Running head: NONMONOTONICITY AND HUMAN PROBABILISTIC REASONING

Nonmonotonicity and Human Probabilistic Reasoning

Niki Pfeifer

Department of Psychology

University of Salzburg

Austria

Gernot D. Kleiter

Department of Psychology

University of Salzburg

Austria

Draft of January 5, 2007

Abstract

Nonmonotonic reasoning allows—contrary to classical (monotone) logic—for withdrawing conclusions in the light of new evidence. Nonmonotonic reasoning is often claimed to mimic human common sense reasoning. Only a few studies, though, have investigated this claim empirically. SYSTEM P is a central, broadly accepted nonmonotonic reasoning system that proposes basic rationality postulates for reasoning about nonmonotonic conditionals. Nonmonotonic conditionals are "if–then" relations for the representation of relations that hold normally or by default. We interpret nonmonotonic conditionals by "high" conditional probabilities, $P(B|A) > .5$.

This study reports a series of four experiments on probabilistic reasoning with inference rules about nonmonotonic conditionals, namely the CAUTIOUS MONOTONICITY, CUT, and the RIGHT WEAKENING rule of SYSTEM P. As a critical condition, we investigated central monotonic properties of classical (monotone) logic, namely MONOTONICITY, TRANSITIVITY, and CONTRAPOSITION. In accordance with previous results, the present study suggests that people reason nonmonotonically rather than monotonically. We propose nonmonotonic reasoning as a competence model of human common sense reasoning.

Nonmonotonicity and Human Probabilistic Reasoning

Introduction

Traditionally, psychological theories on deductive reasoning evaluate the quality of human reasoning by *classical logic* as the normative standard of reference (Braine & O'Brien, 1998; Rips, 1994; Johnson-Laird, 1983). Classical logic has been proposed by Macnamara (1986) as the surest guide towards a competence model for the psychology of reasoning. As compared with the actual reasoning performance (which is biased by memory limitations, limited information processing resources, shifts of attention, etc.), competence refers to ideal reasoning performance. This distinction, here in the domain of reasoning, is analog to the performance/competence distinction in the domain of language introduced by Chomsky (e.g., 1965).

The *psychological plausibility* of classical logic both, as the normative standard of reference and as a competence model, is questionable on *a priori* grounds for several reasons. The two most important reasons are the monotonicity principle and the definition of the "if—then" relation. The monotonicity principle inherent in classical logic does not allow for retracting conclusions in the light of new evidence, while the studies on the suppression of conditional inferences impressively show that subjects are willing to doubt in premises or to withdraw conclusions under certain circumstances (Byrne, 1989; Byrne, Espino, & Santamaría, 1999; Bonnefon & Hilton, 2002, 2004; Dieussaert, De Neys, & Schaeken, 2005; Politzer, 2005). The "if—then" relations as defined in classical logic do not allow dealing with exceptions and uncertainty, while exceptions and uncertainty can almost always be present in common sense reasoning. Thus, e.g., Liu, Lo, and Wu (1996) observed that a high percentage (90%) of the subjects attached a probabilistic interpretation to indicative conditionals. Hadjichristidis et al. (2001), Evans, Handley,

and Over (2003), Over and Evans (2003), and Oberauer and Wilhelm (2003) report data that indicate that subjects interpret common sense indicative conditionals as conditional probabilities.

The present paper proposes to use a probabilistic interpretation of *nonmonotonic reasoning* (Gilio, 2002) as the normative standard of reference for investigating human reasoning, rather than classical logic. This does not mean to abandon logic from psychology, rather then, to enrich the traditional normative standard of reference by nonmonotonic tools to handle uncertainty and the retraction of conclusions. The psychological plausibility of the proposed normative standard of reference is supported in previous studies (Pfeifer & Kleiter, 2005a) and will be supported by four experiments reported in the following sections.

To illustrate what nonmonotonicity means and why it matters, consider the following argument:

(A1) Birds can fly. Therefore, things that are birds and penguins can fly.

A student of classical logic might formalize argument (A1) in the following way:

(CL) $B \rightarrow F \quad \therefore \quad B \wedge P \rightarrow F$,

where $\rightarrow$ ("if–then", material implication), $\therefore$ ("therefore", logical entailment), and $\wedge$ ("and", conjunction) are defined as usual in classical logic. If (A1) is interpreted as (CL), then argument (A1) is a logically valid argument which means that the conclusion ($B \wedge P \rightarrow F$) follows necessarily from the premise ($B \rightarrow F$).

This interpretation is strange however since the premise "Birds can fly" is plausible, while the conclusion "Things that are birds and penguins can fly" is not plausible. By *reductio ad absurdum* one might suggest to reject the premise "Birds can fly". Then, one might replace the premise by "Birds that are not penguins can fly", but this is not satisfactory: what about oistrichs, birds with malfunctioning wings, drunken birds, etc.?

Since there are infinitely many types of exceptions, it is impossible to list all of them explicitly in the premises. In artificial intelligence this problem is called the "frame problem". By "Birds can fly" we do not mean *all* birds fly, rather we mean something like "Birds normally can fly", which allows for *exceptions*.

Nonmonotonic reasoning interprets sentences like "Birds can fly" by formalizing the common sense conditional by the nonmonotonic conditional. The nonmonotonic conditional is denoted by "$\mid\!\sim$" and is read as "*if . . . normally . . .*". Nonmonotonic reasoning does not sanction inferences like in (A1), however, it does sanction a more *cautious* form. In the more cautious form (A1) would be reformulated as

> (A2) Birds can fly. Birds are penguins. Therefore, things that are birds and penguins can fly.

Here, the second premise blocks the inference: since most birds are not penguins the second premise is not plausible, and therefore the conclusion "Birds and penguins can fly" is not guaranteed. Argument (A1) has the problem that all premises are plausible while the conclusion is not plausible. This problem is not present in argument (A2). Argument (A2) is an instance of a nonmonotonic inference rule called "CAUTIOUS MONOTONICITY".

Nonmonotonic reasoning is a promising candidate towards a competence theory of human reasoning that captures both, retraction of conclusions in the light of new evidence and rule guided reasoning. We will present a probabilistic semantics for nonmonotonic reasoning that is based on coherence (Gilio, 2002; Coletti & Scozzafava, 2002, 2005), and report empirical data that investigate its psychological plausibility. One advantage of the probabilistic approach is that it can easily be related to the extensive tradition of *judgment under uncertainty* in psychology (Gilovich, Griffin, & Kahneman, 2002).

Coherence and Nonmonotonic Reasoning

Probability assessments are coherent if they cannot lead to bets with sure losses. An example of a violation of coherence is the conjunction fallacy. The standard task is the Linda Task where about 80% of human subjects commit this fallacy (Tversky & Kahneman, 1983). In this task the subjects rank "Linda is active in the feminist movement and Linda is a bank teller" as more probable than just "Linda is a bank teller". If the subjects interpret the former sentence as a conjunction, then the subjects clearly violate probabilistic coherence. Specifically, the upper probability bound ($min(P(A), P(B))$) is violated, since such a ranking implies $P(A \wedge B) > \min(P(A), P(B))$ which is incoherent, of course. Empirical data of how humans handle the lower bound in a conjunction problem ($max(0, P(A) + P(B) - 1)$) are reported by Pfeifer and Kleiter (2005a).

The coherence interpretation of probability is in the tradition of subjective probability theory, stating that coherent probabilities are *degrees of belief* and not something "objective" in the outside world. "Probability does not exist" (De Finetti, 1974). Degrees of belief are naturally "affine" to psychology. Moreover, coherent conditional probability, $P(B|A)$, is a *primitive* notion. The probability values are assigned *directly*. Coherent conditional probability is not defined—as it is done in the standard concept of conditional probabilities—via the fraction of the "joint" ($P(A \wedge B)$) and the "marginal" ($P(A)$) probabilities,

$$P(B|A) =_{def.} \frac{P(A \wedge B)}{P(A)}\,.$$

If the conditioning event has a probability equal to zero ($P(A) = 0$), then unjustified and incoherent[1] ad-hoc assumptions are adopted in the standard probability approach, since dividing by zero is not defined. The concept of coherent conditional probability avoids such incoherent assumptions. The only requirement on the conditioning event is that it must not be an impossible event. Thus, the extreme probability values 0 and 1 are treated in a natural way. Finally, a complete Boolean algebra is is not required at the very beginning of a probabilistic assessment in the framework of coherence. Only the informations available (i.e., logical and/or probabilistic) are used and then an algebra is generated to the extent needed for the assessment. For introductory notes, recent formal results on combining probability, possibility, fuzzy sets and Spohn's ranking functions in a unified framework under coherence see Coletti and Scozzafava (2002, 2004, 2005).

Gilio (2002) developed a probability semantics for nonmonotonic reasoning that is based on the principle of coherence. We explain the way in which a nonmonotonic inference is cast into a probabilistic format by the standard example of nonmonotonic reasoning.

> Tweety is a bird, and as we know that birds normally fly, we conclude that Tweety presumably flies. When we now learn that Tweety is a penguin, common sense reasoning tells us to retract our earlier, tentative, conclusion that Tweety flies.

Insert Table 1 about here

Table 1 contains a probabilistic version of the Tweety example. The important point is that the preliminary conclusion C1 is retracted in the light of new evidence P3 and P4. This can never happen in classical monotonic inference.

The Tweety example involves three binary events. To specify the joint probability distribution of the eight possible constituents, seven point probabilities would be needed. The eighth probability would result from the constraint that the sum of all eight probabilities is one. As only two probabilities are given, this is a task with incomplete information. The condition of incomplete information is an essential property of nonmonotonic reasoning tasks. This is also the reason why upper and lower probabilities enter the probabilistic representation and not just point probabilities. Since the resources of human information processing are limited, the capacity of dealing with incomplete information is necessary for a competence theory of human reasoning.

While there are many systems of nonmonotonic reasoning[2], SYSTEM P is a basic set of rationality postulates every system of nonmonotonic reasoning should satisfy (Kraus, Lehmann, & Magidor, 1990). Thus, SYSTEM P is broadly accepted in artificial intelligence by the communities that work on nonmonotonic reasoning. For just this system Gilio (2002) gave a probability semantics. We will be concerned with this basic system. We denote the nonmonotonic conditional by "$\mid\!\sim$". "$A \mid\!\sim B$" is read as "B follows normally from A", or short "if A, normally B"; A and B are placeholders for declarative sentences. $\mid\!\sim$ is a *genuine* nonmonotonic operator. $A \mid\!\sim B$ is to be interpreted as the *nonmonotonic conditional* with antecedent (A) and consequent (B). $\mid\!\sim$ is the characteristic junctor which has the least binding strength and may appear at most one time in each formula. Contrary to the material implication $\rightarrow$, the nonmonotonic conditional $\mid\!\sim$ cannot be iterated in a formula. Thus, e.g., $A \rightarrow B \rightarrow C$ is a well formed formula, while $A \mid\!\sim B \mid\!\sim C$ is not a well formed formula.

The two most important semantic families for SYSTEM P are (i) the *preferred model semantics* (Kraus et al., 1990), which is a special kind of possible world semantics, and (ii) the *probability semantics* (Adams, 1975; Goldszmidt & Pearl, 1996; Schurz, 1997; Gilio, 2002; Biazzo, Gilio, Lukasiewicz, & Sanfilippo, 2002; Lukasiewicz, 2005).

In psychology theories of concept representation by prototypes and typicality are closely related to preferred model semantics (Smith & Medin, 1981; Murphy, 2002). The probabilistic approach to human judgment and decision making under uncertainty has a tradition of over half a century now (Kahneman, Slovic, & Tversky, 1982; Gigerenzer, Todd, & the ABC Research Group, 1999; Gilovich et al., 2002). Chater and Oaksford (2001, Oaksford & Chater, 1998) adopted with their probability heuristics model the probabilistic approach to the psychology of reasoning. Recently, studies on subjects' interpretation of "if–then" statements and conditional inferences were conducted from a probabilistic perspective (Liu et al., 1996; Oaksford, Chater, & Larkin, 2000; Liu, 2003; Oberauer & Wilhelm, 2003; Over & Evans, 2003; Evans et al., 2003) which provide support in favor of the probabilistic interpretation of the "if–then" (especially as a conditional probability) and clearly indicate that the subjects do not interpret the "if–then" as a material conditional as defined in logic.

Probability semantics defines $A \mathrel{|\!\sim} B$ within a probability model, such that the conditional probability $P(B|A)$ is "high", that is

$$A \mathrel{|\!\sim} B \quad \text{is interpreted as} \quad P(B|A) > .5\,.$$

The conditional probability $P(B|A)$ is clearly not generally equivalent to the probability of the material implication $P(A \rightarrow B)$. Basically, while the material implication $(A \rightarrow B)$ can be represented by a disjunction $(\neg A \vee B)$ or by a negated conjunction $(\neg(A \wedge \neg B))$, the conditional event $B|A$ cannot be represented by any logical operator.

Let A_i and B_i denote two events and $A_i|B_i$ denote the conditional event A_i *given* B_i. We denote by $\left|A_i\right|$, $\left|B_i\right|$, and $\left|A_i|B_i\right|$ the *indicator function* of the respective sentences, mapping the truth values of the sentences into numbers, typically 1 ("true") and 0 ("false"). What are the truth values of $A_i|B_i$?

If the conditioning sentence is false, then the indicator value of $A_i|B_i$ is the probability of A_i (Coletti & Scozzafava, 2002; Gilio, 2002):

$$\big|A_i|B_i\big| = \begin{cases} 1 & : \text{ if } A_i = 1 \text{ and } B_i = 1, \\ 0 & : \text{ if } A_i = 0 \text{ and } B_i = 1, \\ P(A_i) & : \text{ if } B_i = 0. \end{cases}$$

An obvious consequence is that conditional events cannot be treated by the usual operators of negation, conjunction, and disjunction. There is no logical operator of conditioning. "Logic lacks a conditioning operator corresponding to conditional probability" (Goodman, Nguyen, & Walker, 1991). This is a fundamental property (Lewis triviality result), that distinguishes conditioning from material implication. Thus, the probabilistic versions of inference rules formalizing the conditional as material implication differ from the inference rules using conditional probabilities in the propagation rules, compare, e.g., the material probabilistic version of the MODUS PONENS,

from $P(A \to B) = x$ *and* $P(A) = y$ *infer* $P(B) \in [\max(0, x + y - 1), x]$,

and the corresponding conditional probabilistic version (Pfeifer & Kleiter, 2005b),

from $P(B|A) = x$ *and* $P(A) = y$ *infer* $P(B) \in [xy, 1 - y + xy]$.

