Annals of Mathematics and Artificial Intelligence **34:** 107–130, 2002.

An empirical test of patterns for nonmonotonic inference

Rui Da Silva Neves, Jean-François Bonnefon and Eric Raufaste

Université de Toulouse-Le Mirail, 5 allées Antonio Machado, 31058 Toulouse Cedex, France
E-mail: {neves, bonnefon, raufaste}@univ-tlse2.fr

Human inference can be used to test the inference patterns a reliable nonmonotonic consequence should satisfy, because it appears to be nonmonotonic, it is adaptive and it generally achieves efficiency. In this study, an experiment is conducted to investigate whether human inference tends to be consistent with rationality postulates (System P plus Rational Monotony), especially when it no longer satisfies the Monotony property. The experimental protocol uses a possibilistic semantics for plausible rules. Our results appear to be consistent with all the studied properties. Exceptions are the Cut property (with one kind of content out of two) and Left Logical Equivalence which could not be tested. Moreover, when Monotony was not satisfied by participants' inferences, Cut, Cautious Monotony and And properties were corroborated (Rational Monotony was only plausibly supported and the other properties were not tested). Our results emphasize the psychological plausibility of rationality postulates and support the working hypothesis in Artificial Intelligence that System P plus Rational Monotony offer a plausible basic set of properties for nonmonotonic logics.

Keywords: nonmonotonic inference, System P, human inference

1. Introduction

Several non-standard logical or probabilistic frameworks of inference from default rules have been proposed. They generally do not satisfy Monotony, contrary to classical logic, and thus have been called "nonmonotonic". Subsequently, the question has been raised of the minimal subset of classical deduction rules that a nonmonotonic consequence relationship should satisfy. Such a subset would enable the comparison and classification of the various actual or potential non-standard systems. This line of research has been initiated by Gabbay [10] and Makinson [15], and followed by Kraus, Lehmann and Magidor [13], Lehmann and Magidor [14], and Gärdenfors and Makinson [11]. In particular, three sets of postulates for nonmonotonic consequence relationships emerged from the work of Lehmann and his colleagues: "System C" (C for Cumulative) [13], "System P" (P for Preferential) [13] and "Rational Closure" [14]. According to [13], the former includes the basic properties without which a system should not be considered a logical system, the second one includes the former and seems to occupy a central position in the hierarchy of nonmonotonic systems (see [13]), and the latter includes System P and a "Rational Monotony" postulate.

In this paper, we are interested in the psychological plausibility of these rationality postulates. System P offers a basic core for commonsense reasoning (e.g., in [1]), which

refers at least implicitly to human reasoning. To our knowledge, no empirical study has explicitly tested the "psychological plausibility" of these postulates. Such a test is potentially relevant to the study of nonmonotonic reasoning from an Artificial Intelligence perspective since human inference exhibits nonmonotony, and human inference generally achieves efficiency. Let us illustrate these points.

First, given the following two premises

- *If she has an essay to write then she will study late at the library*;
- *She has an essay to write*;

and given the following three conclusions

- *She will study late at the library*;
- *She will not study late at the library*;
- *She may or may not study late at the library*.

96% of participants select the conclusion '*She will study late at the library*' (Byrne [4]). Yet, if the additional premise "*If the library stays open then she will study late at the library*" is given to participants, only 38% select the same conclusion. Results of this kind have been widely replicated (e.g., in [3,4,21]) and all replications have shown the same patterns of inference. Consequently, this experiment shows nonmonotony in human reasoning. It also shows the role played by implicit knowledge on explaining nonmonotony. Such an importance has been shown experimentally by Elio and Pelletier [8]. In particular, they found that human default reasoning is influenced by the specificity of the given information, its similarity and the size of considered categories. However, according to these authors, nonmonotonic reasoning by human agents depends not only on contexts and contents but also on syntactic cues.

Second, experiments of this kind exhibit the flexibility of human reasoning. Such a flexibility is necessary in order to be efficient and to cope with the dynamic nature of our environment. Changes may occur very quickly. When driving a car, having a conversation or cooking a meal, we must sometimes react at most in a few seconds to a new piece of information. For example, when driving a car, the perception of a clang on the wheel generally leads to stop the car as quickly as possible. However, if we had experienced such a clang before and attributed its origin to a fine gravel close to the disc, and if we know that the trouble is temporary and without consequence, we would behave monotonically and would not stop the car. This example confirms the importance of implicit knowledge on explaining why reasoning is sometimes monotonic and sometimes not, but it also emphasizes the efficiency of human reasoning. Indeed, in the absence of implicit knowledge that informs us of the irrelevance of the clang in relation to the normal course of things, we would stop the car very quickly and possibly prevent some mechanical disaster. On the other hand, given the knowledge of the origin of the clang, despite some (temporary) anxiety caused by the clang, nothing would change in our behavior and neither time nor money would be spent uselessly.

The objective of this paper is to test whether human inference tends to be consistent with rationality postulates (System P plus Rational Monotony), especially when it no longer satisfies the Monotony property.

In order to achieve this objective, two experiments have been conducted. The first one was devoted to the selection of a set of concrete plausible rules that were used in the second experiment for the psychological study of each property. The methodology that was followed for this selection is presented in section 3.1. The choice of concrete rather than abstract material is discussed in section 2.3. The second experiment involved eighty-eight psychology students. Each participant had to evaluate the plausibility of each rule selected at the previous stage independently. The semantics of the possibilistic inference (see section 2.3) was applied in order to evaluate the status of each rule property by property and participant by participant.

Under the hypothesis that participant's inference tends to be consistent with the studied properties, we predicted that participants which endorse the "condition" rules on the left part of a property endorse also the "conclusion rule" on the right part of this property. Moreover, although the "conclusion" rule on the left part of a property might be inferred from another set of rules, we predicted that the non endorsement of the "condition rules" on the left part should lead to a lower endorsement rate of the "conclusion" rule.

In order to verify that rationality postulates are corroborated by human inference when Monotony is not, the endorsement rates of Monotony and of other properties were compared. Comparisons of special interest are those between Monotony, Cautious Monotony and Rational Monotony. Indeed, Rational Monotony is less cautious than Monotony and allows monotonic inferences from irrelevant conditions (see [14]). Detailed predictions are presented in sections 2.4 and 3.2.

In this study, we do not intend to establish that the studied set of properties is as such sufficient to describe human inference considered in all its complexity and variability. As a consequence, we do not claim that any fair nonmonotonic reasoning system must satisfy all the studied properties and only these. Rather, if human inference is consistent with rationality postulates of System P or Rational Closure System, we will be prone to say that a system that satisfies all the studied properties exhibits some suitable properties in order to be flexible and efficient in the human way.

The layout of this paper is as follows. Section 2 outlines how to build an experiment that tests which properties are corroborated by human inference. In particular, in the sequel of this section, we examine how to design an experimental protocol well-suited to our objective. In section 2.1, we briefly recall rationality postulates in System P and Rational Closure, and the Monotony property. Then, the method and the material adopted for the experiments are discussed in sections 2.2 and 2.3. Next, in section 2.4, the principles of the test are summarized. In section 3, the design and procedure of the experimental test are detailed, and in section 4, experimental results are presented and discussed. Finally, section 5 offers a summary of and a conclusion to the study.

