions in five tasks (see Table 6). In tasks C9 and C15 people chose incoherent responses, which involve contradictions and tautologies, which appear difficult to interpret in the context of principles about argument strength. We observed an analogous effect in the corresponding tasks B4 and B8 in the closed response format condition (see Table 4).

The Contingent attack tasks serve to check whether people read the tasks carefully. The Irrelevant premise task was intended to test (**A.gen**) but due to a systematic error in the translation of this argument form into the corresponding tasks, we use it now as a consistency check. In both tasks almost all participants responded as expected. The results of those tasks, which serve to explore directly the connection between probability and strength of attack (i.e., ProbToAttack, AttackToProb, and AttackToProb) were disappointing in the closed response format conditions A and B. In the open response format task C16, which investigates AttackToProb, the majority of participants responded as predicted. In task C17, which investigates ProbToAttack, the majority of only the lower bound responses were coherent. Again, participants scored better in the open response format condition compared to the closed one.

	A5	A6	B4	B6	B7	B8	B11	B12
false	**43.18**	**40.91**	31.25	**47.92**	41.67	**16.67**	31.25	31.25
correct	31.82	22.73	**25.00**	35.42	**31.25**	56.25	**39.58**	**35.42**
undetermined	25.00	36.36	43.75	16.67	27.08	27.08	29.17	33.33

Table 4: Percentages of responses in conditions A ($n_1 = 44$) and B ($n_2 = 48$). Best possible coherent response options are in **bold** (see Table 1).

	C1l	C1u	C2l	C2u	C3l	C3u	C4l	C4u	C5l	C5u	C6l
	0	**.70**	**.30**	**.30**	**.70**	**1**	**.70**	**1**	**.70**	**.70**	**0**
a	.42	.70	.16	.43	.25	.74	.37	.83	.63	.73	.31
b	.33	.20	.20	.36	.33	.18	.33	.22	.22	.08	.35
c	.70	.70	.00	.30	.00	.70	.30	1.00	.70	.70	.00

	C6u	C7l	C7u	C11l	C11u	C13l	C13u	C16l	C16u	C17l	C17u
	.70	**.30**	**.30**	**.30**	**.30**	**0**	**1**	**.30**	**.30**	**.30**	**.30**
a	.77	.17	.51	.14	.48	.24	.92	.31	.51	.28	.57
b	.18	.21	.40	.19	.38	.33	.21	.33	.34	.28	.34
c	.70	.00	.30	.00	.30	.00	1.00	.30	.30	.30	.70

Table 5: Mean (a), standard deviations (b), and medians (c) of lower (L) and upper (U) bound responses in condition C ($n_3 = 47$). Except for the probability responses to task C16, all values are normalized to the value range $[0, 1]$. Best possible coherent response options are in **bold** (see Table 1).

6 Concluding remarks

We showed how the coherence approach to probability can serve to guide the rational selection of qualitative and quantitative attack principles. More research is needed to deepen and to generalize our formal results: e.g., by interpreting implication by conditional probability (or by previsions in conditional random quantities) or by generalizations to fuzzy events. We also presented an experiment to explore the psychological plausibility of the proposed approach. While we are convinced that our approach is intuitive and plausible, we were surprised by the relatively heterogeneous results. Open response format tasks turned out the be more appropriate to investigate quantitative attack principles. 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, 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). Future experimental work is needed to further explore the psychological plausibility of formal attack principles.

References

[1] P. Besnard and A. Hunter. *Elements of argumentation.* MIT Press Cambridge, 2008.

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

	C8	C9	C10	C14	C15	C18	C19
false	**51.06**	23.40	**76.60**	29.79	**14.89**	17.02	12.77
true	19.15	**27.66**	8.51	**42.55**	34.04	**48.94**	**53.19**
undetermined	29.79	48.94	14.89	27.66	51.06	34.04	34.04

Table 6: Percentages of responses to forced choice tasks in condition C ($n_3 = 47$). Best possible coherent response options are in **bold** (see Table 1).

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

[4] E.A. Corsi and C.G. Fermüller. Connecting fuzzy logic and argumentation frames via logical attack principles, submitted.

[5] Sylvie Coste-Marquis, Sébastien Konieczny, Pierre Marquis, and Mohand Akli Ouali. Weighted attacks in argumentation frameworks. In *13th International Conference on the Principles of Knowledge Representation and Reasoning (KR'12)*. AAAI Press, 2012.

[6] P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intelligence*, 77(9):321–357, 1995.

[7] J Michael Dunn and Greg Restall. Relevance logic. In *Handbook of Philosophical Logic*, volume 6, pages 1–128. Springer, 2002.

[8] Paul E Dunne, Anthony Hunter, Peter McBurney, Simon Parsons, and Michael Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2):457–486, 2011.

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

[10] A. Gilio, N. Pfeifer, and G. Sanfilippo. Transitivity in coherence-based probability logic. *Journal of Applied Logic*, 14:46–64, 2016.

[11] U. Hahn and M. Oaksford. The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, 114(3):704–732, 2007.

[12] Ingebrigt Johansson. Der Minimalkalkül, ein reduzierter intuitionistischer Formalismus. *Compositio Mathematica*, 4:119–136, 1937.

[13] M. Oaksford, N. Chater, and U. Hahn. Human reasoning and argumentation: The probabilistic approach. In J. Adler and L. Rips, editors, *Reasoning: Studies of human inference and its foundations*. Cambridge University Press, Cambridge, in press .

[14] N. Pfeifer. The new psychology of reasoning: A mental probability logical perspective. *Thinking & Reasoning*, 19(3–4):329–345, 2013.

[15] N. Pfeifer. Reasoning about uncertain conditionals. *Studia Logica*, 102(4):849–866, 2014.

[16] N. Pfeifer and G. D. Kleiter. Coherence and nonmonotonicity in human reasoning. *Synthese*, 146(1-2):93–109, 2005.

[17] N. Pfeifer and G. D. Kleiter. Framing human inference by coherence based probability logic. *Journal of Applied Logic*, 7(2):206–217, 2009.

[18] N. Pfeifer and L. Tulkki. Conditionals, counterfactuals, and rational reasoning. An experimental study on basic principles. *Minds and Machines*, 27(1):119–165, 2017.

[19] G. Sanfilippo, N. Pfeifer, and A. Gilio. Generalized probabilistic modus ponens. In A. et al. Antonucci, editor, *ECSQUARU 2017*, volume 10369 of *LNCS*, pages 480–490. Springer, 2017.

[20] G. Sanfilippo, N. Pfeifer, D. E. Over, and A. Gilio. Probabilistic inferences from conjoined to iterated conditionals. *International Journal of Approximate Reasoning*, 93:103–118, 2018.

[21] F. Zenker, editor. *Bayesian argumentation: The practical side of probability.* Synthese Library (Springer), Dordrecht, 2013.