The upper and lower probabilities can be determined by elegant or by general methods of linear algebra. If the conditioning event turns out to be false we do not learn anything about the probability of A_i. The probability is equal to its "base rate". It remains the same as if we would know nothing about B_i. Expressed in terms of bets, the bet is called off and the prize of the bet is payed back. A similar proposal was originally made by Ramsey in 1929, which is known as the *Ramsey test*:

> "If two people are arguing "If p will q?" and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that

> basis about q; [...] We can say they are fixing their degrees of belief in q given p. If p turns out false, these degrees of belief are rendered *void*." (Ramsey, 1994, Footnote, p. 155)

Evans et al. (2003); Evans and Over (2004) and Over and Evans (2003) adopted the Ramsey test as a psychological account of how humans interpret and evaluate common sense indicative conditionals.

There are, roughly speaking, three different kinds of probability semantics for reasoning about nonmonotonic conditionals in SYSTEM P, *model-theoretic probabilistic logic*, *infinitesimal* and *non-infinitesimal* probability semantics (Adams, 1975; Schurz, 1997; Gilio, 2002; Biazzo et al., 2002; Lukasiewicz, 2005). Infinitesimal semantics requires probabilities θ infinitesimally close to 1. Psychologically, infinitesimals are hard to communicate to the subjects. Non-infinitesimal probability semantics requires only *practically high* probabilities, e.g., $\theta > .5$. More specifically, the nonmonotonic conditional is then written as "$A \mid\!\sim_x B$", where x is an element of the *interval* $[x', x'']$, that is

$$A \mid\!\sim_x B \quad \text{is interpreted as} \quad P(B|A) \in [x', x''] \,,$$

where x' is the *lower* and x'' is the *upper bound* of the probability interval $[x', x'']$, and $0 \leq x' \leq x'' \leq 1$. Point probabilities are treated as special cases of intervals, such that $x' = x''$. Our studies will focus on a non-infinitesimal probability semantics only.

The rules of SYSTEM P serve as a minimal set of basic rationality postulates for nonmonotonic reasoning. They are

- REFLEXIVITY (Axiom): $A \mid\!\sim A$,
- CAUTIOUS MONOTONICITY: *from* $A \mid\!\sim C$ *and* $A \mid\!\sim B$ *infer* $A \wedge B \mid\!\sim C$,
- RIGHT WEAKENING: *from* $A \mid\!\sim B$ *and* $\models B \rightarrow C$ *infer* $A \mid\!\sim C$,
- CUT: *from* $A \mid\!\sim B$ *and* $A \wedge B \mid\!\sim C$ *infer* $A \mid\!\sim C$,
- LEFT LOGICAL EQUIVALENCE *from* $\models A \leftrightarrow B$ *and* $A \mid\!\sim C$ *infer* $B \mid\!\sim C$,

- OR: *from* $A \mathrel{|\!\sim} C$ *and* $B \mathrel{|\!\sim} C$ *infer* $A \vee B \mathrel{|\!\sim} C$, and,
- AND rule (derived rule): *from* $A \mathrel{|\!\sim} B$ *and* $A \mathrel{|\!\sim} C$ *infer* $A \mathrel{|\!\sim} B \wedge C$.

"$\mathrel{|\!\sim}$" denotes the nonmonotonic conditional, "$\models$" denotes logical validity, "$\leftrightarrow$" ("if and only if") denotes the material equivalence , "$\rightarrow$" ("if–then") the material implication, "$\vee$" ("or") the disjunction and "$\wedge$" ("and") denotes the conjunction, which are defined as usual in classical logic. REFLEXIVITY is the only axiom in the system. Because of its triviality we discard it here and focus on the inference rules only.

The present paper focuses on empirical investigations of the CAUTIOUS MONOTONICITY rule, CUT rule, and of the RIGHT WEAKENING rule of SYSTEM P. For the critical test, whether subjects endorse nonmonotonic inference rules and do not endorse monotonic inference argument forms, we investigate three basic properties of classical (monotone) logic. These properties are MONOTONICITY, TRANSITIVITY, and CONTRAPOSITION. They are, of course, not valid in SYSTEM P. Table 2 presents the SYSTEM P rules used in the present paper and the monotonic argument forms, and gives an overview of the series of experiments reported below.

Gilio (2002) proved propagation rules for the probability bounds in SYSTEM P. The propagation rules specify how the coherent probability bounds in the *conclusion* ($z \in [z', z'']$) can be inferred from the coherent upper and lower probability bounds in the *premises* ($x \in [x', x''], y \in [y', y'']$). The propagation rules used in the present paper are presented in Table 2.

Insert Table 2 about here

CAUTIOUS MONOTONICITY is the nonmonotonic counterpart to MONOTONICITY. CAUTIOUS MONOTONICITY is "cautious", because of the presence of the first premise, $A \mathrel{|\!\sim}_x B$. Contrary to its "incautious" or monotone counterpart, this premise cannot be

dropped. Argument (A1) above is an instance of MONOTONICITY. Argument (A2) is syntactically an instance of the CAUTIOUS MONOTONICITY rule, however the second premise that claims that birds are normally penguins is implausible. This blocks monotonicity.

CUT and RIGHT WEAKENING are the nonmonotonic counterparts to the TRANSITIVITY argument form. CUT is more cautious than TRANSITIVITY since CUT applies when *additionally* the "$A \wedge$"-part in the second premise is fulfilled. RIGHT WEAKENING is more cautious than TRANSITIVITY, since—as one can see in the second premise—it applies only when a *logical entailment* ($\models B \rightarrow C$) holds instead of just the nonmonotonic conditional ($B \mathrel{|\!\sim_y} C$) as in case of TRANSITIVITY. Logical entailment, $\models X \rightarrow Y$, is a very strong condition, since it means that $X \rightarrow Y$ is a tautology (Y follows from X under *all* truth valuations of X and Y).

Note that the propagation rule of the CUT rule is identical to that of the probabilistic MODUS PONENS (*from* $P(C|A) \in [x', x'']$ *and* $P(A) \in [y', y'']$ *infer* $P(C) \in [x'y', 1 - y' + x''y']$). If the "$B$"-part is dropped in the CUT rule, then the MODUS PONENS is an instance of the CUT. While the CUT has three propositional variables (A, B, C), the MODUS PONENS has only two propositional variables (A, C). Human subjects can reduce the complexity of the CUT rule by representing and processing it as a MODUS PONENS.

CONTRAPOSITION is a central property of classical monotone logic and adding it to SYSTEM P would make SYSTEM P monotonic. Furthermore, CONTRAPOSITION holds for the material implication ($A \rightarrow B$ if, and only if $\neg B \rightarrow \neg A$) but of course not for conditional probabilities. $P(A \rightarrow B) = P(\neg B \rightarrow \neg A)$, but $P(B|A) = P(\neg A|\neg B)$ does not hold necessarily.

For investigating the critical condition whether human subjects endorse the inference rules of SYSTEM P and whether human subjects do not endorse monotonic

argument forms, we now introduce the property "probabilistic informativeness". Probabilistic informativeness differentiates the SYSTEM P rules from monotonic inference argument forms. We call an inference rule *probabilistically informative* if the coherent probability interval of its conclusion is *not* necessarily equivalent to the unit interval, $[0, 1]$ (Pfeifer & Kleiter, 2006a). If an inference rule is probabilistically informative, then the premises *inform* about the probability of the conclusion. Consequently, an inference rule is probabilistically *not* informative, if the assignment of the unit interval to its conclusion is necessarily coherent. All the rules of SYSTEM P are probabilistically informative. Furthermore, if all probabilities in the premises of a SYSTEM P rule are equal to 1, then the probability of the conclusion is equal to 1. In their material version (i.e., if the nonmonotonic conditional $\mid\sim$ is replaced by the material implication $\rightarrow$) the rules of SYSTEM P are logically valid. Logical validity does not guarantee probabilistic informativeness, as in case of MONOTONICITY, TRANSITIVITY and CONTRAPOSITION. Furthermore, probabilistic informativeness does not guarantee logical validity: the probabilistic versions of the DENYING THE ANTECEDENT and AFFIRMING THE CONSEQUENT are not logically valid, but they are probabilistic informative. The propagation rules for the conditional syllogisms (MODUS PONENS, MODUS TOLLENS, DENYING THE ANTECEDENT, AFFIRMING THE CONSEQUENT) and the relationship between logical validity and probabilistic informativeness are reported by Pfeifer and Kleiter (2006a).

SYSTEM P is an example of a set of inference rules that are both, logically valid and probabilistically informative. Moreover, SYSTEM P is nonmonotonic with respect to the following two characteristics:

- it contains genuine nonmonotonic *conditionals*, $\mid\sim$, and
- the set of conclusions does not monotonically grow when further premises are added.

The goal of the present study is the investigation of human reasoning about nonmonotonic conditionals with inference rules of SYSTEM P. Furthermore, if the subject do endorse the nonmonotonic SYSTEM P rules, then we next want to see whether in the control condition the subjects do not endorse monotonic argument forms. If our predictions hold both in the test and in the control condition, then it is reasonable to infer that the subjects accept the basic rationality postulates of nonmonotonic reasoning.

In the present study, we do not ask whether subjects retract conclusions in the light of new evidence, rather, we ask how subjects reason from nonmonotonic conditionals and whether they do it in a rational way. Specifically, we investigate empirically the probabilistic interpretation of the CAUTIOUS MONOTONICITY rule, the CUT rule, and the RIGHT WEAKENING rule of SYSTEM P and three "incautious" counterparts, namely MONOTONICITY, TRANSITIVITY and CONTRAPOSITION, which are not contained in SYSTEM P. Our bold hypothesis is that the subjects propagate the probabilities from the premises in a coherent way to the conclusions, and that they infer probabilistically not informative intervals ($[0, 1]$) from the premises of the monotonic argument forms.

Empirical Studies on SYSTEM P

Theoretical papers on nonmonotonic reasoning often motivate their work by claiming to mimic human common sense. What is the empirical status of this claim? Studies on the suppression of conditional inferences (Byrne, 1989; Byrne et al., 1999; Bonnefon & Hilton, 2002, 2004; Dieussaert et al., 2005; Politzer, 2005), reasoning with uncertain premises (R. Stevenson & Over, 1995; George, 1997; R. J. Stevenson & Over, 2001; Politzer & Bourmaud, 2002), and belief revision (Elio & Pelletier, 1997; Politzer & Carles, 2001; Dieussaert et al., 2005) are closely related to nonmonotonic reasoning since they are concerned with withdrawing conclusions. These studies show that subjects suppress inferences if disabling premises are present, that subjects are willing to assign

degrees of confidence in conclusions if the premises are uncertain, and that subjects revise their beliefs in premises if contradicting conclusions arise. Now we are mentioning studies which explicitly claim to investigate nonmonotonic reasoning.

Elio and Pelletier (1993), Pelletier and Elio (2003), and Schurz (2005) conducted empirical studies on basic problems of nonmonotonic reasoning. Ford (2005), Ford and Billington (2000), and Vogel (1996) worked on human nonmonotonic inheritance reasoning. The studies used (unspecified) phrases like "normally" (Elio & Pelletier, 1993; Vogel, 1996), "usually" (Ford, 2005; Ford & Billington, 2000), or "generally" (Benferhat, Bonnefon, & Da Silva Neves, 2005). Verbal phrases like "normally", "usually", or "typically" are ambiguous. Dieussaert, Ford, and Horsten (2004), e.g., observed that "typically" is interpreted to be stronger than "usually". The interpretation of verbal phrases is also context dependent. Furthermore, interindividual differences can lead to different interpretations: for some subjects phrases like "normally" might mean something like "almost-every time" for another subject "most of the time" or something else. If phrases like "normally" are presented in the instruction, the subject can only infer what the experimenter has in mind and has to reason *to an interpretation* of "normally". Investigating reasoning to an interpretation is important for understanding how subjects form representations. In the present study, however, we are concerned with reasoning *from an interpretation*, i.e., how subjects manipulate representations from a fixed interpretation. Therefore we prefer to communicate non-ambiguous numerical values to the subjects and not verbal paraphrases like "normally". Furthermore, instead of "*if–then*" formulations, we use formulations that come closer to the probabilistic interpretation of the nonmonotonic conditional as a conditional probability. The importance of the distinction between reasoning to an interpretation and reasoning from an interpretation was stressed by Stenning and van Lambalgen (2004, 2005).

There are two studies on SYSTEM P using a possibilistic semantics (Da Silva Neves,

Bonnefon, & Raufaste, 2002; Benferhat et al., 2005). Da Silva Neves et al. (2002) investigated all rules of SYSTEM P except REFLEXIVITY. RATIONAL MONOTONICITY (*from* $A \mathrel{|\!\sim} C$ *and* $\neg(A \mathrel{|\!\sim} \neg B)$ *infer* $A \wedge B \mathrel{|\!\sim} C$) was tested as well, which is an extension of SYSTEM P. In addition the MONOTONICITY argument form was tested. According to the possibilistic semantics, a conditional $A \mathrel{|\!\sim} B$ is *judged to be plausible* if the possibility Π of both A and B is judged to be greater than the possibility of both A and $\neg B$, that is, $A \mathrel{|\!\sim} B$ *if, and only if* $\Pi(A \wedge B) > \Pi(A \wedge \neg B)$ (Dubois & Prade, 1988; Benferhat, Dubois, & Prade, 1997; Benferhat, Saffiotti, & Smets, 2000). The subjects rated pairs of phrases like "To which degree do you judge possible that Simon A. is a vegetarian and enjoys bull fights", and "... *does not* enjoy bull fights". The plausibility of twenty-four similar rules were rated in a pilot study by forty subjects. Ratings were made on a line with the labels *not possible at all* and *entirely possible* at the left and right hand side of the line. In the main experiment the premises and conclusions were *not* distinguishable by the subjects, since they were scattered around the test material. That is because the authors do "... not see these patterns [of SYSTEM P] as direct inference rules,... but as general emerging properties of the inferential apparatus" (Da Silva Neves et al., 2002, p. 110).

The main experiment of Da Silva Neves et al. (2002) involved eighty-eight first-year psychology students. Three kinds of monotonicity M_1, M_2, M_3, which differed in content, were tested. In the same sense, CUT and CAUTIOUS MONOTONICITY have been tested twice. The authors investigated whether the endorsement of the premises is preferentially associated with the endorsement of the conclusions for each pattern of SYSTEM P. M_1, CUT_2, $\text{CAUTIOUS MONOTONICITY}_1$, $\text{CAUTIOUS MONOTONICITY}_2$, RIGHT WEAKENING, OR , and AND were corroborated. There was a problem with the LEFT LOGICAL EQUIVALENCE rule. While it is an easy rule, no participant endorsed both, the premises and the conclusion. M_2, M_3 and RATIONAL MONOTONICITY were not corroborated. There was a strong influence of the content: CUT_1 and CUT_2, $\text{CAUTIOUS MONOTONICITY}_1$ and

CAUTIOUS MONOTONICITY$_2$, M$_1$, M$_2$ and M$_3$ showed significantly different data patterns.