2. Principles of an experimental study of human nonmonotonic inference

Checking whether human inference corroborates (i.e., tends to be consistent with) a given set of properties requires an appropriate experimental setting. No current experimental device can provide relevant and direct observation of the human inferential system "at work". Yet, we are able to observe the conclusions derived by human participants in the context of a given set of premises. This is the usual approach in the study of human deductive reasoning. Experimentalists first have participants endorse some set of premises and, given this set of premises, ask them to complete one of the following tasks, (a) judging the validity of a given conclusion, (b) selecting a correct conclusion from a pool of given conclusions, and (c) deriving a correct conclusion by themselves. Arguments involved in such tasks are usually constructed so that the premises "entail" a conclusion according to some valid (e.g., Modus Ponens, Modus Tollens) or fallacious (e.g., Negation of the Antecedent) rule of inference. For example, using such methodology, it has been demonstrated that Modus Ponens is almost universally endorsed (with a 100% endorsement rate in numerous studies, never below 89%) whereas the endorsement rates of Modus Tollens range from 41% to 81% (see [9]).

Despite its psychological interest, such a protocol may appear ill-adapted to our present purpose. Indeed, even if we expect human inference to corroborate these properties, we know of no sufficient reason to think that lay reasoners would recognize any rationality postulate as valid, neither that they would conscientiously use them to guide their reasoning. In other words, we do not see these patterns as direct inference rules, which could be investigated in the usual way described above, but as general emerging properties of the inferential apparatus. We therefore refer to "the left part" (LP) and "the right part" (RP) of the properties instead of using the terms "premises" and "conclusion" of a pattern.

2.1. *Patterns for nonmonotonic inference*

We adopt the following notational conventions: A plausible rule "if α then plausibly β" is denoted by $\alpha \mid\sim \beta$; the material equivalence is denoted by $\alpha \Leftrightarrow \beta$; the material implication is denoted by $\alpha \Rightarrow \beta$. Other classical propositional connectives are referred to by the symbols $\neg$, $\wedge$, and $\vee$.

System C includes the five following patterns:

1. $\alpha \mid\sim \alpha$ (Reflexivity).
2. $\alpha \wedge \beta \mid\sim \gamma, \alpha \mid\sim \beta \Rightarrow \alpha \mid\sim \gamma$ (CUT).
3. $\alpha \mid\sim \beta, \alpha \mid\sim \gamma \Rightarrow \alpha \wedge \beta \mid\sim \gamma$ (Cautious Monotony: CM).
4. $\models \alpha \Leftrightarrow \beta, \alpha \mid\sim \gamma \Rightarrow \beta \mid\sim \gamma$ (Left Logical Equivalence: LLE).
5. $\models \alpha \Rightarrow \beta, \gamma \mid\sim \alpha \Rightarrow \gamma \mid\sim \beta$ (Right Weakening: RW).

System P includes system C, and the following pattern:

6. $\alpha \mid\sim \gamma, \beta \mid\sim \gamma \Rightarrow \alpha \vee \beta \mid\sim \gamma$ (OR).

Rational Closure system includes patterns of System P, and Rational Monotony.

7. $\alpha \mid\sim \gamma, \neg(\alpha \mid\sim \neg\beta) \Rightarrow \alpha \wedge \beta \mid\sim \gamma$ (Rational Monotony: RM)

An interesting pattern that derives from the axioms of these systems is the conjunction rule AND.

8. $\alpha \mid\sim \beta, \alpha \mid\sim \gamma \Rightarrow \alpha \mid\sim \beta \wedge \gamma$ (AND).

All these patterns except reflexivity are tested in this study. From now on, for notation convenience, we refer to this set of seven "Target Properties" as "TP". Reflexivity has been excluded because of pragmatic reasons that will be discussed in section 3.2.

In addition to TP, we test Monotony:

$$\alpha \mid\sim \gamma \Rightarrow \alpha \wedge \beta \mid\sim \gamma.$$

2.2. *Some candidate methods for the empirical test of TP*

As an example, consider Cautious Monotony (CM: $\alpha \mid\sim \beta, \alpha \mid\sim \gamma \Rightarrow \alpha \wedge \beta \mid\sim \gamma$). Given the material implication between the left and right parts (noted respectively LP and RP) of the property, it can be said that CM is not corroborated when reasoners endorse both $\alpha \mid\sim \beta$ and $\alpha \mid\sim \gamma$, but do not endorse $\alpha \wedge \beta \mid\sim \gamma$. Such reasoners could be said to violate the CM property. However, this analysis assumes that RP should logically follow from LP from reasoners' point of view. Yet, as already mentioned, we doubt that in the normal course of their life people would conscientiously use properties of TP. Instead, we assume that human inference is constrained by knowledge organization in memory and that its formal properties emerge from a spreading activation process operating directly on knowledge structures. We make the hypothesis that this spreading activation process is by and large consistent with TP. This change in point of view has important methodological consequences for the test of the seven properties in TP. Indeed, consider a classical question like "does this conclusion follow from these premises?" and the case of a positive answer . In such a case, whatever the reasons (logically or knowledge based), the response can be said consistent with the property under consideration. But consider now a negative answer. In such a case, if people base their conclusion on the formal properties of the argument, the rejection of the logical link between the conclusion and the premises is inconsistent with the pattern and constitutes a violation of this pattern. On the other hand, if people base their negative answer on the non existence of a "path" in memory between LP and RP (that is the non activation of RP from LP), nothing can be said about the violation or not of the pattern. Indeed, if people do not endorse LP, it is sufficient to reject that "RP follows from LP", without any violation of the pattern. Such default endorsement has been experimentally shown by George [12], although the experimenter explicitly asked the participants to endorse the premises.

At least two alternative methodologies might be appropriate.

The first one makes participants learn a set of unfamiliar instantiated rules that are formally equivalent to the plausible rules in LP. Then we ask the participants to infer a set of new rules that logically follow from the previous ones. Finally, we compare the set of inferred rules with the formal extension of the learned rules.

A second methodology makes participants independently evaluate the plausibility of every rule involved in the properties of TP (regardless of their position in the

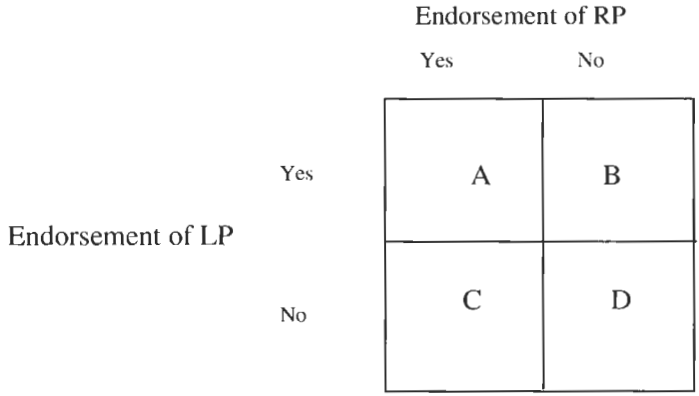

Figure 1. Contingency table.

property), and tests whether the endorsement of LP is preferentially associated with the endorsement of RP. Classically, the information relevant to the test of an association between two variables can be summarized in a contingency table (figure 1), where A, B, C and D represent the number of participants who did endorse, or not, the plausible rules within LP and RP. Let P be the situation where LP is endorsed, and let Q be the situation where RP is endorsed. The degree of association between LP and RP can be computed from the frequencies of "$P \wedge Q$" (the cell A of the contingency table), "$\neg P \wedge \neg Q$" (cell D), "$\neg P \wedge Q$" (cell C), and "$P \wedge \neg Q$" (cell B). From a statistical standpoint, the A and D cells positively contribute to the association hypothesis, whereas the B and C cells contribute negatively. See section 3.2 for details about the statistical test.