Benferhat et al. (2005) report an experiment where the rules were presented in argument form. Subjects' reasoning was consistent with the LEFT LOGICAL EQUIVALENCE, RIGHT WEAKENING, OR, AND, and the CUT rule. The data was not conclusive about the CAUTIOUS MONOTONICITY and the RATIONAL MONOTONICITY rule.

Pfeifer and Kleiter (2005a) report a series of four studies on the AND, LEFT LOGICAL EQUIVALENCE, and the OR rule where the verbal indicators of nonmonotonicity were replaced by explicit numerical probabilities. The probabilities associated with the premises were presented in terms of percentages and the subjects inferred the probability associated to the conclusion. Subjects were free either to respond by an exact percentage or in terms of interval percentages. A good agreement between actual human inferences and the probabilistic interpretation of SYSTEM P was observed. The conjunction fallacy (i.e. violations of the upper probability bound) that is frequently reported in the literature (Gilovich et al., 2002) was not prominent in the data. More "lower bound" than "upper bound" violations were observed. If the premises contained interval percentages, subjects mean lower bound responses were higher than in the point percentage condition. A separate study showed that this effect is not produced by a matching with respect to those numbers already contained in the explanation of the task.

Practically all subjects endorsed the LEFT LOGICAL EQUIVALENCE, which makes the LEFT LOGICAL EQUIVALENCE rule attractive for mental rule theories. Mental rule theories postulate that the human inference engine is driven by basic formal rules like MODUS PONENS (Rips, 1994; Braine & O'Brien, 1998). Some of the inference rules may be applied faster and more accurate than others. LEFT LOGICAL EQUIVALENCE is one of the most easy inference rules. Furthermore, LEFT LOGICAL EQUIVALENCE is a candidate for a "hard-wired" rule of a *mental probability logic* (Pfeifer & Kleiter, 2005b). Mental probability logic is a psychological *competence theory* about how common sense

conditionals (*if A, then B*) are interpreted, premises of everyday life arguments are represented and how inferences are drawn ideally from the premises. Contrary to actual reasoning performance, reasoning competence refers to *ideal* reasoning performance.

The next sections report a series of four experiments on the CAUTIOUS MONOTONICITY, the CUT and the RIGHT WEAKENING rule. They are compared with central monotonic argument forms, namely the MONOTONICITY, CONTRAPOSITION (both directions) and two forms of the TRANSITIVITY argument form. In all experiments, the rules were packed into cover stories. The interpretation of "normally" was fixed by percentages. We used percentages for representing probabilities to use something "neutral" between frequencies and probabilities. Contrary to frequencies, percentages are normed between zero and one hundred, and they are easier to communicate to the subjects than probabilities. To keep context effects constant, we did not vary the covers-stories within and between the conditions. Only the percentages presented in the premises were varied within each condition. The subjects were free to respond either by point or by interval percentages (from at least ...% to at most ...%).

In Experiment 1 we investigated the CAUTIOUS MONOTONICITY rule and the MONOTONICITY argument form. If subjects reason nonmonotonically, then they should understand that the MONOTONICITY argument form is probabilistically not informative and they should infer wide intervals from the premises in the MONOTONICITY tasks. In Experiment 2 we investigated the CUT and the RIGHT WEAKENING rule of SYSTEM P. In addition, CONTRAPOSITION was investigated, which is a basic property of monotone logic. In Experiment 3 we investigated TRANSITIVITY, MODUS BARBARA, and CONTRAPOSITION. They are properties of monotone logic, probabilistically not informative, and they are not valid in SYSTEM P. Finally, in Experiment 4, CUT, TRANSITIVITY and RIGHT WEAKENING were investigated with more rigorous and improved cover stories.

Experiment 1: CAUTIOUS MONOTONICITY and MONOTONICITY

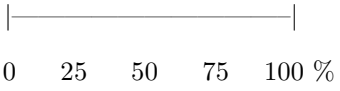
Subjects

Forty students of the University of Salzburg participated in the study. No students with special logical or mathematical education were included.

Method and Procedure

Each subject received a booklet containing a general introduction, one example explaining the response modality with point percentages, and one example explaining the response modality with interval percentages. Three target tasks were presented on separate pages. Eleven additional target tasks were presented in tabular form. The first three target tasks were of the following kind:

> About the guests at a prom we know the following:
>
> *exactly 89%* wear glasses
> *exactly 91%* wear a black suit
>
> Imagine all the persons of *this prom* that *wear glasses*.
>
> How many percent of the persons *wear a black suit*,
> **given** they are at *this prom* **and** *wear glasses*?
>
> The answer is either a point percentage or a percentage between two boundaries (*from at least ... to at most ...*):
>
> a.) If you think that the correct answer is an *point* percentage, please fill in your answer here:

	Point percentage
Exactly ...% of the persons *wear a black suit*, **given** they are at *this prom* **and** *wear glasses*.	

b.) If you think that the correct answer lies within two boundaries (*from at least ... to at most ...*), please mark the two values here:

Within the bounds of:

At least ...%, and *at most* ...%, of the persons *wear a black suit,* **given** they are at *this prom* **and** *wear glasses.*

|———————————————|

0 25 50 75 100 %

The next two tasks were formulated accordingly. The percentages of the two premises in the first task were 89 and 91%, in the second task 99 and 63%, and in the third task the percentages were 64 and 98%, respectively.

In addition, the subjects were asked to solve eleven analogous tasks presented in tabular form, as follows:

The following table lists eleven proms (a-k). It is known how many percent are *wearing a black suit*, and it is known how many percent are *wearing glasses*.

Please determine now how many percent of the persons *wear a black suit* **given** they are at *this prom* **and** *wear glasses*:

prom	*glasses*	*black suit*	*wear a black suit* **given** *at this prom* **and** *wear glasses*
a	60%	70%	
b	72%	63%	
c	90%	55%	
d	56%	99%	
e	63%	72%	
f	60%	60%	
g	100%	100%	
h	51%	51%	
i	88%	79%	
j	56%	77%	
k	79%	88%	

The percentages contained in the premises of these tasks were deliberately chosen to be identical to the percentages presented in Study 1 by Pfeifer and Kleiter (2005a). The subjects were divided into two groups, twenty subjects received the CAUTIOUS MONOTONICITY tasks, as just described, and twenty subjects received the MONOTONICITY tasks. The MONOTONICITY tasks were formulated exactly as the CAUTIOUS MONOTONICITY tasks, with the difference that the first premise (indicating the percentage of the persons wearing glasses) was dropped. The booklets were mixed and assigned arbitrarily to the subjects.

The subjects were tested individually in a quiet room in the psychology department. Subjects were told to take as much time as they want. In case of questions, the subjects were asked to reread the instructions carefully.

Results

The data of one subject of the CAUTIOUS MONOTONICITY condition was excluded from further analysis because of an incomplete response sheet. At the end of each session the subjects rated the overall comprehensibility of the tasks, how sure they were that their answers were correct, and the overall difficulty of the tasks, on a rating scale from one (very comprehensible, very certain, and very easy, respectively) to five (very incomprehensible, very uncertain, and very difficult, respectively). The mean comprehensibility of the CAUTIOUS MONOTONICITY tasks was 1.68 ($SD = .75$), the mean confidence in the correctness of the answers was 3.16 ($SD = 1.30$), and the mean difficulty was 2.53 ($SD = .96$). The mean comprehensibility of the MONOTONICITY tasks was 2.20 ($SD = 1.01$), the mean confidence was 2.80 ($SD = 1.40$), and the mean difficulty was 2.75 ($SD = .85$).

Insert Table 3 about here

We denote the first three tasks by $A1, A2$, and $A3$, and the eleven tasks presented in tabular form by $B1, \ldots, B11$. Table 3 lists subjects' mean lower and upper bound responses for the CAUTIOUS MONOTONICITY and the MONOTONICITY condition. In the CAUTIOUS MONOTONICITY condition, 37.22% of the subjects on the average responded by point values ($M = 7.07, SD = 3.56$, 14 tasks, $n = 19$). In the MONOTONICITY condition, 31.07% of the subjects on the average responded by point values ($M = 6.21, SD = 0.80$, 14 tasks, $n = 20$). That the majority responded by interval percentages indicates that most subjects understand that the premises provide only incomplete informations about the probability of the conclusion.

In the MONOTONICITY condition, more than half of the subjects responded by lower bounds ≤ 1. More than half of the subjects responded by upper bounds that are equal to the values presented in the premises of the tasks. Table 3 shows the frequencies of the interval responses with both, lower bounds ≤ 1 and upper bounds ≥ 91. On the average, 27% of the subjects responded by intervals with lower bounds ≤ 1 and upper bounds ≥ 91 ($M = 5.36, SD = 4.07$, $n = 14$ tasks). Most subjects understand that the lower bound can be practically zero and most subjects use a matching heuristic for inferring the upper bound.

To further investigate whether subjects infer probabilistically non-informative intervals in the MONOTONICITY tasks, we computed the mean interval size (i.e., upper minus lower bounds) for each task. Table 4 contains the mean interval sizes in both conditions. The intervals in the MONOTONICITY condition are clearly larger than in the CAUTIOUS MONOTONICITY condition. The subjects infer tighter intervals in the CAUTIOUS MONOTONICITY condition, and infer large intervals in the MONOTONICITY condition, as

postulated by the coherence interpretation of SYSTEM P. This speaks for the psychological plausibility in the CAUTIOUS MONOTONICITY rule of SYSTEM P.

Insert Table 4 about here

Insert Table 5 about here

In the CAUTIOUS MONOTONICITY condition the correlation between the mean lower bound responses and the normative lower bounds over all fourteen tasks was very high ($r = 0.92, t(12) = 8.04, p < .001$). The correlation between the mean upper bound responses and the normative upper bounds was not computed since ten out of the fourteen normative upper bounds were equal to 100.

If the normative bounds are greater than 0% and smaller than 100%, then there are six possible categories of interval responses, as illustrated in Figure 1. Of the six possible interval categories, only one category contains coherent intervals.

Table 5 reports the frequencies of interval responses of the CAUTIOUS MONOTONICITY condition in 3×3 tables. Each 3×3 table contains the six possible interval responses together with the according empirical frequencies of the interval responses. The *columns* designate whether the subjects' lower bounds are below (LB), within (LW), or above (LA) the normative intervals. The *rows* designate whether the subjects' upper bounds are above (UA), within (UW) or below (UB) the normative intervals.

Insert Figure 1 about here

In the CAUTIOUS MONOTONICITY condition, 51.89% of the subjects responded by coherent intervals on the average over all 14 tasks (cell ***e***, $M = 9.86, SD = 4.97, n = 14$). For ten of the fourteen tasks no upper bound violation is possible. In the four tasks with normative bounds smaller than 100%, no too wide intervals were observed and no interval was given with both bounds above the normative upper bounds. Practically all incoherent interval responses are violations of the normative lower bounds.

If subjects responses were equally distributed over all six possible response categories, then 16.67% would be in each cell. The coherent category contains the majority of responses. The frequency of responses belonging to this category clearly exceeds the guessing level. This is interpreted as a good agreement of our subjects with the CAUTIOUS MONOTONICITY rule.

For the MONOTONICITY condition a classification of the interval responses into coherent and incoherent inferences is not sensible, since all values between 0 and 100% are coherent.

In sum, subjects are moderately close to the normative intervals. Subjects are sensitive to the probabilistic non-informativeness of the MONOTONICITY argument form and are cautious with monotonicity.

Discussion

The subjects in the CAUTIOUS MONOTONICITY condition infer tight intervals, close to the normative intervals of the coherence semantics. The mean interval size (mean upper bound minus mean lower bound) in the MONOTONICITY condition is significantly larger than its nonmonotonic counterpart, indicating that the subjects understand the probabilistic non-informativeness of the MONOTONICITY argument form. This suggests that the subjects endorse the nonmonotonic CAUTIOUS MONOTONICITY rule and do not endorse the monotonic MONOTONICITY argument form. The result, that more than 50% of

the subjects responded by coherent intervals is similar to the investigations on the AND rule of SYSTEM P, where the same endorsement rate is reported (Pfeifer & Kleiter, 2005a).

In the next section we report data on two nonmonotonic versions of transitivity, namely CUT and RIGHT WEAKENING. CONTRAPOSITION was tested as well, which is a central property of classical monotone logic.

Experiment 2: CUT, RIGHT WEAKENING, CONTRAPOSITION

In the second experiment we investigated the CUT rule and the RIGHT WEAKENING rule of SYSTEM P, and two forms of the CONTRAPOSITION which are not valid in SYSTEM P. We call the two versions of the CONTRAPOSITION the "AN-CONTRAPOSITION" (affirmative premise and negations in the conclusion; *from* $A \mathrel{|\!\sim} B$ *infer* $\neg B \mathrel{|\!\sim} \neg A$) and the reversed version "NA-CONTRAPOSITION" (negations in the conclusion and affirmative premise; *from* $\neg B \mathrel{|\!\sim} \neg A$ *infer* $A \mathrel{|\!\sim} B$). As noted in the introduction, CUT and RIGHT WEAKENING are the nonmonotonic versions of TRANSITIVITY, and are therefore of special interest.

Subjects

Forty students of the University of Salzburg participated in the study. No students with special logical or mathematical education were included.

Method and Procedure

The method and procedure of Experiment 2 are the same as in Experiment 1. Subjects were tested individually and received a booklet with reasoning problems.

Twenty subjects were assigned to the CUT condition and twenty subjects were assigned to the RIGHT WEAKENING condition. In the CUT condition subjects were asked to imagine the following situation:

> *Exactly 89%* of the cars on a *big parking lot* are *blue*.
>
> *Exactly 91%* of *blue* cars that are on the *big parking lot* have *grey tire-caps*.

> Imagine all the cars that are on the *big parking lot*. How many percent of these cars have *grey tire-caps*?
>
> The answer is either a point percentage or a percentage between two boundaries (*from at least ... to at most ...*):

As in Experiment 1 subjects were free to respond either by point percentages or interval percentages. The first three tasks were presented on separate pages and the eleven subsequent tasks in tabular form. The percentages of all the respective premises were equal to the percentages of Experiment 1. After the fourteen CUT tasks the following NA-CONTRAPOSITION task was presented:

> Please imagine the following situation:
>
> > *Exactly 93%* of the cars that are **not** on a *big parking lot* are **not** *red*.
>
> Imagine all the cars that are *red*. How many percent of the *red* cars are on the *big parking lot*?

The RIGHT WEAKENING condition was in parallel to the CUT condition with the following two exceptions. First, the second premises of the CUT tasks (*Exactly 91%* of *blue* cars that are on the *big parking lot* have *grey tire-caps*.) were replaced by "*All blue* cars have *grey tire-caps*.". Second, after the fourteen RIGHT WEAKENING tasks the AN-CONTRAPOSITION task was presented, which was formulated as follows:

> Please imagine the following situation:
>
> > *Exactly 93%* of the cars that are *red* are on a *big parking lot*.
>
> Imagine all the cars that are **not** on the *big parking lot*. How many percent of the cars that are **not** on the *big parking lot* are **not** *red*?