The calculation of such a degree of association allows only for the test of an equivalence between LP and RP. Therefore, the finding of a significant degree of association is not sufficient to conclude about a property because this degree can be due to the reverse inference ("RP $\Rightarrow$ LP"). In order to conclude about the "LP $\Rightarrow$ RP" implication, it is necessary to focus on the comparisons between cells A and B, A and C, and B and C. Indeed, whatever the property of TP, LP being endorsed by participants, LP would imply RP preferentially if RP is significantly endorsed rather than non endorsed (i.e., the cell A value is significantly more important than the cell B value). In addition, the inference "RP $\Rightarrow$ LP" has no significant contribution to the association degree if there is no significant difference between cells A and C. Finally, a significant difference between cells B and C (with B greater than C) is an indication of a strongest contribution to the association degree of the "LP $\Rightarrow$ RP" inference than the "RP $\Rightarrow$ LP" inference.

In brief, assuming that all four cells have the same weight in the computation of the degree of association between LP and RP, it is noteworthy that a very low degree of association (close to null or even negative) would lead to the conclusion that participants' inferential processes failed to corroborate the considered property. On the other hand, a statistically significant degree of association is not sufficient to establish that the considered property is corroborated. Thus, in order to conclude about the consistency of

participants inferences with a given property, the following rules are applied:

1. If (LP and RP are not significantly positively associated)
or (A is not significantly greater than B, with B close to 0)
then (the property under consideration is not corroborated).

2. If (LP and RP are significantly positively associated)
and (A is significantly greater than B, with B close to 0)
and (A is not significantly different from C)
then (the property under consideration is corroborated).

3. If (LP and RP are significantly positively associated)
and (A is significantly greater than B, with B close to 0)
and (A is significantly different from C)
and (C is significantly different from B)
then (the property under consideration tends to be corroborated).
In such a case, the inference "RP $\Rightarrow$ LP" contribute also to the association degree but less than the "LP $\Rightarrow$ RP" inference.

In addition, in order to reason at a system level rather than at a single property level, given the separate computation of degrees of association for each property, we must consider which properties are corroborated or rejected on a participant by participant analysis.

2.3. Criteria that the material of the experiment should satisfy

The methodology retained in the previous section enables us to evaluate from a statistical point of view to which degree it is reasonable to accept or not the hypothesis that a given property (Monotony or any property belonging to TP) is corroborated by human inference. Next issue to be dealt with is the choice of appropriate material.

The first question about the experimental material to be used is whether it should be of abstract or concrete nature. Abstract material can only be processed successfully by abstract reasoning rules. Thus, using abstract material would only be relevant if the target properties were considered as actual inference rules applying to all kinds of materials. Again, we do not conceive target properties as actual human reasoning rules. To the contrary, we conceive them as emergent properties of the spontaneous mental activity. Therefore, material must be concrete in order to be implanted to some degree in memory. Now, how many concrete instances of an abstract rule are necessary in order to conduct our test? Two levels must be considered in the analysis. The first one is the rule level. The second one is the property level.

At the rule level, what we need to control is whether rules in LP and RP are endorsed or not. A simple semantics for plausible rules, yet one allowing strict measures,

has been proposed by Benferhat, Dubois and Prade [1], and Dubois and Prade [6] within the framework of Possibility theory.[1] According to this semantics, each conditional assertion $\alpha \mid\sim \beta$ can be viewed as a constraint expressing that a situation where α and β are both true has a greater possibility (Π) than a situation where α is true and β is false – that is, the counter-example situation $\alpha \wedge \neg\beta$ is strictly less possible than the "normal state of affairs", $\alpha \wedge \beta$. Formally, $\alpha \mid\sim \beta$ is a plausible rule iff $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. Using this semantics, it can easily be verified whether a given rule $\alpha \mid\sim \beta$ is plausible or not, for a given participant. It only requires to know the degree to which the participant judges possible that $\alpha \wedge \beta$ is true, and the degree to which the same participant judges possible that $\alpha \wedge \neg\beta$ is true. If the possibility granted to $\alpha \wedge \beta$ is strictly greater than the possibility granted to $\alpha \wedge \neg\beta$, then $\alpha \mid\sim \beta$ is a plausible rule. Moreover, following [1], it can be assumed that a rule is a material implication if the possibility of $\alpha \wedge \neg\beta$ is null, and a rule is a material equivalence if both the possibilities of $\alpha \wedge \neg\beta$ and $\neg\alpha \wedge \beta$ are null. In all remaining situations, the rule is not considered plausible. An experimental device that allows such a test for each rule involved in a given property can be easily constructed. However, can we assume that participants would give "possibility judgments" in the sense of possibility theory? In previous empirical studies that made use of the linguistic marker "possible", it was shown that judgments of uncertainty in lay persons (in [2]) and in medical experts (in [18]) conformed to the axioms of possibility theory.

At the property level, with the experimental procedure and the decision criteria introduced in section 2.2, and with a device that would enable measuring the degrees of possibility related to the rules involved in each property, how many instances of each property would be needed? In the case of abstract material, a single instance might suffice.

Consider now the case of some concrete material: Are there any reasons to think that a property could be endorsed with some concrete material and not endorsed with some other concrete material? The answer is yes, of course. Indeed, suppose that we judge plausible some concrete rule of the form $\alpha \mid\sim \beta$. Because of the plausibility of the rule, in the presence of α and in the absence of any additional knowledge about abnormal conditions or defaults to the rule, whatever the conditions belonging to the normal course of things, we will continue to believe β. However, if some additional information that allows to infer that some default to the rule or abnormal condition is met, then, we shall stop to believe β or shall feel at least serious doubts on its subject.

Now, why should a property be endorsed with some concrete material and not endorsed with some other concrete material? The answer has been given above: because, in the first case, there is no available implicit knowledge that would offer some reasons to defeat the rule, while in the second case, such implicit knowledge is available.

The fact is that a single instance might suffice in the case of abstract material because such a material is not related to any implicit knowledge. Therefore, if some concrete material is such that no implicit knowledge is available, a single instance should suffice also in order to test a given property.

[1] See the appendix for a brief recall.

Consider the following rules, related to knowledge about Lawyers:

"Lawyers do not have their hair dyed in red,"

"Lawyers have a large income,"

and let suppose that these rules are plausible ones. The conclusion

"Lawyers having a large income do not have their hair dyed in red"

normally follows (Cautious Monotony) except if some implicit knowledge is available about a particular kind of lawyers having a large income and having their hair dyed in red. However, implicit knowledge should not be considered if the kind of lawyers under consideration is explicitly the most typical one, which should be the case if we are asked to imagine that the lawyers under consideration had been randomly selected from a phone book.