Results

At the end of the sessions the subjects rated the overall task comprehensibility, how sure they were that their answers were correct, and the overall difficulty of the tasks on a rating scale from one (very comprehensible, very certain, and very difficult, respectively) to five (very incomprehensible, very uncertain, and very easy, respectively). Table 6 reports the mean ratings. One subject in the RIGHT WEAKENING condition did not rate the tasks and her answers.

Insert Table 6 about here

The task comprehensibility of the CUT and the RIGHT WEAKENING tasks was judged to be "good", and the subjects were intermediately confident in the correctness of their answers. The RIGHT WEAKENING tasks were judged to be easier than the CUT tasks ($t(37) = 4.58, p < .0001$), easier than the NA-CONTRAPOSITION task ($t(37) = 2.83, p < .01$) and easier than the AN-CONTRAPOSITION task ($t(37) = 2.16, p < .05$). In the NA-CONTRAPOSITION task the subjects were less confident in the correctness of their answers than in the RIGHT WEAKENING tasks ($t(37) = 2.69, p < .05$).

Insert Table 7 about here

Table 7 presents the mean upper and lower bound responses in the CUT condition and in the RIGHT WEAKENING condition. 25.71% of the subjects in the CUT condition responded by point values on the average ($M = 5.14, SD = 4.75$, 14 tasks, $n = 20$). 41.43% of the subjects in the RIGHT WEAKENING condition responded by point values on the average ($M = 8.29, SD = 3.17$, 14 tasks, $n = 20$).

Insert Table 8 about here

Table 8 reports the frequencies of the six possible interval response categories in the CUT tasks. 55.35% of the subjects responded by coherent intervals on the average in the CUT tasks (cell ***e***, $M = 11.07, SD = 5.17$, $n = 20$, 14 tasks). The mean frequencies of the incoherent response categories of cell ***a*** are 1.43 ($SD = 0.94$), ***b*** are 6.0 ($SD = 4.27$), ***c*** are 0.71 ($SD = 0.91$), ***d*** are 0.57 ($SD = 0.65$), and of cell ***f*** are 0.14 ($SD = 0.36$). We observe more upper bound violations than lower bound violations.

The frequencies of the coherent interval response categories in the RIGHT WEAKENING tasks are presented in Table 9. 87.15% of the subjects responded by coherent intervals on the average (cell ***e***, $M = 17.43, SD = 0.94$, fourteen tasks). No upper bound violations are possible in the RIGHT WEAKENING task.

Insert Table 9 about here

The correlation between the mean lower bound responses and the normative lower bounds over all fourteen tasks ($n = 14$) was $r = .98$. Since the normative upper bounds in all RIGHT WEAKENING tasks are equal to 100, the corresponding correlation between the upper bound responses and the normative upper bounds cannot be computed.

To estimate the reliability of the data we calculated correlations between the responses in those tasks which have normatively equivalent bounds in the conclusions. The propagation rule of the normative lower bound of the CUT rule is commutative ($xy = yx$, see Table 2). There are two pairs of tasks in which the percentages in the premises are interchanged (tasks $B2$ and $B5$, and $B9$ and $B11$, respectively). The

correlations between the lower bound responses of these pairs of tasks are $r = 0.93$ ($B2$ and $B5$, $n = 20$) and $r = 0.93$ ($B9$ and $B11$, $n = 20$). Tasks $B2$ and $B11$ have practically the same normative upper bounds, 82.36 and 81.52, respectively. The correlation between the upper bound responses of task $B2$ and task $B11$ was $r = 0.95$ ($n = 20$). The correlations indicate a high reliability of the data.

The mean lower bound of the NA-CONTRAPOSITION task was 6.50 ($SD = 20.58$). Only one subject gave a lower bound greater than 7 (namely "93"). The mean upper bound was 62.39 ($SD = 46.43$). Eight subjects gave an upper bound smaller than 93 (all of these eight subjects gave an upper bound equal to 7). In the AN-CONTRAPOSITION task, 11 subjects (i.e, 55%) gave interval responses with both, lower bounds ≤ 7 and upper bounds ≥ 93.

A similar result was observed in the AN-CONTRAPOSITION. More than half of the subjects responded by lower bounds ≤ 7.00 ($M = 30.20, SD = 42.39$). More than half of the subjects responded by upper bounds ≥ 93.00 ($M = 74.70, SD = 39.09$). 9 subjects (i.e, 45%) gave interval responses with both, lower bounds ≤ 7 and upper bounds ≥ 93.

In both versions of the CONTRAPOSITION tasks many subjects understand that the CONTRAPOSITION argument form is probabilistically not informative, that only lower bounds close to zero and upper bounds close to one hundred can be inferred.

Discussion

The data of Experiment 2 clearly endorse the RIGHT WEAKENING rule that is valid in SYSTEM P and clearly do not endorse the monotonic CONTRAPOSITION argument forms that are not valid in SYSTEM P. The endorsement rate of the RIGHT WEAKENING rule is very similar to that of the LEFT LOGICAL EQUIVALENCE rule of SYSTEM P (Pfeifer & Kleiter, 2005a). Practically all subjects endorse these rules. Both, the RIGHT WEAKENING and the LEFT LOGICAL EQUIVALENCE are important in the framework of SYSTEM P since

they allow the incorporation of logical knowledge into the reasoning process (Kraus et al., 1990). Furthermore, since these rules are naturally drawn by the subjects they are attractive for the mental probability logic. In a meta-analysis reported by Evans, Newstead, and Byrne (1993) the non-probabilistic version of the MODUS PONENS is endorsed by 89-100% of the subjects, which comes close to the endorsement rate of the RIGHT WEAKENING and the LEFT LOGICAL EQUIVALENCE.

A possible flaw of the RIGHT WEAKENING tasks is, that the logical tautology in the second premise of the RIGHT WEAKENING rule was formulated by a contingent All-statement, "*All blue* cars have *grey tire-caps*" (a contingent sentence is of course weaker than a logical tautology). In Experiment 4 we investigated the RIGHT WEAKENING more rigorous with a logical tautology.

The majority of the subjects in Experiment 2 responded by coherent intervals to the CUT tasks, which is comparable to the results of the CAUTIOUS MONOTONICITY tasks in Experiment 1 and the results of the AND tasks reported by Pfeifer and Kleiter (2005a). Subjects endorse the nonmonotonic rules and do not endorse the monotonic argument forms.

The next section investigates TRANSITIVITY and CONTRAPOSITION as critical monotonic conditions.

Experiment 3: TRANSITIVITY, MODUS BARBARA and CONTRAPOSITION

In the third experiment we investigated two versions of the TRANSITIVITY argument form. The two versions differed in the order of the premises. The first one was the TRANSITIVITY (or HYPOTHETICAL SYLLOGISM) and the second one had the form of the MODUS BARBARA, an argument scheme well known in the Aristotelian syllogistics (for the classical syllogistics and generalizations see Pfeifer (2006)). Since, formally, the order of the premises does not matter, both argument forms are formally equivalent. Both

TRANSITIVITY (*from* $A \mid\!\sim_x B$ *and* $B \mid\!\sim_y C$ *infer* $A \mid\!\sim_{[0,1]} C$) and MODUS BARBARA (*from* $B \mid\!\sim_x C$ *and* $A \mid\!\sim_y B$ *infer* $A \mid\!\sim_{[0,1]} C$) are monotonic argument forms and hence probabilistically not informative. Experiment 3 investigates whether human subjects are sensitive to the probabilistic non-informativeness of these argument forms and whether the order of the premises does influence human reasoning. Furthermore, we tried to replicate the results of the CONTRAPOSITION task in Experiment 2.

Subjects

Forty students of the University of Salzburg participated in the study. No students with special logical or mathematical education were included.

Method and Procedure

The method and procedure of Experiment 3 are the same as in Experiment 1. Subjects were tested individually and received a booklet with reasoning problems.

Twenty subjects were assigned to the TRANSITIVITY condition and twenty subjects were assigned to the MODUS BARBARA condition. In the TRANSITIVITY condition subjects were asked to imagine the following situation:

> Please imagine the following situation:
>
> > *Exactly 89%* of the cars on a *big parking lot* are *blue*.
> > *Exactly 91%* of the *blue* cars have *grey tire-caps*.
>
> Imagine all the cars that are on the *big parking lot*. How many percent of these cars have *grey tire-caps*?
>
> The answer is either a point percentage or a percentage between two boundaries (*from at least ... to at most ...*)

The rest of the instruction and the procedure was the same as in the previous experiments. After the last TRANSITIVITY task we presented again the NA-CONTRAPOSITION as in Experiment 2.

The MODUS BARBARA condition was formulated in the same way as the TRANSITIVITY condition with the exception of two differences. First, the order of the

premises of the first fourteen tasks was reversed. Second, the AN-CONTRAPOSITION was presented instead of the NA-CONTRAPOSITION.

Results

At the end of the sessions the subjects rated the overall task comprehensibility, how sure they were that their answers were correct, and the overall difficulty of the tasks. Table 10 reports the mean ratings.

Insert Table 10 about here

Subjects were less confident in the correctness of their responses in the MODUS BARBARA tasks than in the TRANSITIVITY tasks ($t(38) = 3.16, p < .01$).

Table 11 presents the mean upper and lower bound responses in the TRANSITIVITY and in the MODUS BARBARA condition. 45.71% of the subjects in the TRANSITIVITY tasks responded by point values on the average ($M = 9.14, SD = 2.98$, 14 tasks, $n = 20$). 63.93% of the subjects in the MODUS BARBARA tasks responded by point values on the average ($M = 12.79, SD = 2.22$, 14 tasks, $n = 20$). The mean values of the upper and lower bound responses, and the high percentage of point value responses indicate that the subjects do not understand the probabilistic non-informativeness of these monotonic argument forms.

Insert Table 11 about here

The mean lower bound responses of the AN-CONTRAPOSITION task was 11.30 ($SD = 28.11$) and the mean upper bound responses was 71.70 ($SD = 42.84$). Eighteen of the twenty subjects responded by a lower bound ≤ 7. Fourteen subjects responded by an

upper bound ≥ 93. Twelve subjects responded by intervals which have both, lower bounds ≤ 7 and upper bounds ≥ 93. Thus, 60% of the subjects understand that the AN-CONTRAPOSITION argument form is probabilistically not informative.

In the NA-CONTRAPOSITION task the mean lower bound responses was 38.70 ($SD = 41.52$) and the mean upper bound responses was 71.60 ($SD = 39.90$). Eleven of the twenty subjects responded by a lower bound ≤ 7. Thirteen subjects responded by an upper bound ≥ 93. Seven subjects responded by intervals which have both, lower bounds ≤ 7 and upper bounds ≥ 93. Thus, 35% of the subjects understand that the NA-CONTRAPOSITION argument form is probabilistically not informative.

Compared with Experiment 2, the percentage of subjects understanding the probabilistic non-informativeness of the CONTRAPOSITION varies from 35% to 60%. Practically all of the subjects who did not infer wide intervals responded either by lower and upper bounds that are close to zero, or by lower and upper bounds that are close to one hundred.

Adams (1975) stressed the probabilistic invalidity of the TRANSITIVITY and suggested to interpret TRANSITIVITY in everyday life argumentation as CUT. Bennett (2003) justified Adams' suggestion in terms of conversational implicatures (Grice, 1989). If a speaker first utters a premise of the form $A \mathrel{|\!\sim_x} B$ and then utters as the second premise $B \mathrel{|\!\sim_y} C$, the speaker actually means by the second premise a sentence of the form ($\underline{A \textit{ and}}$ B) $\mathrel{|\!\sim_y} C$. The speaker does not mention "*A and*" to the addressat because *A and* is already conversationally implied and "clear" from the context. Suppose we speak about cars on a *big parking lot* that are *blue*, and suppose we then add that

Exactly 91% of the *blue* cars have *grey tire-caps* ,

it is highly plausible to assume that we are speaking about blue cars *that are on the big parking lot*, even if we do not mention this explicitly.

This interpretation explains why subjects do not infer wide intervals close to the unit interval. If the conversational implicature hypothesis is correct, then the subjects actually interpret both forms of the TRANSITIVITY tasks as instances of the CUT rule. Table 12 presents the mean interval response frequencies of the TRANSITIVITY and of the MODUS BARBARA condition, classified by the coherent CUT bounds. In the TRANSITIVITY condition 62.14% of the subjects gave coherent interval responses in accordance of the CUT rule, in the MODUS BARBARA condition 50.00% of the subjects responded coherently. The results are quite similar to those of the original CUT condition in Experiment 2, where we observed 55.35% coherent interval responses.

Insert Table 12 about here

Discussion

Experiment 3 replicated the results of Experiment 2 on the CONTRAPOSITION argument forms. The subjects are sensitive to the probabilistic non-informativeness of both CONTRAPOSITION versions. The subjects do not understand the probabilistic non-informativeness of the TRANSITIVITY and the MODUS BARBARA. Under the hypothesis, that the subjects interpret these tasks as instances of the CUT rules we observed an endorsement rate similar to the endorsement rate in the CUT condition of Experiment 2. The cover-story somehow invites to read the task as if the premise "... of the *blue* cars have *grey tire-caps*" is about the cars on the big parking lot, and not—as we actually intended—about all blue cars in general (not only those on the big parking lot). This "misunderstanding" between the subjects and the experimenter could be caused by not explicitly mentioning the universe of discourse. We tried to overcome this problem by explicitly mentioning the universe of discourse in Experiment 4.

Experiment 4: CUT, TRANSITIVITY, and RIGHT WEAKENING

In Experiment 4 we investigated the CUT, the TRANSITIVITY and RIGHT WEAKENING rule. The cover stories of the tasks now explicitly stated the universe of discourse. This makes the tasks more comprehensible to the subjects and aims to decrease the influence of (implicit) conversational implicatures. In the RIGHT WEAKENING condition of Experiment 2 we used an contingent statement instead of a logical tautology. In Experiment 4 we investigated the RIGHT WEAKENING rule with a logical tautology.

Subjects

Forty students of the University of Salzburg participated in the study. No students with special logical or mathematical education were included.

Method and Procedure

As in the previous experiments, the subjects were tested individually and received a booklet with reasoning problems. Twenty subjects were assigned to the CUT condition and twenty subjects were assigned to the TRANSITIVITY condition.

In the CUT condition the subjects were asked to imagine the following situation:

> In Christmas time there are many skiers in the Arlberg ski-resort. The cable cars transport the skiers to the mountains every hour.
> It is known, that:
>
> > *Exactly* 99% of all the skiers in the *Galzig-cable car* have a *blue suite*.
> > *Exactly* 63% of all the skiers in the *Galzig-cable car* that have a *blue suite* are *ski instructors*.
>
> Imagine all the skiers that are in the *Galzig-cable car*. Please try to determine how many percent of the skiers in the *Galzig-cable car* are *ski instructors*?

Then, the subjects were informed that the answer is either a point percentage or a percentage between two boundaries (*from at least ... to at most ...*). The response modalities were the same as in Experiment 1.

The TRANSITIVITY condition was identical to the CUT condition with the exception that the second premise ("*Exactly* 63%...") was replaced by

> *Exactly* 63% of all the skiers in the Arlberg ski-resort that have a *blue suite* are *ski instructors*.

Here, the phrase "in the Arlberg ski-resort" makes the universe of discourse explicit and does not denote the subset of skiers in the Galzig cable car, as in the CUT condition.

After the CUT and the TRANSITIVITY tasks, respectively, we presented one RIGHT WEAKENING task to the subjects. In the RIGHT WEAKENING task the subjects were asked to imagine the following:

> In a small town in Southern Germany each inhabitant owns a car.
>
> > *All blue Volkswagen* cars are *blue cars*.
> > *Exactly 70%* of all the *inhabitants* of the small town own a *blue Volkswagen*.
>
> Please imagine now all the inhabitants of this small town.
> Please try now to determine how many percent of the *inhabitants* have a *blue car*.

Subjects were free to respond either in terms of point values or in terms of intervals.

Results

The data of four subjects in the CUT condition and one subject in the TRANSITIVITY condition was excluded from the data analysis because they did not answer all tasks. Task comprehension, confidence in the correctness of the responses, and the overall difficulty of the tasks were not investigated in Experiment 4.

Table 13 presents the mean upper and lower bound responses of the CUT-condition ($n = 16$) and the TRANSITIVITY condition ($n = 19$). In the TRANSITIVITY condition 51.13% of the subjects responded by point values on the average ($M = 9.71, SD = 2.13$, 14 tasks, $n = 19$). In the CUT condition 70.54% of the subjects responded by point values on the average ($M = 11.29, SD = 1.49$, 14 tasks, $n = 16$).

The mean interval responses in the TRANSITIVITY condition are slightly larger than those in the CUT condition; t-tests comparing the mean interval sizes in TRANSITIVITY tasks and in the CUT tasks were not significant. As in Experiment 3, the subjects did not understand the probabilistic non-informativeness of the TRANSITIVITY argument form.

Insert Table 13 about here

The frequencies of the interval response categories of the CUT tasks are presented in Table 14. 65.63% of the subjects in the CUT tasks responded by coherent intervals on the average (cell $\boldsymbol{e}$, $M = 10.50, SD = 2.59$, 14 tasks). This is a "better" performance compared with the 55.35% coherent interval responses in the CUT tasks of Experiment 2. The mean frequencies of the incoherent response categories of cell $\boldsymbol{a}$ are 0.86 ($SD = 0.36$), $\boldsymbol{b}$ are 6.71 ($SD = 0.61$), $\boldsymbol{c}$ are 0.07 ($SD = 0.27$), $\boldsymbol{d}$ are 1.14 ($SD = 0.86$), and of cell $\boldsymbol{f}$ are 2.71 ($SD = 2.13$).

Insert Table 14 about here

Similar as in Experiment 2, the correlation between the mean lower bound responses and the normative lower bounds for all fourteen CUT tasks was $r = .95$.

As in Experiment 2, we estimate the reliability of the data by the correlations between the responses in those tasks which have normatively equivalent bounds. The correlation between the lower bound responses in the tasks $B2$ and $B5$ is $r = .97$, and in the tasks $B9$ and $B11$ the correlation is $r = .99$. The correlation between the upper bound responses of the tasks $B2$ and $B11$ was $r = .80$. These results are similar to Experiment 2 indicating that the reliability of the data is high.

The RIGHT WEAKENING task was presented after the CUT tasks and after the TRANSITIVITY tasks. In the RIGHT WEAKENING task in CUT condition, fifteen of the sixteen subjects responded by the coherent lower bound "70" ($M = 67.50, SD = 10.00$) and 11 responded by the coherent upper bound "100" ($M = 88.12, SD = 20.40$). Eleven

subjects responded by the optimal coherent interval (70-100%), only one subject was incoherent because of a violation of the lower bound. All subjects except one endorsed the RIGHT WEAKENING rule.

In the RIGHT WEAKENING task in the TRANSITIVITY condition, eighteen of the nineteen subjects responded coherently "70" as the lower bound. Eight subjects responded by the optimal coherent upper bound, namely "100", and eleven responded by "70" which is coherent but not optimal ($M = 82.63, SD = 15.22$). Eight subjects responded by the optimal coherent interval. All subjects except one endorsed the RIGHT WEAKENING rule.

The formulation as a logical tautology makes the RIGHT WEAKENING task more easier for the subjects compared with the formulation of Experiment 2 where we observed an endorsement of 87.15% of the RIGHT WEAKENING rule.

Discussion

In Experiment 4 we replicated the results of the CUT and the RIGHT WEAKENING tasks of Experiment 2. The improved cover stories yielded a better result, producing a higher rate of coherent interval responses. Explicitly mentioning the universe discourse in the revised version of the TRANSITIVITY task did not help the subjects to understand the probabilistic non-informativeness of the TRANSITIVITY argument form.

There are several explanations. One explanation is, that the conversational implicatures are stronger and override the informations about the universe of discourse. Another explanation is that both forms of the TRANSITIVITY are proper three variable problems. Subjects thus reduce the processing demands by representing the TRANSITIVITY tasks as CUT, since the CUT can be reduced to a two variable problem, as explained in the introduction.

General Discussion

In the present study we investigated human probabilistic reasoning about nonmonotonic conditionals. A series of four studies was investigating the understanding of elementary rules of a central formal system of nonmonotonic reasoning called SYSTEM P. A total of 154 subjects was investigated. For the investigation whether subjects endorse nonmonotonic inference rules and do not endorse monotonic argument forms, we investigated three nonmonotonic inference rules valid in SYSTEM P and three central properties of classical (monotone) logic which are not valid in SYSTEM P. While nonmonotonic inference rules are probabilistically informative, the monotonic argument forms are probabilistically not informative. Anything in the not informative unit interval, $[0, 1]$, can be inferred from the premises of a monotonic argument.

Gilio (2002) proved probability propagation rules for the rules of SYSTEM P which determine how the coherent lower and upper probability bounds in the conclusion are inferred from the premises. The normative lower and upper bounds derived from the propagation rules were used to evaluate the rationality of the subjects' inferences.

RIGHT WEAKENING, CAUTIOUS MONOTONICITY, and CUT are nonmonotonic inference rules which are valid in SYSTEM P. All subjects with very few exceptions inferred probabilistically informative intervals in the nonmonotonic tasks. Practically all subjects perfectly endorse the RIGHT WEAKENING rule of SYSTEM P by inferring coherent intervals from the premises to the conclusion. More than 50% of the subjects inferred coherent intervals from the premises of the CAUTIOUS MONOTONICITY rule of SYSTEM P and from the premises of the CUT rule of SYSTEM P. This is a rather good result. Only one out of six possible categories of interval responses contains coherent intervals and just this category contains the majority of the interval responses.

MONOTONICITY, CONTRAPOSITION and TRANSITIVITY are monotonic argument forms which are not valid in SYSTEM P. A critical result of the present study is that

neither of the two monotonic argument forms MONOTONICITY and CONTRAPOSITION was endorsed by the subjects. In these tasks subjects understand the probabilistic non-informativeness of these two argument forms. The subjects did not understand the probabilistic non-informativeness of the monotonic TRANSITIVITY. We explained this by conversational implicatures, that the TRANSITIVITY tasks are actually interpreted by the subjects as instances of the CUT rule. We analyzed the TRANSITIVITY data under this assumption and observed a similar endorsement rate as in the CUT tasks.

In the present study, we were concerned with reasoning from a fixed interpretation of the nonmonotonic conditionals. We fixed the interpretation by communicating non-ambiguous percentages and used instead of "*if–then*"-statements careful formulations that come closer to the probabilistic interpretation of the nonmonotonic conditional.

If it is correct that humans interpret indicative common sense conditionals probabilistically, then the results in the CONTRAPOSITION tasks support the hypothesis that humans do not interpret common sense conditionals as probabilities of the material implication. Since $A \rightarrow B$ is logically equivalent with $\neg B \rightarrow \neg A$, the probabilities of the conditional and the contrapositive conditionals are equal, $P(A \rightarrow B) = P(\neg B \rightarrow \neg A)$. If subjects interpret the common sense conditional as probability of the material implication, they would have responded in the both versions of the CONTRAPOSITION task simply by the value presented in the premise, and not by non-informative intervals. Our data add to the results reported by Evans et al. (2003), where subjects assessed the probability of conditionals (*If* A *then* B) and of their contrapositives (*If* $\neg B$ *then* $\neg A$) independently.

Adding these results to the results to those reported in previous studies on the AND, LEFT LOGICAL EQUIVALENCE, and on the OR rule of SYSTEM P (Pfeifer & Kleiter, 2005a) supports the hypothesis, that the basic rationality postulates of SYSTEM P represent cornerstones in a competence theory of human reasoning. As reported above, there are also empirical studies on possibilistic semantics, that support the psychological plausibility

of SYSTEM P properties (Bonnefon & Hilton, 2002; Benferhat et al., 2005).

Raufaste, Da Silva Neves, and Mariné (2003) argue that standard probability theory cannot express partial ignorance in cases when the uncertainty about an event is poorly correlated with the uncertainty about the opposite event. In possibility theory, uncertainty about an event A is expressed such that A is entirely possible and simultaneously, $\neg A$ is entirely possible too. The knowledge of the probability of an event A is expressed by $P(A)$. If $P(A)$ is given then $P(\neg A)$ is determined. In conditional probability logic, however, it is straightforward to express partial ignorance about $P(A)$ by interval valued probabilities (Pfeifer & Kleiter, 2006a) or by second order probability distributions (Pfeifer & Kleiter, 2006b). Complete ignorance about the probability value of A, e.g., is expressed by $P(A) \in [0, 1]$, or by a uniform second order probability density function, respectively. In our experiments, the interval probabilities are determined by lower and upper bounds that are precise. Second order probability distributions allow to generalize our experimental paradigm by providing tools to express and reason from imprecise bounds (Pfeifer & Kleiter, 2006b).

In the present framework conditional probabilities are primitive concepts and are not defined over the fraction of absolute probabilities as in the classical approach, where $P(A|B)$ is defined by $P(A \wedge B)/P(B)$. Thus, formally zero probabilities are treated in a natural way (Biazzo et al., 2002; Coletti & Scozzafava, 2002; Gilio, 1999, 2002) and problems arising from conditioning events that equal zero are avoided. In the coherence based approach to probability it is natural to use only probabilities that are actually stated in the premises. No algebra of events closed under union and intersection of atomic events is required. Our subjects are supposed to make inferences from probabilities actually stated in the premises only. Moreover, in the present framework probabilities are *degrees of belief* and not an objective measures, of something external. Degrees of belief are more affine to the psychology of reasoning. In the cover stories of Experiment 1 and

Experiment 4 we used epistemic cues to introduce the degrees of beliefs ("... we know the following..." and "It is known that"). Phrases like "suppose, you belief that" would have been more appropriate for communicating the degrees of belie to the subjects.

In the experiments we observed high standard deviations which indicate interindividual differences among the subjects. The high standard deviations may be a consequence of the use of both point and interval responses. The subject could actually have some imprecise value in mind, nevertheless she responds by some mean value just to reduce the complexity of the task. Thus, the point value responses bias the mean lower bound and upper bound responses. In future experiments one might remove the possibility to give a point value response, which then would force the subjects to respond by intervals.

Of special interest are the tasks in which all premises are sure. This is the case in those tasks in which the percentages of the lower and upper bounds in the premises are equal to 100. Table 15 summarizes the mean lower and upper bound responses in the SYSTEM P rules and in the monotonic argument forms which are not valid in SYSTEM P.

Insert Table 15 about here

In the tasks with sure premises, practically all subjects endorse the SYSTEM P rules. The high endorsement rates are comparable to the endorsement rates of the non-probabilistic version of the MODUS PONENS (89–100%; Evans et al. (1993)).

In the MONOTONICITY task with sure premises the interval responses are large, which means that many subjects understand the probabilistic non-informativeness of the MONOTONICITY argument form. In the case of the monotonic argument forms TRANSITIVITY and MODUS BARBARA the mean lower bounds are very high. As discussed above, subjects might interpret these argument forms as CUT.

In the present study we did not investigate whether subjects actually withdraw

conclusions in the light of new evidence. Rather, reasoning from nonmonotonic conditionals was investigated and basic rationality postulates of nonmonotonic reasoning were corroborated. The critical condition of the monotonic argument forms indicate that most human subjects do not reason monotonically.

The rules of SYSTEM P may constitute the basic inference engine as postulated by the mental probability logic. Mental probability logic claims that the common sense conditional, *if A, then B*, is interpreted as a subjective conditional probability, $P(B|A)$. Based on the available informations, the premises are evaluated and represented by coherent point probabilities (x), coherent interval probabilities ($x \in [x', x'']$, where x' is the lower and x'' is the upper probability bound), second order probability distributions (Pfeifer & Kleiter, 2006b) or logical informations. Humans ideally assign coherent probabilities to the premise set and draw coherent conclusions basically by the use of formal inference rules of SYSTEM P.

While actual reasoning performance can be observed directly, reasoning competence cannot be inspected directly. Since the rules of SYSTEM P constitute the minimal properties of reasoning competence and since these rules are actually endorsed by most subjects, we think that these rules are at the core of the human inference engine.

In sum, nonmonotonic reasoning represents cornerstones of a competence theory of human reasoning that captures both, retraction of conclusions in the light of new evidence and rule guided reasoning.

References

Adams, E. W. (1975). *The logic of conditionals*. Dordrecht: Reidel.

Antoniou, G. (1997). *Nonmonotonic reasoning*. Cambridge, MA: MIT Press.

Benferhat, S., Bonnefon, J.-F., & Da Silva Neves, R. (2005). An overview of possibilistic handling of default reasoning, with experimental studies. *Synthese*, *1-2*, 53-70.

Benferhat, S., Dubois, D., & Prade, H. (1997). Nonmonotonic reasoning, conditional objects and possibility theory. *Artificial Intelligence*, *92*, 259-276.

Benferhat, S., Saffiotti, A., & Smets, P. (2000). Belief functions and default reasoning. *Artificial Intelligence*, *122*, 1-69.

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Oxford University Press.

Biazzo, V., Gilio, A., Lukasiewicz, T., & Sanfilippo, G. (2002). Probabilistic logic under coherence, model-theoretic probabilistic logic, and default reasoning in System P. *Journal of Applied Non-Classical Logics*, *12*(2), 189-213.

Bonnefon, J.-F., & Hilton, D. J. (2002). The suppression of modus ponens as a case of pragmatic preconditional reasoning. *Thinking & Reasoning*, *8*(1), 21-40.