Our position is that such a demand invites not to refer to implicit knowledge about exceptions to the rules. Consequently, if a large sample of participants (according to statistical criteria) endorse that "lawyers (drawn from a phone book) having a large income do not have their hair dyed in red", it can be attributed to a general property of the thinking of the studied sample of participants.

Nevertheless, the effect of material on properties endorsement rates should be tested. Given the methodology presented at the end of section 2.2, and given only one set of rules for the test of each property, such an opportunity occurs naturally for CM and CUT properties because these two properties involve the same set of formal rules. In addition, the Monotony property might be also tested given the concrete sets of rules necessary to CM and CUT tests. A third test of Monotony is possible with the material used for the test of Rational Monotony.

Finally, using alternate experimental devices instead of testing properties with various materials within the same experimental device would be an interesting way to investigate both content effects and the robustness of results. Such a refinement will be looked for in further experiments.

2.4. *Principle of the test*

(a) *We must be able to infer, for each property and each participant, the status of any rule involved in the test.*

Solution: For each rule $\alpha \mid\sim \beta$, the degrees to which each participant feels possible that $\alpha \wedge \beta$ is true, and that $\alpha \wedge \neg\beta$ is true are measured. Along with possibilistic inference semantics, the status of the plausible rule $\alpha \mid\sim \beta$ is inferred as follows: $\alpha \mid\sim \beta$ is a plausible rule for a given participant iff this participant judges that $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$. Considering the test of the Right Weakening property, $\alpha \mid\sim \beta$ is a material implication $(\alpha \Rightarrow \beta)$ if $\Pi(\alpha \wedge \neg\beta) = 0$. Considering the test of the Left Logical Equivalence, $\alpha \mid\sim \beta$ is a material equivalence $(\alpha \Leftrightarrow \beta)$ if $\Pi(\alpha \wedge \neg\beta) = 0$ and $\Pi(\neg\alpha \wedge \beta) = 0$.

(b) *We must be able to determine whether participants' judgments tend to* corroborate *or not to* corroborate *properties under interest.*

Endorsement of LP in the CM property.	Endorsement of RP in the CM property: Yes	Endorsement of RP in the CM property: No
Yes	Number of participants judging that: $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ $\Pi(\alpha \wedge \gamma) > \Pi(\alpha \wedge \neg\gamma)$ $\Pi(\alpha \wedge \beta \wedge \gamma) > \Pi(\alpha \wedge \beta \wedge \neg\gamma)$	Number of participants judging that: $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ $\Pi(\alpha \wedge \gamma) > \Pi(\alpha \wedge \neg\gamma)$ $\Pi(\alpha \wedge \beta \wedge \gamma) \leq \Pi(\alpha \wedge \beta \wedge \neg\gamma)$
No	Number of participants judging that: $\Pi(\alpha \wedge \beta) \leq \Pi(\alpha \wedge \neg\beta)$ $\Pi(\alpha \wedge \gamma) \leq \Pi(\alpha \wedge \neg\gamma)$ $\Pi(\alpha \wedge \beta \wedge \gamma) > \Pi(\alpha \wedge \beta \wedge \neg\gamma)$ or that: $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ $\Pi(\alpha \wedge \gamma) \leq \Pi(\alpha \wedge \neg\gamma)$ $\Pi(\alpha \wedge \beta \wedge \gamma) > \Pi(\alpha \wedge \beta \wedge \neg\gamma)$ or that: $\Pi(\alpha \wedge \beta) \leq \Pi(\alpha \wedge \neg\beta)$ $\Pi(\alpha \wedge \gamma) > \Pi(\alpha \wedge \neg\gamma)$ $\Pi(\alpha \wedge \beta \wedge \gamma) > \Pi(\alpha \wedge \beta \wedge \neg\gamma)$	Number of participants judging that: $\Pi(\alpha \wedge \beta) \leq \Pi(\alpha \wedge \neg\beta)$ $\Pi(\alpha \wedge \gamma) \leq \Pi(\alpha \wedge \neg\gamma)$ $\Pi(\alpha \wedge \beta \wedge \gamma) \leq \Pi(\alpha \wedge \beta \wedge \neg\gamma)$ or that: $\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$ $\Pi(\alpha \wedge \gamma) \leq \Pi(\alpha \wedge \neg\gamma)$ $\Pi(\alpha \wedge \beta \wedge \gamma) \leq \Pi(\alpha \wedge \beta \wedge \neg\gamma)$ or that: $\Pi(\alpha \wedge \beta) \leq \Pi(\alpha \wedge \neg\beta)$ $\Pi(\alpha \wedge \gamma) > \Pi(\alpha \wedge \neg\gamma)$ $\Pi(\alpha \wedge \beta \wedge \gamma) \leq \Pi(\alpha \wedge \beta \wedge \neg\gamma)$

Figure 2. Attribution of participants to the four cells of the contingency table as a function of their judgments – the example of the Cautious Monotony property ($\alpha \vdash \beta; \alpha \vdash \gamma \Rightarrow \alpha \wedge \beta \vdash \gamma$).

The solution includes six steps:

1. For each property, a set of concrete rules that matches the formal set of nonmonotonic entailments involved in the property is chosen. These rules must be such that some people endorse them as plausible and some other people do not.
2. For each rule, a large sample of participants rate the degrees of possibility for $\alpha \wedge \beta$ and $\alpha \wedge \neg\beta$.
3. The status each rule has for each participant is determined according to the semantics of the possibilistic inference, as exposed in step (a) above.

4. For each target property, there is a repartition of the participants within each cell of a contingency table:

   A Those who endorse LP and RP;

   B Those who endorse LP but not RP;

   C Those who do not endorse LP but endorse RP;

   D Those who endorse neither LP nor RP.

5. The coefficient of association between the variables "endorsement of LP" (Yes versus No) and "endorsement of RP" (Yes versus No) is computed. In addition, differences between cells *A* and *B*, cells *A* and *C*, and cells *B* and *C* are also computed.

6. A conclusion is drawn as a function of the value of this coefficient and these differences (details related to this step are given in sections 2.2 and 3.2).

2.5. *Summary*

First, we assume that human inference is constrained by knowledge organization in memory and that its formal properties emerge from a spreading activation process operating directly on knowledge structures. We make the hypothesis that this spreading activation process is by and large consistent with TP. The test of the consistence of the spreading activation process with TP includes the following steps:

1. For each property belonging to TP, we chose a set of concrete rules $\alpha_i \mid\sim \beta_i$ that matches the formal set of nonmonotonic entailments involved in the property. Each rule must be such that some people endorse it as plausible and some other people do not. Using abstract material would only be relevant if the target properties were considered as actual inference rules applying to all kinds of materials. Because we do not conceive target properties as actual human reasoning rules but as emergent properties of the spontaneous mental activity, the material must be concrete in order to be implanted to some degree in memory. In addition, we assume that a single concrete instance should suffice in order to test a given property if the concrete material is such that no implicit knowledge is available or if the instances of the rules refer explicitly to the most typical one (see section 2.3). Nevertheless, the effect of material on the endorsement of the properties must be tested.