Bonnefon, J.-F., & Hilton, D. J. (2004). Consequential conditionals: Invited and suppressed inferences from valid outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 28-37.

Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Mahwah, NJ: Erlbaum.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*, 61-83.

Byrne, R. M. J., Espino, O., & Santamaría, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language*, *40*, 347-373.

Chater, N., & Oaksford, M. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Science*, *5*(8), 349-357.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Coletti, G., & Scozzafava, R. (2002). *Probabilistic logic in a coherent setting*. Dordrecht: Kluwer.

Coletti, G., & Scozzafava, R. (2004). Conditional probability, fuzzy sets, and possibility: A unifying view. *Fuzzy Sets and Systems*, *144*, 227-249.

Coletti, G., & Scozzafava, R. (2005). Conditioning in a coherent setting: Theory and applications. *Fuzzy Sets and Systems*, *155*, 26-49.

Coletti, G., Scozzafava, R., & Vantaggi, B. (2001). Probabilistic reasoning as a general unifying tool. In S. Benferhat & P. Besnard (Eds.), *Lecture notes LNAI* (Vol. 2143, p. 120-131).

Da Silva Neves, R., Bonnefon, J.-F., & Raufaste, E. (2002). An empirical test of patterns for nonmonotonic inference. *Annals of Mathematics and Artificial Intelligence*, *34*, 107-130.

De Finetti, B. (1974). *Theory of probability* (Vols. 1, 2). Chichester: John Wiley & Sons. (Original work published 1970)

Delgrande, J. P. (1988). An approach to default reasoning based on a first-order conditional logic: Revised report. *Artificial Intelligence*, *36*, 63-90.

Dieussaert, K., De Neys, W., & Schaeken, W. (2005). Suppression and belief revision, two sides of the same coin? *Psychologica Belgica*, *45*(1), 29-46.

Dieussaert, K., Ford, M., & Horsten, L. (2004). Influencing nonmonotonic reasoning by modifier strength manipulation. In *Proceedings of the 26th annual conference of the*

cognitive science society. (p. 315-320). Chicago, IL. (`http://www.cogsci.northwestern.edu/cogsci2004/papers/paper337.pdf`)

Dubois, D., & Prade, H. (1988). *Possibility theory. An approach to computerized processing of uncertainty.* New York: Plenum Press.

Elio, R., & Pelletier, F. J. (1993). Human benchmarks on ai's benchmark problems. In *Proceedings of the 15th annual conference of the cognitive science society* (p. 406-411). Boulder, CO: Morgan Kaufmann.

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science*, *21*(4), 419-460.

Evans, J. S. B. T., Handley, S. H., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 321-355.

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning.* Hove, UK: Erlbaum.

Evans, J. S. B. T., & Over, D. E. (2004). *If.* Oxford: Oxford University Press.

Ford, M. (2005). Human nonmonotonic reasoning: The importance of seeing the logical strength of arguments. *Synthese*, *1-2*, 71-92.

Ford, M., & Billington, D. (2000). Strategies in human nonmonotonic reasoning. *Computational Intelligence*, *16*(3), 446-468.

Gabbay, D. M., & Hogger, C. J. (Eds.). (1994). *Handbook of logic in artificial intelligence and logic programming* (Vol. 3. Nonmonotonic reasoning and uncertain reasoning.). Oxford: Clarendon Press.

Gärdenfors, P., & Makinson, D. (1994). Nonmonotonic inference based on expectation orderings. *Artificial Intelligence*, *65*, 197-245.

George, C. (1997). Reasoning from uncertain premises. *Thinking and Reasoning*, *3*, 161-189.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gilio, A. (1999). Probabilistic relations among logically dependent conditional events. *Soft Computing*, *3*, 154-161.

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

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases. The psychology of intuitive judgment*. Cambridge, UK: Cambridge University Press.

Goldszmidt, M., & Pearl, J. (1996). Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*, *84*, 57-112.

Goodman, I. R., Nguyen, H. T., & Walker, E. A. (1991). *Conditional inference and logic for intelligent systems. A theory of measure-free conditioning*. Amsterdam: North-Holland.

Grice, H. P. (Ed.). (1989). *Studies in the way of words*. Cambridge, Massachusetts: Harvard University Press.

Hadjichristidis, C., Stevenson, R. J., Over, D. E., Sloman, S. A., Evans, J. S. B. T., & Feeney, A. (2001). On the evaluation of *if p then q* conditionals. In *Proceedings of the 23rd annual meeting of the cognitive science society.* Edinburgh, Scotland. (`http://www.hcrc.ed.ac.uk/cogsci2001/pdf-files/0381.pdf`)

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, UK: Cambridge University Press.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Konolige, K. (1994). Autoepistemic logic. In D. M. Gabbay, C. Hogger, & H. Robinson (Eds.), *Handbook of logic in artificial intelligence and logic programming: Nonmonotonic reasoning and uncertain reasoning* (Vol. 3, p. 218-295). Oxford: OUP.

Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, *44*, 167-207.

Lehmann, D., & Magidor, M. (1992). What does a conditional knowledgebase entail? *Artificial Intelligence*, *55*(1), 1-60.

Lifschitz, V. (1994). Circumscription. In D. Gabbay, C. Hogger, & H. Robinson (Eds.), *Handbook of logic in artificial intelligence and logic programming: Nonmonotonic reasoning and uncertain reasoning* (Vol. 3, p. 297-352). Oxford: OUP.

Liu, I.-M. (2003). Conditional reasoning and conditionalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 694-709.

Liu, I.-M., Lo, K.-C., & Wu, J.-T. (1996). A probabilistic interpretation of 'If—Then'. *The Quarterly Journal of Experimental Psychology*, *49*(A), 828-844.

Lukasiewicz, T. (2005). Weak nonmonotonic probabilistic logics. *Artificial Intelligence*, *168*, 119-161.

Macnamara, J. (1986). *The place of logic in psychology*. Cambridge, MA: MIT Press.

McCarthy, J. (1980). Circumscription: A form of non-monotonic reasoning. *Artificial Intelligence*, *13*, 27-39.

McDermott, D., & Doyle, J. (1980). Non-monotonic logic I. *Artificial Intelligence*, *13*, 41-72.

Moore, R. C. (1985). Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, *25*, 75-94.

Murphy, G. L. (2002). *The big book of concepts.* MIT Press.

Nute, D. (1994). Defeasible logic. In D. Gabbay, C. Hogger, & H. Robinson (Eds.), *Handbook of logic in artificial intelligence and logic programming: Nonmonotonic reasoning and uncertain reasoning* (Vol. 3, p. 353-395). Oxford: OUP.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 883-899.

Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 680-693.

Over, D. E., & Evans, J. S. B. T. (2003). The probability of conditionals: The psychological evidence. *Mind & Language*, *18*, 340-358.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1990). System Z. In M. Y. Vardi (Ed.), *Proceedings of theoretical aspects of reasoning about knowledge* (p. 21-135). Santa Mateo, CA: Morgan Kaufmann.

Pelletier, F. J., & Elio, R. (2003). Logic and computation: Human performance in default reasoning. In P. Gärdenfors, K. Kijania-Placet, & J. Wolenski (Eds.), *Logic, methodology, and philosophy of science.* Dordrecht: Kluwer.

Pfeifer, N. (2002). *Psychological investigations on human nonmonotonic reasoning with a focus on System P and the conjunction fallacy.* Unpublished master's thesis, Institut für Psychologie, Universität Salzburg.

Pfeifer, N. (2006). Contemporary syllogistics: Comparative and quantitative syllogisms. In G. Kreuzbauer & G. J. W. Dorn (Eds.), *Argumentation in Theorie und Praxis: Philosophie und Didaktik des Argumentierens* (p. 57-71). Wien: LIT.

Pfeifer, N., & Kleiter, G. D. (2005a). Coherence and nonmonotonicity in human reasoning. *Synthese*, *146*(1-2), 93-109.

Pfeifer, N., & Kleiter, G. D. (2005b). Towards a mental probability logic. *Psychologica Belgica*, *45*(1), 71-99. (Updated version at: `http://users.sbg.ac.at/~pfeifern/`)

Pfeifer, N., & Kleiter, G. D. (2006a). Inference in conditional probability logic. *Kybernetika*, *42*, 391-404.

Pfeifer, N., & Kleiter, G. D. (2006b). Towards a probability logic based on statistical reasoning. In *Proceedings of the 11th IPMU conference (Information processing and management of uncertainty in knowledge-based systems)* (p. 2308-2315). Paris, France.

Politzer, G. (2005). Uncertainty and the suppression of inferences. *Thinking & Reasoning*, *11*(1), 5-33.

Politzer, G., & Bourmaud, G. (2002). Deductive reasoning from uncertain conditionals. *British Journal of Psychology*, *93*, 345-381.

Politzer, G., & Carles, L. (2001). Belief revision and uncertain reasoning. *Thinking & Reasoning*, *7*(3), 217-234.

Pollock, J. (1994). Justification and defeat. *Artificial Intelligence*, *67*, 377-407.

Poole, D. (1980). A logical framework for default reasoning. *Artificial Intelligence*, *36*, 27-47.

Ramsey, F. P. (1994). General propositions and causality (1929). In D. H. Mellor (Ed.),

Philosophical papers by F. P. Ramsey (p. 145-163). Cambridge: Cambridge University Press.

Raufaste, E., Da Silva Neves, R., & Mariné, C. (2003). Testing the descriptive validity of possibility theory in human judgments of uncertainty. *Artificial Intelligence*, *148*(1-2), 197-218.

Reiter, R. (1980). A logic of default reasoning. *Artificial Intelligence*, *13*, 81-132.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, Massachusetts: MIT Press.

Schurz, G. (1997). Probabilistic default reasoning based on relevance and irrelevance assumptions. In D. Gabbay et al. (Ed.), *Qualitative and quantitative practical reasoning* (p. 536-553). Berlin: Springer.

Schurz, G. (1998). Probabilistic semantics for Delgrande's conditional logic and a counterexample to his default logic. *Artificial Intelligence*, *102*, 81-95.

Schurz, G. (2005). Non-monotonic reasoning from an evolution-theoretic perspective: Ontic, logical and cognitive foundations. *Synthese*, *1-2*, 37-51.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, Mass.: Harvard University Press.

Stenning, K., & van Lambalgen, M. (2004). A little logic goes a long way: Basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cognitive Science*, *28*, 481-529.

Stenning, K., & van Lambalgen, M. (2005). Semantic interpretation as computation in nonmonotonic logic: The real meaning of the suppression task. *Cognitive Science*, *29*, 919-960.

Stevenson, R., & Over, D. E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology*, *48A*, 613-643.

Stevenson, R. J., & Over, D. E. (2001). Reasoning from uncertain premises: Effects of expertise and conversational context. *Thinking and Reasoning*, *7*, 367-390.

Touretzky, D. S. (1986). *The mathematics of inheritance systems*. Los Altos, CA: Morgan Kaufmann.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293-315.

Vogel, C. (1996). Human reasoning with negative defaults. In D. Gabbay & H. J. Ohlbach (Eds.), *Practical reasoning, lecture notes in artificial intelligence, 1085* (p. 606-621). Berlin: Springer.

Author Note

This study was supported by the AKTION (Austrian-Czech exchange program, 45p16).

Footnotes

[1]Some authors suggest that $P(B|A) = 1$, if $P(A) = 0$. This is incoherent since it follows that if $P(A) = 0$, then $P(B|A) = 1 = P(\neg B|A)$, which is incoherent of course, since $P(B|A) + P(\neg B|A)$ should sum up to 1 (Coletti, Scozzafava, & Vantaggi, 2001; Gilio, 2002).

[2]*Default reasoning* (Reiter, 1980; Poole, 1980), *Autoepistemic (nonmonotonic reasoning) logic* (McDermott & Doyle, 1980; Moore, 1985; Konolige, 1994), *Circumscription* (McCarthy, 1980; Lifschitz, 1994), *Defeasible reasoning* (Pollock, 1994; Nute, 1994), *Default inheritance reasoning* (Touretzky, 1986), *Possibility theory* (Dubois & Prade, 1988; Benferhat et al., 1997, 2000), and *Conditional* and *preferential entailment: conditional logic*-based (Delgrande, 1988; Schurz, 1998), *preferential model*-based (Kraus et al., 1990; Lehmann & Magidor, 1992), *expectation-ordering*-based (Gärdenfors & Makinson, 1994), and *probabilistic entailment*-based approaches (Adams, 1975; Pearl, 1988, 1990; Schurz, 1998; Gilio, 2002; Lukasiewicz, 2005), just to mention some of the most important. For an overview refer to Gabbay & Hogger (1994), or Antoniou (1997).

Table 1

Probabilistic version of the Tweety example.

P1	$P[\text{Fly}(x)\|\text{Bird}(x)] = .95.$	(*x is a bird*) $\mid\!\sim_{.95}$ (*x can fly*)
P2	Bird(Tweety).	*Tweety is a bird.*
C1	$P[\text{Fly}(\text{Tweety})] = .95.$	$\mid\!\sim_{.95}$ (*Tweety can fly*)
P3	Penguin(Tweety).	*Tweety is a penguin.*
P4	$P[\text{Fly}(x)\|\text{Penguin}(x)] = .01.$	(*x is a penguin*) $\mid\!\sim_{.01}$ (*x can fly*)
P5	$P[\text{Bird}(x)\|\text{Penguin}(x)] = .99.$	(*x is a penguin*) $\mid\!\sim_{.99}$ (*x is a bird*)
C2	$P[\text{Fly}(\text{Tweety}) \mid \text{Bird}(\text{Tweety}) \wedge \text{Penguin}(\text{Tweety})] \in [0, .01].$	
	(*Tweety is a bird* $\wedge$ *Tweety is a penguin*)	$\mid\!\sim_{[0,.01]}$ (*Tweety can fly*)

Note. P1-P2 are the initial premises, C1 is the preliminary conclusion, P3-P5 are the additional premises, C2 is the final conclusion. The inference from P1 and P2 to C1 is justified by the probabilistic MODUS PONENS (Pfeifer & Kleiter, 2006a) and the inference from P1-P5 to C2 is justified by the CAUTIOUS MONOTONICITY rule of SYSTEM P, see Table 2 below (P4 and P5 are instantiations of the premises and C2 is an instance of the conclusion of the CAUTIOUS MONOTONICITY.