2. Participants independently rate the degrees of possibility for $\alpha_i \wedge \beta_i$ and $\alpha_i \wedge \neg\beta_i$. It allows to evaluate the plausibility of every rule involved in the properties of TP (regardless of their position in the property) for each participant according to the semantics of the possibilistic inference (see section 2.3).

3. For each property, (i) a contingency table is constructed according to the principles given in section 2.4, (ii) the coefficient of association between the variables "endorsement of LP" and "endorsement of RP", and the differences between cells *A* and *B*, cells *A* and *C*, and cells *B* and *C* are computed.

4. The consistency of human inference (i.e., the spreading activation process) with a property is checked according to the set of rules described at the end of the section 2.2.

5. In order to test the psychological validity of system C, System P and Rational Closure, we must observe that the sets of properties involved in these systems are corroborated by human inference when the monotony property is not satisfied.

3. Experiment

According to the procedure defined in the previous section, a set of rules was chosen to investigate each property. The selected rules were such that some participants judged them plausible and other participants did not, "judging a rule as plausible" being defined according to the semantics of the possibilistic inference. This choice was made by means of the following pre-experiment.

3.1. Experimental protocol for the pre-selection of the rules

We first designed (on an intuitive basis, and following the syntax of the rules involved in the left part of the seven properties) a large set of rules that could potentially be interpreted as plausible rules by a sample drawn from the student population of the University of Toulouse-Le Mirail.

Participants

Forty First-year Psychology students at the University of Toulouse-Le Mirail, all native speakers in French, contributed to this study. None of them had previously received any formal training in logic.

Material

The material consisted of 24 pairs of questions related to 24 rules of the kind "Vegetarians do not enjoy bullfights". Each question involved an unknown character (e.g., Mathilde B., Simon A., ...). Participants were asked to imagine that these characters had been randomly selected from a phone book. Participants had to answer pairs of questions like "To which degree do you judge possible that Simon A. is a vegetarian and enjoys bullfights?", and "To which degree do you judge possible that Simon A. is a vegetarian and does not enjoy bullfights?" Participants were invited to draw their judgments of possibility upon an axis with no graduation as shown by figure 3.

Design and procedure

Questions were presented within a booklet, in random order for half the participants and in reverse order for the other half. Participants were informed that they could answer by either checking the point of the axis that best matched their judgment, or drawing an

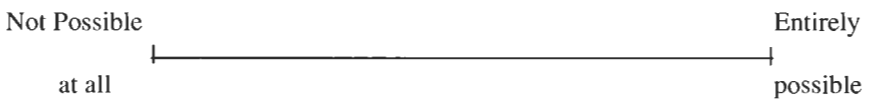

Figure 3. Axis for participants' answers.

ellipsoid around the portion of the axis that contained their answer, if they had only an imprecise idea of the localization of this answer.

Results

We first applied an eleven-point scale grid to the response axis and then encoded participants' response as an interval $[\Pi_{inf}, \Pi_{sup}]$. If the answer was an ellipsoid, the lower and upper values of the interval were given by the graduations that were closest to the intersection point between the ellipsoid and the axis. When the intersection point was equidistant from two graduations, the retained value was the lowest one in the case of Π_{inf}, and the upper one in the other case. When the answer was a single check, the closest graduation provided both the lower and upper values of the interval. When the check was equidistant from two graduations, the left one gave the lower value and the right one gave the upper one. We then computed the min value of each interval, its average value and its max value. Because subsequent treatments led to the same pattern of results for these three measures, we limit our presentation to the average model.

For each participant, and each rule (e.g., "Vegetarians do not enjoy bullfights"), we applied the criteria derived from the possibilistic semantics to the possibility judgments expressed as answers to the questions (e.g., "to which degree do you judge possible that Simon A. is a vegetarian and enjoys bullfights?"; and "to which degree do you judge possible that Simon A. is a vegetarian and does not enjoy bullfights?").

For each rule, we computed the percentages of participants who interpreted the rule as:

- a plausible rule ($\Pi(\alpha \wedge \beta) > \Pi(\alpha \wedge \neg\beta)$),
- a non plausible rule ($\Pi(\alpha \wedge \beta) \leqslant \Pi(\alpha \wedge \neg\beta)$),
- a material implication ($\Pi(\alpha \wedge \neg\beta) = 0$),
- and a material equivalence ($\Pi(\alpha \wedge \neg\beta) = 0$ and $\Pi(\neg\alpha \wedge \beta) = 0$).

On the basis of these results, we selected some pairs of rules so that the product of their rates of endorsement would be close to 0.5. Selected rules (translated from French) appear in table 1, with LP in normal case and expected RP in italics.

3.2. *Experimental testing of TP*

Participants

Eighty-eight first-year psychology students at the University of Toulouse-Le Mirail, all native speakers in French, took part in this study. None of them had received any logical training. These participants did not take part in the pre-experiment.

Material

TP included seven properties, CUT, CM, LLE, RW, OR, RM, and AND. Each involved three rules: two rules in LP and one in RP. The 14 LP rules were selected from the pre-experiment. The 7 RP rules were derived according to target properties. The status these 21 rules had for each participant was checked by means of two questions.

Table 1
Rules selected to test CUT, CM, LLE, RW, OR, RM and AND properties.

CUT		
	Smokers have some light most of the time. (PR)	
	Smokers having some light most of the time rarely ask for light. (PR)	RP
	Smokers rarely ask for a light.	LP
CM		
	Lawyers do not have their hair dyed in red. (PR)	LP
	Lawyers have a large income. (PR)	
	Lawyer having a large income do not have their hair dyed in red.	RP
LLE		
	Vegetarians do not eat meat. (ME)	
	Vegetarians do not enjoy bullfights. (PR)	
	People who do not eat meat do not enjoy bullfights.	
RW		
	Students entering high school are minors. (PR)	
	Minors do not have the right to vote. (MI)	
	Students entering high school do not have the right to vote.	
OR		
	People eating a lot of chocolate and candies have cavities. (PR)	
	People who rarely brush their teeth have cavities. (PR)	
	People eating a lot of chocolate and candies or who rarely brush theirs teeth (or both) have cavities.	
RM		
	Lawyers do not speak Italian. (NPR)	
	Lawyers have a large income. (PR)	LP
	Lawyers who speak Italian have a large income.	RP
AND		
	Lawyers do not have their hair dyed in red. (PR)	
	Lawyers have a large income. (PR)	
	Lawyers do not have their hair dyed in red and have a large income.	

Rules in italic were expected to be the right part of the property. "PR" stands for "plausible rule", "MI" for "material implication", "ME" for "material equivalence", and "NPR" for "non plausible rule". In the CUT, CM and RM properties, LP and RP indicate the rules used for the test of the Monotony property.

Consequently, the material consisted of 42 questions. The Reflexivity property was not included in TP because questions of the kind "to which degree do you judge possible that Simon A. is a minor and is a minor?" would make no sense for participants. The test of Monotony was based on CUT, CM and RM materials.

Design and procedure

Design and procedure were the same as in the pre-experiment. The 42 questions participants had to answer were presented within booklet form. In order to control an order effect of questions' presentation, the 42 questions were presented in random order for half of the participants and in reverse order for the other half.

As an example, the following questions for the test of Right Weakening were scattered across the answer booklet:

"To which degree do you judge possible that Christophe C. is entering high school and is minor?",

"To which degree do you judge possible that Christophe C. is entering high school and is not minor?",

"To which degree do you judge possible that Christophe C. is minor and has the right to vote?",

"To which degree do you judge possible that Christophe C. is minor and has no right to vote?",

"To which degree do you judge possible that Christophe C. is entering high school and has the right to vote?"

"To which degree do you judge possible that Christophe C. is entering high school and has not the right to vote?".

Predictions

The encoding of answers was the same as in the pre-experiment. We wanted to know whether the endorsement of LP (ELP) was preferentially associated with the endorsement of RP (ERP) as described in section 2.4. The degree of association between ELP and ERP was computed by means of the Phi coefficient (ϕ) and statistically tested by means of Chi-square (χ^2) (see [19]).

The coefficient ϕ for a 2×2 table (see figure 1) is defined as

$$\phi = \frac{|AD - BC|}{((A + B)(C + D)(A + C)(B + D))^{1/2}}.$$

For each property, the statistical hypothesis H_0 (null hypothesis) was "there is no significant association between ELP and ERP". The alternate hypothesis, H_1, was "there is a significant positive association between ELP and ERP". H_0 was rejected if the probability of obtaining a value as large as the observed χ^2 was not greater than 0.05, as it is usual in experimental psychology.

Under the general hypothesis that participants' inference tends to corroborate CUT, CM, LLE, RW, OR, RM, and AND, we predicted that H_0 would be rejected for each property.

Moreover, according to the decision criteria introduced in section 2.4, we computed a χ^2 coefficient in order to test the differences between cells A and B, cells A and C, and cells B and C. Finally, under the general hypothesis that human inference is consistent with rationality postulates (System P plus Rational Monotony), but not with Monotony property, we predicted that H_0 could not be rejected for Monotony property.

However, given the independence of statistical tests for each property, each single property could be statistically corroborated while being rejected by a non negligible number of participants. Under the hypothesis that participant judgments are consistent with TP, we predicted that a high proportion of participants would not commit any violation.

Experimental results

We computed ELP and ERP rates for each property belonging to TP, and gathered the results in seven contingency tables, one per property (see figure 4). CUT and CM have special contingency tables because the material used to test CUT can also be used to test CM, and vice versa: numbers between parentheses in the contingency table of CUT refer to the material that was primarily used to test CM, and vice versa. In addition, we computed ELP and ERP rates for Monotony with CUT, CM and RM materials (see figure 4).

A table by table examination of results shows that a problem occurred with LLE. Although few violations of this property have been made, no participant endorsed both LP and RP of LLE. This suggests a problem with the plausibility of the left part of the property that was not detected during the pre-experiment. Another explanation is that the sentence "vegetarians do not eat meat" was not felt by subjects as an equivalence. Indeed, not to eat meat may be the consequence of a medical recommendation.

Over all target properties (except monotony), strict violations – that is endorsements of LP but not of RP (B cells) – were only 8.7%, and 8.4% without LLE. Endorsements of both LP and RP (A cells) were 46.3%, and 51.8% without LLE. Non endorsements of both RP and LP (D cells) were 26.3%, and 19.2% without LLE. Finally, non endorsements of LP and endorsement of RP (C cells) were 18.7%, and 21.2% without LLE. As a whole, endorsement rates of LP were as expected from the pre-experiment.

Focusing on TP properties, the ϕ degrees of association reported in table 2 show a significant association between LP and RP for all the properties belonging to TP – except for LLE – whatever the material used. The probability of mistakenly rejecting H_0 is lower than 0.05 in each case (excluding LLE, though), and even lower than 0.01 in most cases.

Focusing on Monotony, table 2 shows that there is no significant degree of association between LP and RP for the Monotony property tested with CM (Monotony2) and RM material (Monotony3). This is to compare with CM and RM significant degrees of association. On the other hand, the Monotony property tested with CUT material (Monotony1) exhibits a significant level of association at a level close to CM with the same material.

As a consequence, we may reject the null hypothesis that there is no significant association between the left and right parts of CUT, CM, RW, OR, RM and AND. CUT and CM are of a particular interest because their test has been conducted with two different sets of rules. Although both sets lead to significant results, table 2 and figure 4 show very different patterns of endorsement as a function of material. The material used to test a property then appears to influence participants' pattern of endorsement.

The material effect is confirmed with the test of the Monotony property. The null hypothesis that there is no significant association between the left and right parts of Monotony is rejected only in the case of CUT material (Monotony1).

According to our decision criteria, in addition to a significant degree of association between LP and RP, in order to conclude that a property was corroborated by human

CUT

24 (59)	27 (3)
6 (20)	31 (6)

Cautious Monotony

59 (24)	4 (4)
18 (32)	7 (28)

Right Weakening

38	11
22	17

Left Logical Equivalence

0	8
8	72

OR

50	2
20	16

Rational Monotony

47	10
15	16

AND

61	2
13	12

Monotony1 (CUT material)

28	2
38	20

Monotony2 (CM material)

62	6
15	5

Monotony3 (RM material)

52	24
10	2

Figure 4. Contingency tables for each property. In each table, the upper left cell figures the frequency of endorsing both LP and RP. The upper right cell figures the frequency of endorsing LP but not RP. The lower left cell figures the frequency of endorsing RP but not LP. The lower right cell figures the frequency of endorsing neither LP nor RP. In each table, the total number of participants is $N = 88$.

inference, cells A and B ($A > B$ with B close to 0) must be significantly different. Figure 4 examination shows that cell B has the lowest value in all contingency tables except in the test of CUT1, LLE, Monotony2 and Monotony3. In addition, table 3 shows that all the computed differences between cells A and B are significant except

Table 2
Values of ϕ and significance of the association between LP and RP of each property ($N = 88$). LLE could not relevantly be tested because no participant simultaneously endorsed LP and RP. "ns" means that the ϕ value is nonsignificant at the 0.05 level.

	ϕ	Significance
CUT	0.32 (0.27)	0.002 (0.01)
Monotony1 (with CUT material)	0.30	0.004
CM	0.29 (0.31)	0.006 (0.002)
Monotony2 (with CM material)	0.20	ns
RW	0.22	0.034
LLE	–	–
OR	0.49	0.000
RM	0.36	0.001
Monotony3 (with RM material)	−0.11	ns
AND	0.55	0.000

Table 3
χ^2 values and significance (in brackets) of the differences between cells A and B, A and C, and B and C. $N = 88$ and $DF = 1$ for all the comparisons. "ns" means that the difference is nonsignificant at the 0.05 level.