Table 2
Inference rules and propagation rules for the probabilities in the premises $(0 \leq x \leq 1,\ 0 \leq y \leq 1)$ *to the coherent probability (interval* z*) of the conclusion investigated in the present study.*

Inference rule		*Propagation rule*	#
RIGHT WEAKENING:	*from* $A \mathrel{|\!\sim}_x B$ *and* $\models B \rightarrow C$ *infer* $A \mathrel{|\!\sim}_z C$	$x \leq z \leq 1$	2,4
CAUTIOUS MONOTONICITY:	*from* $A \mathrel{|\!\sim}_x B$ *and* $A \mathrel{|\!\sim}_y C$ *infer* $A \wedge B \mathrel{|\!\sim}_z C$	$\left.\begin{array}{ll} \frac{x+y-1}{x}, & if \quad x+y>1 \\ 0, & if \quad x+y \leq 1 \end{array}\right\} \leq z \leq \left\{\begin{array}{ll} \frac{y}{x}, & if \quad y<x \\ 1, & if \quad y \geq x \end{array}\right.$	1
CUT:	*from* $A \mathrel{|\!\sim}_x B$ *and* $A \wedge B \mathrel{|\!\sim}_y C$ *infer* $A \mathrel{|\!\sim}_z C$	$xy \leq z \leq 1-x+xy$	2,4
MONOTONICITY:	*from* $A \mathrel{|\!\sim}_x B$ *infer* $A \wedge C \mathrel{|\!\sim}_z B$	$0 \leq z \leq 1$	1
TRANSITIVITY:	*from* $A \mathrel{|\!\sim}_x B$ *and* $B \mathrel{|\!\sim}_y C$ *infer* $A \mathrel{|\!\sim}_z C$	$0 \leq z \leq 1$	3,4
MODUS BARBARA:	*from* $B \mathrel{|\!\sim}_x C$ *and* $A \mathrel{|\!\sim}_y B$ *infer* $A \mathrel{|\!\sim}_z C$	$0 \leq z \leq 1$	3
AN-CONTRAPOSITION:	*from* $A \mathrel{|\!\sim}_x B$ *infer* $\neg B \mathrel{|\!\sim}_z \neg A$	$0 \leq z \leq 1$	2,3
NA-CONTRAPOSITION:	*from* $\neg B \mathrel{|\!\sim}_x \neg A$ *infer* $A \mathrel{|\!\sim}_z B$	$0 \leq z \leq 1$	2,3

Note. The rules above the line are contained in SYSTEM P and are probabilistically informative, while the argument forms below the line are not contained in SYSTEM P and are not probabilistically informative. $\models X \rightarrow Y$ means that X *implies logically* Y (i.e. $X \rightarrow Y$ is a tautology). $\wedge$ ("and") and $\neg$ ("not") are defined as usual in classical propositional logic. For simplicity, the propagation rules are formulated for point values in the premises (see Gilio (2002) for interval-valued premises). Column # designates the number of the experiment in which the respective rule was investigated.

Table 3
Subjects' mean lower and upper bound responses and frequencies of $[L \leq 1, 91 \leq U]$*-interval responses in the* MONOTONICITY *tasks of Experiment 1.*

		A1L	*A1U*	*A2L*	*A2U*	*A3L*	*A3U*	*B1L*
CM	*Mean:*	76.42 (89.89)	90.05 (100)	57.55 (62.63)	68.60 (63.64)	68.35 (96.88)	74.93 (100)	43.79 (50.00)
	SD:	20.06	5.24	15.29	13.68	22.66	16.81	16.72
M	*Mean:*	26.75 [12]	87.05	19.10 [2]	65.10	32.45 [12]	92.25	17.90 [3]
	SD:	39.81	15.35	27.07	18.41	45.29	19.44	28.49
		B1U	*B2L*	*B2U*	*B3L*	*B3U*	*B4L*	*B4U*
CM	*Mean:*	65.63 (100)	40.87 (48.61)	62.82 (87.50)	43.11 (50.00)	58.16 (61.11)	66.17 (98.21)	71.79 (100)
	SD:	18.02	17.84	10.19	15.72	11.63	25.92	21.70
M	*Mean:*	68.25	17.07 [3]	63.67	13.45 [3]	56.56	25.05 [12]	90.55
	SD:	20.47	25.72	19.63	22.20	22.53	39.97	21.21
		B5L	*B5U*	*B6L*	*B6U*	*B7L*	*B7U*	*B8L*
CM	*Mean:*	49.68 (55.56)	66.37 (100)	35.91 (33.33)	62.95 (100)	100 (100)	100 (100)	23.21 (3.92)
	SD:	21.97	19.03	17.94	19.45	.00	.00	23.45
M	*Mean:*	18.25 [3]	70.10	14.90 [3]	60.10	41.25 [10]	92.10	14.00 [3]
	SD:	29.25	19.57	24.54	22.59	46.63	19.31	20.97
		B8U	*B9L*	*B9U*	*B10L*	*B10U*	*B11L*	*B11U*
CM	*Mean:*	54.26 (100)	64.90 (76.14)	80.72 (89.77)	47.47 (58.93)	66.47 (100)	70.20 (84.81)	80.05 (100)
	SD:	23.74	20.65	7.47	18.89	19.07	17.38	17.05
M	*Mean:*	55.35	20.40 [3]	75.25	19.05 [3]	73.40	21.75 [3]	81.05
	SD:	21.28	31.98	19.76	31.15	20.55	36.04	23.28

Note. L and U designate subjects' lower and upper bound responses, respectively. The normative lower and upper bounds are given in round parentheses. In each MONOTONICITY task, the normative lower and upper bounds are 0 and 100, respectively. The frequencies of the interval responses with both, lower bounds ≤ 1 and upper bounds ≥ 91, are in square bracket. CM = CAUTIOUS MONOTONICITY condition ($n = 19$). M = MONOTONICITY condition ($n = 20$).

Table 4

Subjects' mean interval size of Experiment 1.

	*A1*M	*A1*CM	*A2*M	*A2*CM	*A3*M	*A3*CM	*B1*M
Mean	60.30***	13.63***	46.00***	11.05***	59.80***	6.59***	50.35**
SD	24.46	14.82	8.66	1.61	25.84	5.85	8.02
Cohen's d	2.31		5.61		2.84		4.96
	*B1*CM	*B2*M	*B2*CM	*B3*M	*B3*CM	*B4*M	*B4*CM
Mean	21.84**	46.60*	21.94*	43.10**	15.06**	65.50***	5.62***
SD	1.30	6.09	7.64	.33	4.09	18.76	4.23
Cohen's d		3.57		9.74		4.40	
	*B5*M	*B5*CM	*B6*M	*B6*CM	*B7*M	*B7*CM	*B8*M
Mean	51.85***	16.68***	45.20^{ns}	27.04^{ns}	50.85***	.00***	41.35^{ns}
SD	9.68	2.95	1.94	1.51	27.31	.00	.32
Cohen's d	4.92		10.45		2.63		33.70
	*B8*CM	*B9*M	*B9*CM	*B10*M	*B10*CM	*B11*M	*B11*CM
Mean	31.06^{ns}	54.85***	15.82***	54.35**	19.00**	59.30***	9.85***
SD	.29	12.23	13.18	10.59	.18	12.76	.33
Cohen's d		3.07		4.72		5.48	

Note. CM = CAUTIOUS MONOTONICITY condition ($n = 19$). M = MONOTONICITY condition ($n = 20$). "***" designate t-tests comparing M and CM significant at $p < .001$, "**" at $p < .01$, "*" at $p < .05$, and "^{ns}" means "not significant".

Table 5
Frequencies of the interval responses in the CAUTIOUS MONOTONICITY *condition of Experiment 1 ($n = 19$).*

	Task A1 (89.89-100)			*Task A2 (62.63-63.64)*			*Task A3 (96.89-100)*			*Task B1 (50.00-100)*			*Task B2 (48.61-87.50)*		
	[89 and 91]			*[99 and 63]*			*[64 and 98]*			*[60 and 70]*			*[72 and 63]*		
UA	-	-	-	0	3	0	-	-	-	-	-	-	0	0	0
UW	11	**6**	-	2	**14**	-	1	**5**	-	9	**8**	-	8	**10**	-
UB	2	-	-	0	-	-	13	-	-	2	-	-	1	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task B3 (50.00-61.11)			*Task B4 (98.21-100)*			*Task B5 (55.56-100)*			*Task B6 (33.33-100)*			*Task B7 (100-100)*		
	[90 and 55]			*[56 and 99]*			*[63 and 72]*			*[60 and 60]*			*[100 and 100]*		
UA	0	2	0	-	-	-	-	-	-	-	-	-	-	-	-
UW	2	**15**	-	1	**6**	-	8	**8**	-	8	**10**	-	0	**19**	-
UB	0	-	-	12	-	-	3	-	-	1	-	-	0	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task B8 (3.92-100)			*Task B9 (76.14-89.77)*			*Task B10 (58.93-100)*			*Task B11 (84.81-100)*			***Task***		
	[51 and 51]			*[88 and 79]*			*[56 and 77]*			*[79 and 88]*			*[**Pr.1** and **Pr.2**]*		
UA	-	-	-	0	1	0	-	-	-	-	-	-	***a***	***b***	***c***
UW	0	**19**	-	11	**5**	-	9	**8**	-	3	**5**	-	***d***	***e***	-
UB	0	-	-	2	-	-	2	-	-	11	-	-	***f***	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*

Note. The percentages presented in the premises are in the square brackets and the normative intervals are in the round parentheses. *UA*: the subjects' upper bound response is *above* the normative upper bound, *UW*: upper bound response is *within* the normative interval, *UB*: upper bound response is *below* the normative lower bound; *LA*, *LW*, and *LB*: same for the subjects' lower bound responses. ***a***: too wide interval responses, ***b***: lower bound responses coherent, ***c***: both bound responses above, ***d***: upper bound responses coherent, ***e***: both bound responses coherently within ±5% (bold), ***f***: both bound responses below the normative lower bounds. For a graphical representation of the six interval response categories see Figure 1.

Table 6

Mean ratings of overall task comprehensibility, confidence in the correctness of the answers, and task difficulty in the CUT *($n = 20$) and in the* RIGHT WEAKENING *($n = 19$) condition in Experiment 2.*

	Comprehensibility	*Confidence*	*Difficulty*
CUT	2.10 (1.07)	3.40 (1.43)	2.25 (0.91)
RIGHT WEAKENING	1.89 (0.88)	2.37 (1.21)	3.58 (0.90)
NA-CONTRAPOSITION	2.55 (1.28)	3.55 (1.50)	2.55 (1.32)
AN-CONTRAPOSITION	2.47 (1.12)	3.05 (1.35)	2.89 (1.05)

Note. The tasks were rated by the subjects on a scale from one (very comprehensible, very certain, and very difficult, respectively) to five. The *SD*s are in the parentheses.

Table 7
Subjects' mean lower and upper bound responses in the CUT *(n = 20) and in the* RIGHT WEAKENING *(n = 20) condition of Experiment 2.*

		A1L	*A1U*	*A2L*	*A2U*	*A3L*	*A3U*	*B1L*
CUT	*Mean:*	75.06 (80.99)	93.65 (89.99)	59.05 (62.37)	79.96 (99.37)	61.55 (62.72)	90.72 (64.72)	42.80 (42.00)
	SD:	21.31	5.07	13.71	17.64	16.87	13.27	11.99
RW	*Mean:*	80.10 (89.00)	95.60 (100)	94.05 (99.00)	99.55 (100)	60.80 (64.00)	83.80 (100)	59.50 (60.00)
	SD:	27.39	5.53	22.14	0.51	14.31	18.38	25.64
		B1U	*B2L*	*B2U*	*B3L*	*B3U*	*B4L*	*B4U*
CUT	*Mean:*	74.90 (72.00)	47.79 (45.36)	70.37 (82.36)	49.90 (49.50)	72.22 (94.50)	53.08 (55.44)	87.45(56.44)
	SD:	22.28	13.68	19.40	16.28	20.49	13.66	18.92
RW	*Mean:*	88.00 (100)	50.75 (63.00)	80.55 (100)	46.75 (55.00)	82.00 (100)	84.15 (56.00)	99.60 (100)
	SD:	15.08	25.17	25.17	20.15	22.62	36.27	0.50
		B5L	*B5U*	*B6L*	*B6U*	*B7L*	*B7U*	*B8L*
CUT	*Mean:*	45.20 (45.36)	74.18 (73.36)	42.85 (36.00)	70.75 (76.00)	95.05 (100)	100 (100)	41.70 (26.01)
	SD:	12.45	20.88	18.84	23.09	22.14	0.00	28.08
RW	*Mean:*	61.20 (63.00)	88.80 (100)	51.00 (60.00)	84.00 (100)	95.00 (100)	100 (100)	43.35 (51.00)
	SD:	26.38	14.07	21.98	20.10	22.36	0.00	18.68
		B8U	*B9L*	*B9U*	*B10L*	*B10U*	*B11L*	*B11U*
CUT	*Mean:*	68.16 (75.01)	68.50 (69.52)	84.41 (90.52)	43.92 (43.12)	75.36 (66.12)	67.06 (69.52)	86.30 (81.52)
	SD:	27.58	16.86	12.36	13.70	22.52	16.77	11.27
RW	*Mean:*	80.40 (100)	67.15 (79.00)	91.60 (100)	65.45 (56.00)	90.80 (100)	74.80 (79.00)	95.20 (100)
	SD:	24.63	28.94	10.56	28.21	11.56	32.24	6.03

Note. L and *U* designate the subjects' lower and upper bound responses, respectively. The normative lower and upper bounds are given in parentheses. RW = RIGHT WEAKENING condition.

Table 8

Frequencies of the interval responses in the CUT *condition of Experiment 2 ($n = 20$).*

	Task A1 (80.99-89.99)			*A2 (62.37-99.37)*			*A3 (62.72-64.72)*			*B1 (42.00-72.00)*			*B2 (45.36-82.36)*		
	[89 and 91]			*[99 and 63]*			*[64 and 98]*			*[60 and 70]*			*[72 and 63]*		
UA	1	6	0	0	0	0	2	13	2	1	11	0	1	4	0
UW	2	**11**	-	1	**19**	-	0	**3**	-	0	**8**	-	1	**14**	-
UB	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task B3 (49.50-94.50)			*B4 (55.44-56.44)*			*B5 (45.36-73.36)*			*B6 (36.00-76.00)*			*B7 (100-100)*		
	[90 and 55]			*[56 and 99]*			*[63 and 72]*			*[60 and 60]*			*[100 and 100]*		
UA	1	4	1	3	11	1	3	8	0	1	3	1	-	-	-
UW	0	**14**	-	1	**4**	-	1	**8**	-	0	**15**	-	1	**19**	-
UB	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task B8 (26.01-75.01)			*B9 (69.52-90.52)*			*B10 (43.12-66.12)*			*B11 (69.52-81.52)*			***Task***		
	[51 and 51]			*[88 and 79]*			*[56 and 77]*			*[79 and 88]*			*[**Pr.1** and **Pr.2**]*		
UA	1	2	3	2	4	0	2	10	1	2	9	1	***a***	***b***	***c***
UW	0	**14**	-	0	**13**	-	0	**6**	-	1	**7**	-	***d***	***e***	-
UB	0	-	-	1	-	-	1	-	-	0	-	-	***f***	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*

Note. The percentages presented in the premises are in the square brackets and the normative intervals are in the round parentheses. For the abbreviations see Table 5.