	A vs. B	A vs. C	B vs. C
CUT1	ns	10.8 (0.001)	13.4 (0.000)
CUT2 (with CM material)	50.6 (0.000)	19.3 (0.000)	12.6 (0.000)
Monotony1 (Cut material)	22.5 (0.000)	ns	32.4 (0.000)
CM1	48 (0.000)	21.8 (0.000)	8.9 (0.003)
CM2 (with CUT material)	14.3 (0.000)	1.14 (0.28)	21.8 (0.000)
Monotony2 (CM material)	46.1 (0.000)	28.7 (0.000)	3.9 (0.05)
RW	14.8 (0.000)	4.3 (0.039)	3.7 (0.05)
OR	44.3 (000)	12.9 (0.000)	14.7 (0.000)
RM	24 (0.000)	16.6 (0.000)	ns
Monotony3 (RM material)	10.3 (0.001)	28.5 (0.000)	5.8 (0.016)
AND	55.3 (0.000)	31.1 (0.000)	8.1 (0.004)

for CUT1. It means that, for all these properties, when participants accepted the rules on the left part, they tended also to accept the rules on the right part. Moreover, all the computed differences between cells A and C are significant except for Monotony1 and CM2. It means that participants who accepted rules on the right part tended also

Table 4
Percentages of participants as a function of the number of violations ($N = 88$).

Number of violations	0	1	2	3	4
Percentage of participants	47	37	14	1	1
Cumulated percentages of participants	47	84	98	99	100

to accept rules on the left part. However, the comparisons between "A vs. B" and "A vs. C" columns in table 3 show that χ^2 computed for A and C cells are lower than χ^2 computed for A and C cells except for CUT1 and Monotony3. Finally, all the computed differences between cells B and C are significant at the 0.05 level except for RW and RM.

Given these results and the decision rules in section 2.2:

- LLE, CUT1, Monotony2 and Monotony3 were not corroborated by participants' inference (ϕ degrees not significant).
- Monotony1 and CM2 were corroborated (ϕ to significant, A significantly greater than B with B close to zero, and A not significantly greater than C).
- CUT2, CM1, RW, OR and AND were corroborated (ϕ significant, A significantly greater than B with B close to zero, A significantly greater than C but C significantly greater than B).
- RM was not corroborated according to our criteria because cell C value is not significantly greater than cell B value. However, the χ^2 computed for A and B is greater than the χ^2 computed for A and C, and cell C value is greater than cell B value.

Focusing at the system level, and no longer at the individual property level, we tallied how many violations (out of 7 possible) each participant made. In order to keep all properties equally weighted, we only considered the material that was explicitly designed for each property. That is, we did not consider the material for CUT when looking for violation of CM, and vice versa. Otherwise, CUT and CM would have been double-weighted.

Results appear in table 4. A proportion of 46.6% of participants made no violation at all, and 84.1% participants made one violation or none. These percentages are quite high if we consider the unavoidable imperfection of our experimental apparatus, material, and sampling of participants. They suggest that most participants draw inferences from their own knowledge in a manner that is consistent with TP.

In summary, our results have exhibited that whatever the considered material, CM, RW, OR and AND properties were fully corroborated by participants' judgments. CUT was only with one kind of material and RM was not fully corroborated according to our criteria. Moreover, Monotony was corroborated with one kind of material out of three (CUT material). As a consequence, the material used to test a property appeared to influence participants' pattern of endorsement. Finally, a proportion of 46.6% of participants made no violation at all, and 84.1% participants made one violation or none. These percentages suggest that most participants draw inferences from their own knowledge in a manner that is consistent with TP. These results are discussed in the next section.

4. Interpretation of results

Our results have indicated that CM (whatever the material), RW, OR and AND were fully corroborated by participants' inference. CUT was fully corroborated with only one kind of material (CUT2) because with the other kind of material (CUT1) an important proportion (30%) of participants who endorse the rules on the left part did not endorse the rule on the right part. Moreover, with this material, participants who endorse the rule on the right part also tend to endorse rules on the left part. In other words, participants who judge a rule of the form $\alpha \mid\sim \gamma$ plausible also endorse $\alpha \wedge \beta \mid\sim \gamma$ (Monotony) and $\alpha \mid\sim \beta$. This result is consistent with the finding that Monotony1 is corroborated. Moreover, it seems intuitively consistent that *Smokers that rarely ask for a light* ($\alpha \mid\sim \gamma$) *have some light most of the time* ($\alpha \mid\sim \beta$)". A plausible explanation for the non corroboration of CUT1 is that participants believed that *Smokers have some light most of the time* but experienced (in a French University) that *Smokers ask quite often for a light*. The fact is that the impression of "often" may come from a large sample of smokers rather than from only one. Another material effect seems to have occurred with LLE. Indeed, despite the pre-experiment devoted to the selection of the material on the left part of the properties, no participant endorsed the two rules on the left part of LLE. This property should be tested in a new experiment.

Another property to discuss is the RM property. Indeed, results have exhibited an association between the left and right parts of RM. Moreover, cell A represents 53% of the participants' responses and the percentage of participants that endorsed the left part but not the right part (cell B) of the property is lower than the percentage of participants that endorse its right part but not its left part (cell C). These percentages represent 11% and 17% of participants' responses, which is not significantly different. Thus, our results are consistent with RM but they are also consistent with the inference of LP from RP. It infers $\alpha \mid\sim \gamma$ from $\alpha \wedge \beta \mid\sim \gamma$, when it is endorsed that from α we cannot plausibly infer non β. This inference is an acceptable inference.

In addition, at the System level, we found that about half of our participants made judgments that were consistent with TP, and that about 85% of participants made judgments that did not violate more than one property. It is noteworthy that the violated properties differed from one participant to another. This is an important result because the probabilities to observe no violation or only one violation would be low if human inference was not consistent with TP. Indeed, from a logical standpoint, a single violation is sufficient to conclude that the system as a whole is not corroborated. However, the variability that is inherent in the human cognitive system can explain most violation cases. Actually, it would have been very surprising not to observe such violations. Yet, the relatively important number of violations observed with CUT1 suggests that an explanation based on variability is not sufficient, which has been already discussed.

So, our results appear to be consistent with all the studied properties, except with the CUT property with one kind of material (and of course except LLE). However, a question remains: what happens when Monotony is not corroborated? Such a case has been observed twice with Monotony2 and Monotony3 materials, materials which are

not independent and which allowed the test of CM1 and CUT2. Moreover, they were also related to the material used for the test of AND and RM. Our results show that, on the one hand, Monotony2 and Monotony3 were rejected, and, on the other hand, CM1, CUT2, and AND were corroborated and RM was not rejected.

These results emphasize the psychological plausibility of rationality postulates and give support to the working hypothesis in Artificial Intelligence that System P plus Rational Monotony offer a plausible basic set of properties for nonmonotonic logics.

5. Conclusion

The objective of this paper was to test whether human inference tends to be consistent with rationality postulates (System P plus Rational Monotony), especially when it no longer satisfies the Monotony property. In order to achieve this objective, we first conducted a pre-experiment devoted to the selection of a set of concrete plausible rules used in a second experiment for the investigation of the psychological plausibility of System P plus Rational Monotony. The selected rules were such that they formally fit the left part of a considered property. Moreover, they were such that some participants judged them plausible and other participants did not, "judging a rule as plausible" being defined according to the semantics of possibilistic inference. Then, the selected rules were combined according to the general patterns of inference and the obtained conclusion rules were added to the previous ones. In the second experiment, 88 students in Psychology had to evaluate the plausibility of all the selected and inferred rules at the previous stage. The semantics of possibilistic inference was applied again in order to evaluate the status of each rule, property by property, and participants by participants. Then, the following decision rule was applied:

- a property was not corroborated if no association was found between the left part and the right part of the property (condition 1) or if no significant difference was found between the proportions of participants that endorsed both left and right parts of the property and participants that endorsed the left part but not the right part (condition 2),
- a property was corroborated by participants' inferences if neither condition 1 nor 2 occurred and if a significant difference was found between the proportions of participants that (i) endorsed the left part but not the right part of the property and (ii) endorsed the right part but not the left part of the property.