Table 9

Frequencies of the interval responses in the RIGHT WEAKENING *condition of Experiment 2 ($n = 20$).*

	Task A1 (89.00-100)			*Task A2 (99.00-100)*			*Task A3 (64.00-100)*			*Task B1 (60.00-100)*			*Task B2 (63.00-100)*		
UW	2	**18**	-	1	**19**	-	3	**17**	-	3	**16**	-	3	**16**	-
UB	0	-	-	0	-	-	0	-	-	0	-	-	1	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task B3 (55.00-100)			*Task B4 (56.00-100)*			*Task B5 (63.00-100)*			*Task B6 (60.00-100)*			*Task B7 (100-100)*		
UW	3	**17**	-	3	**17**	-	3	**17**	-	3	**17**	-	3	**17**	-
UB	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task B8 (51.00-100)			*Task B9 (79.00-100)*			*Task B10 (56.00-100)*			*Task B11 (79.00-100)*			***Cells***		
UW	1	**19**	-	3	**17**	-	3	**17**	-	3	**17**	-	***d***	***e***	-
UB	0	-	-	0	-	-	0	-	-	0	-	-	***f***	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*

Note. The normative bounds are in the round parentheses. The percentages presented in the premises are identical to the normative lower bounds. Since the normative upper bounds are equal to 100, no violations of the upper bounds are possible. For the abbreviations see Table 5.

Table 10

Mean ratings of overall task comprehensibility, confidence in the correctness of the answers, and task difficulty in the TRANSITIVITY *(n = 20) and in the* MODUS BARBARA *(n = 20) condition in Experiment 3.*

	Comprehensibility	*Confidence*	*Difficulty*
TRANSITIVITY	2.35 (1.04)	2.75 (1.16)	2.80 (0.95)
MODUS BARBARA	2.80 (0.95)	3.95 (1.23)	2.40 (0.82)
NA-CONTRAPOSITION	3.15 (1.35)	3.00 (0.86)	2.50 (0.76)
AN-CONTRAPOSITION	3.35 (1.14)	3.90 (1.41)	2.20 (0.89)

Note. The tasks were rated by the subjects on a scale from one (very comprehensible, very certain, and very difficult, respectively) to five. The *SD*s are in the parentheses.

Table 11

Subjects' mean lower and upper bound responses in the TRANSITIVITY *(n = 20) and in the* MODUS BARBARA *(n = 20) condition of Experiment 3.*

		A1L	*A1U*	*A2L*	*A2U*	*A3L*	*A3U*	*B1L*
TRANSITIVITY	*Mean:*	71.20	83.60	60.27	66.71	58.41	79.34	41.15
	SD:	25.17	20.30	6.11	15.13	15.10	20.47	18.39
MODUS BARBARA	*Mean:*	71.45	84.95	61.05	71.95	64.79	72.00	44.15
	SD:	27.10	10.88	16.93	16.69	26.08	19.93	16.72
		B1U	*B2L*	*B2U*	*B3L*	*B3U*	*B4L*	*B4U*
TRANSITIVITY	*Mean:*	67.00	45.14	63.77	45.92	58.45	50.06	78.26
	SD:	23.29	15.94	21.15	11.59	15.76	20.14	22.54
MODUS BARBARA	*Mean:*	55.75	52.00	63.65	61.05	70.10	48.85	63.80
	SD:	22.47	20.08	24.09	24.91	23.96	16.23	19.00
		B5L	*B5U*	*B6L*	*B6U*	*B7L*	*B7U*	*B8L*
TRANSITIVITY	*Mean:*	43.06	67.96	42.25	65.95	95.00	100.00	33.25
	SD:	19.24	21.70	19.31	22.80	22.36	0.00	20.00
MODUS BARBARA	*Mean:*	47.95	60.65	49.15	58.75	95.00	100.00	39.90
	SD:	16.10	23.84	25.29	28.91	22.36	0.00	21.02
		B8U	*B9L*	*B9U*	*B10L*	*B10U*	*B11L*	*B11U*
TRANSITIVITY	*Mean:*	60.35	65.63	78.25	40.68	66.97	61.23	79.45
	SD:	25.56	21.39	18.26	17.15	23.95	25.57	20.07
MODUS BARBARA	*Mean:*	53.80	71.50	79.90	42.25	57.20	66.45	76.75
	SD:	31.75	20.50	15.88	15.68	26.43	18.18	15.42

Note. *L* and *U* designate subjects' lower and upper bound responses, respectively. The normative lower and upper bounds are in all tasks 0 and 100, respectively. The percentages in the premises are the identical to those in Experiment 1 (see the values in square brackets in Table 5).

Table 12

Mean Frequencies and percentages of the interval responses in the TRANSITIVITY *($n = 20$) and in the* MODUS BARBARA *($n = 20$) condition assumed that the argument forms are interpreted as* CUT *(Experiment 3).*

TRANSITIVITY (order of premises: $A \mathrel{|\!\sim} B$, $B \mathrel{|\!\sim} C$)			
	LB	LW	LA
UA	5.00%	20.00%	2.50%
	$M = 1.00$ ($SD = 1.04$)	$M = 4.00$ ($SD = 3.70$)	$M = 0.50$ ($SD = 0.52$)
UW	3.93%	62.14%	—
	$M = 0.79$ ($SD = 1.05$)	$M = 12.43$ ($SD = 5.18$)	—
UB	6.43%	—	—
	$M = 1.29$ ($SD = 0.73$)	—	—
MODUS BARBARA (order of premises: $B \mathrel{|\!\sim} C$, $A \mathrel{|\!\sim} B$)			
	LB	LW	LA
UA	0.36%	0.36%	0.36%
	$M = 0.07$ ($SD = 0.27$)	$M = 0.07$ ($SD = 0.27$)	$M = 0.07$ ($SD = 0.27$)
UW	10.71%	50.00%	—
	$M = 2.14$ ($SD = 1.03$)	$M = 10.00$ ($SD = 4.76$)	—
UB	40.00%	—	—
	$M = 8.00$ ($SD = 4.24$)	—	—

Note. Explanation of the abbreviations see Table 5. The mean values of each interval response category are calculated over the 14 tasks.

Table 13
Subjects' mean lower and upper bound responses in the CUT *(n = 16) condition and the* TRANSITIVITY*-condition (n = 19) of Experiment 4.*

		A1L	*A1U*	*A2L*	*A2U*	*A3L*	*A3U*	*B1L*
CUT	*Mean:*	72.38 (80.99)	81.75 (89.99)	57.25 (62.37)	71.12 (99.37)	58.00 (62.72)	68.12 (64.72)	37.62 (42.00)
	SD:	22.60	12.11	15.37	16.84	19.36	16.94	14.98
TRANSITIVITY	*Mean:*	59.42 (0.00)	86.05 (100)	43.70 (0.00)	72.88 (100)	50.25 (0.00)	74.11 (100)	30.05 (0.00)
	SD:	36.55	5.38	27.80	16.49	30.07	17.31	22.52
		B1U	*B2L*	*B2U*	*B3L*	*B3U*	*B4L*	*B4U*
CUT	*Mean:*	48.19 (72.00)	40.88 (45.36)	50.69 (82.36)	44.06 (49.50)	52.44 (94.50)	45.62 (55.44)	54.56 (56.44)
	SD:	20.86	15.63	20.82	13.18	14.49	18.11	19.18
TRANSITIVITY	*Mean:*	52.73 (100)	33.30 (0.00)	55.18 (100)	35.03 (0.00)	59.03 (100)	43.34 (0.00)	61.54 (100)
	SD:	17.66	23.03	18.27	20.88	21.64	30.56	17.19
		B5L	*B5U*	*B6L*	*B6U*	*B7L*	*B7U*	*B8L*
CUT	*Mean:*	40.69 (45.36)	49.81 (73.36)	43.81 (36.00)	53.75 (76.00)	93.75 (100)	93.75 (100)	36.38 (26.01)
	SD:	16.39	20.41	17.07	21.06	25.00	25.00	16.01
TRANSITIVITY	*Mean:*	34.55 (0.00)	58.37 (100)	29.05 (0.00)	48.84 (100)	77.95 (0.00)	94.74 (100)	21.55 (0.00)
	SD:	25.64	16.34	22.96	19.38	37.98	15.77	20.88
		B8U	*B9L*	*B9U*	*B10L*	*B10U*	*B11L*	*B11U*
CUT	*Mean:*	46.25 (75.01)	56.50 (69.52)	65.31 (90.52)	37.88 (43.12)	47.44 (66.12)	55.69 (69.52)	64.12 (81.52)
	SD:	24.07	23.95	22.46	13.81	19.61	25.08	23.58
TRANSITIVITY	*Mean:*	39.08 (100)	49.92 (0.00)	76.13 (100)	41.74 (0.00)	53.76 (100)	47.76 (0.00)	76.03 (100)
	SD:	22.37	31.75	13.42	52.13	17.84	33.43	12.73

Note. *L* and *U* designate subjects' lower and upper bound responses, respectively. The normative lower and upper bounds are given in parentheses.

Table 14
Frequencies of the interval responses in the CUT *condition of Experiment 4 ($n = 16$).*

	Task A1 (80.99-89.99)			*A2 (62.37-99.37)*			*A3 (62.72-64.72)*			*B1 (42.00-72.00)*			*B2 (45.36-82.36)*		
	[89 and 91]			*[99 and 63]*			*[64 and 98]*			*[60 and 70]*			*[72 and 63]*		
UA	1	0	0	0	0	0	1	2	1	1	1	0	1	1	0
UW	1	**8**	-	2	**13**	-	1	**10**	-	0	**12**	-	1	**10**	-
UB	6	-	-	1	-	-	1	-	-	2	-	-	3	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task B3 (49.50-94.50)			*B4 (55.44-56.44)*			*B5 (45.36-73.36)*			*B6 (36.00-76.00)*			*B7 (100-100)*		
	[90 and 55]			*[56 and 99]*			*[63 and 72]*			*[60 and 60]*			*[100 and 100]*		
UA	1	0	0	1	1	0	1	1	0	1	1	0	-	-	-
UW	0	**13**	-	3	**7**	-	1	**10**	-	2	**12**	-	0	**15**	-
UB	2	-	-	4	-	-	3	-	-	0	-	-	1	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task B8 (26.01-75.01)			*B9 (69.52-90.52)*			*B10 (43.12-66.12)*			*B11 (69.52-81.52)*			***Task***		
	[51 and 51]			*[88 and 79]*			*[56 and 77]*			*[79 and 88]*			*[**Pr.1** and **Pr.2**]*		
UA	1	1	0	1	0	0	1	1	0	1	1	0	***a***	***b***	***c***
UW	1	**13**	-	2	**7**	-	1	**10**	-	1	**7**	-	***d***	***e***	-
UB	0	-	-	6	-	-	3	-	-	6	-	-	***f***	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*

Note. The percentages presented in the premises are in the square brackets and the normative intervals are in the round parentheses. For the abbreviations see Table 5.

Table 15

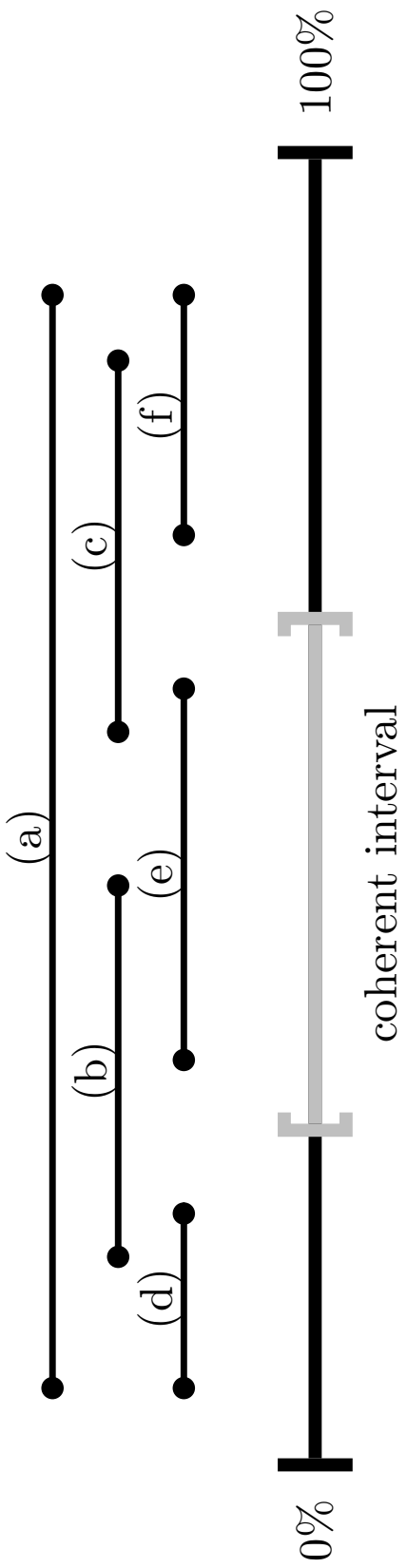
Mean interval responses in the tasks with sure premises.

	Lower bound		Upper bound		
	M	*SD*	*M*	*SD*	*n*
CAUTIOUS MONOTONICITY (Experiment 1)	100	0.00	100	0.00	19
CUT (Experiment 2)	95.05	22.14	100	0.00	20
CUT (Experiment 4)	93.75	25.00	93.75	25.00	16
RIGHT WEAKENING (Experiment 2)	95.00	22.36	100	0.00	20
AND (Pfeifer, 2002)	75.30	43.35	90.25	29.66	40
AND (Pfeifer, 2002)	87.18	33.87	97.44	16.01	39
OR	99.63	1.83	99.97	0.18	30
MONOTONICITY (Experiment 1)	41.25	46.63	92.10	19.31	20
TRANSITIVITY (Experiment 3)	95.00	22.36	100	0.00	20
MODUS BARBARA (Experiment 3)	95.00	22.36	100	0.00	20
TRANSITIVITY (Experiment 4)	77.95	37.98	94.74	15.77	19

Note. The data of the OR task stems from an unpublished study. *n* denotes the sample size. In all tasks the probabilities in the premises were equal to 100%. For the SYSTEM P rules above the line the coherent probability of the conclusion equals to 100%. For the monotone argument forms below the line (MONOTONICITY, TRANSITIVITY, and MODUS BARBARA) the coherent probability interval of the conclusion is the unit interval, 0–100%.

Figure Captions

Figure 1. Categories of possible interval responses. Coherent interval responses (e). Incoherent interval responses: too wide intervals (a), intervals with both bounds too low (d) or too high (f), and intervals where only the upper (b) or only the lower bound (c) are coherent.


100%
(a)
(b)
(c)
(d)
(e)
(f)
coherent interval
0%