Given these criteria, we found that all the studied properties, except CUT (with one kind of material) and LLE (which could not be tested) were corroborated by participants' inferences. Moreover, our results confirmed that at least CM, CUT (with only one kind of material), AND and plausibly RM were corroborated when Monotony was not.

In addition, we considered the percentages of participants whose judgments were consistent with all properties, with all but one, all but two, and so on. We found that about half our participants made judgments that were fully consistent with TP, and that about 85% of participants made judgments that did not violate more than one property. It is noteworthy that the violated properties differed from one participant to another.

Finally, an influence of the content of our concrete material has been observed. Indeed, the duplicated measures for CUT, CM and Monotony have shown sensibly different distributions of participants within the cells of contingency tables.

To our knowledge, the present contribution is the very first empirical testing of System P and Rational Monotony. We do not claim that the question is closed. At least, it is now open. These preliminary results encourage us to engage in the search for new evidence using other materials, with the same experimental device as well as with new ones. We hope that these new studies will lead to conclusive data. However, in the line of the idea expressed by Makinson [16], it may be that there is not any unique set of properties that an efficient human nonmonotonic inference should satisfy. Moreover, it is doubtful that human nonmonotonic inference be always perfectly efficient. It may be also that this kind of work is premature. Indeed, a serious critic against this kind of work is that some progress must be done in the understanding of the principles that actually govern the individual inferences made in nonmonotonic reasoning (see, for example, Pollock [17]) before to test generalizations about the structure of reasoning. Further experiments will certainly need to take these remarks into account. Finally, we believe that, in the long run, this new line of research is of interest for both the psychology and artificial intelligence communities.

Acknowledgements

This research was supported by granted number 90N35/005 of the CNRS – "GIS Science de la Cognition" and by grant number 99001558 of "Conseil Régional de la Région Midi Pyrénées". This paper has benefited from fruitful discussions with Salem Benferhat, Didier Dubois and Henri Prade. We are also indebted to the three anonymous referees for their helpful comments.

Appendix

Possibilistic logic (e.g., Dubois and Prade [7]) is an extension of classical logic where propositions are weighted by possibility and necessity degrees which belong to the [0, 1] interval. A possibility measure $\Pi(A)$ represents the qualitative degree to which the proposition A belongs to the normal course of things. $\Pi(A) = 0$ means that A is not possible under the current set of information. $\Pi(A) = 1$ means that A is entirely possible. By convention, $\Pi(T) = 1$ and $\Pi(\bot) = 0$, where T and $\bot$ respectively denote tautology and contradiction. The possibility measure satisfies the following axiom: $\Pi(A \vee B) = \max(\Pi(A), \Pi(B))$. This means that the possibility to observe event A, event B or their simultaneous occurrence is given by the higher possibility between A and B. The necessity measure N is associated by duality with a possibility measure by the following axiom: $N(A) = 1 - \Pi(\neg A)$. Moreover, $N(A \wedge B) = \min(N(A), N(B))$. This means that the certainty that A and B are simultaneously true is the lowest of the certainty that A alone is true and the certainty that B alone is true. Possibilistic logic

copes with partial inconsistency, and its inference machinery turns out to express a preferential entailment à la Shoham [20]. Indeed, the semantics of possibilistic logic can be expressed in terms of a complete ordering of interpretations. A conditional knowledge base made of default rules of the form "generally, if α_i then β_i" can be viewed as a set of constraints stating that $\alpha_i \wedge \beta_i$ is strictly more plausible than $\alpha_i \wedge \neg\beta_i$. The default rules can then be turned into possibilistic logic formulas $(\neg\alpha_i \vee \beta_i, w_i)$ were the weight w_i reflects a rule priority, computed from the least informed possibility measure compatible with the set of constraints. Under this ranking of rules, possibilistic logic entailment is equivalent to rational closure entailment. When we consider all the possibility measures compatible with the set of constraints induced by the conditional assertions instead of the least informed one, System P is recovered.

References

[1] S. Benferhat, D. Dubois and H. Prade, Representing default rules in possibilistic logic, in: *Proc. of the 3rd Internat. Conf. on Principles of Knowledge Representation and Reasoning KR'92* (Cambridge, MA, 1995) pp. 673–684.

[2] S. Benferhat, R. Da Silva Neves, D. Dubois, H. Prade and E. Raufaste, Qualitative approaches to reasoning under uncertainty: Formal developments and experimental validations, part 1: Possibility theory, Research Report LTC-CERPP-IRIT/00-17 R (2000).

[3] J.F. Bonnefon and D.J. Hilton, The suppression of Modus Ponens as a case of pragmatic preconditional reasoning. To appear in Thinking and Reasoning.

[4] R.M.J. Byrne, Suppressing valid inferences with conditionals, Cognition 31 (1989) 61–83.

[5] D. Chan and F. Chua, Suppression of valid inferences: syntactic views, mental models, and relative salience, Cognition 53 (1994) 217–238.

[6] D. Dubois and H. Prade, Conditional objects, possibility theory and default rules, in: *Conditionals: From Philosophy to Computer Sciences*, eds. G. Crocco, L. Fariñas del Cerro and A. Herzig (Oxford University Press, Oxford, 1995) pp. 301–336.

[7] D. Dubois and H. Prade, Possibility theory: Qualitative and quantitative aspects, in: *Quantified Representation of Uncertainty and Imprecision*, Handbook of Defeasible Reasoning and Uncertainty Management, Vol. I (Kluwer, Dordrecht, 1998).

[8] R. Elio and F.J. Pelletier, The effect of syntactic form on simple belief revisions and updates, in: *Proceedings of the 16th Annual Conference of the Cognitive Science Society* (Lawrence Erlbaum, Hillsdale, NJ, 1994) pp. 260–265.

[9] J.St.B.T. Evans, S.E. Newstead and R.M.J. Byrne, *Human Reasoning: The Psychology of Deduction* (Lawrence Erlbaum, London, 1993).

[10] D.M. Gabbay, Theoretical foundations for non-monotonic reasoning in expert systems, in: *Logics and Models of Concurrent Systems*, ed. K.R. Apt (Springer, 1985) pp. 439–457.

[11] P. Gärdenfors and D. Makinson, Nonmonotonic inference based on expectations, Artif. Intell. 65 (1994) 197–245.

[12] C. George, The endorsement of the premises: Assumption-based or belief-based reasoning, British J. Psychology 86 (1995) 93–111.

[13] S. Kraus, D. Lehmann and M. Magidor, Nonmonotonic reasoning, preferential models and cumulative logics, Artif. Intell. 44 (1990) 167–207.

[14] D. Lehmann and M. Magidor, What does a conditional knowledge base entail?, Artif. Intell. 55 (1992) 1–60.

[15] D. Makinson, General theory of cumulative inference, in: *Proceedings Second International Workshop on Non-Monotonic Reasoning*, Lectures Notes in Computer Science, eds. M. Reinfrank and J. De Kleer (Springer, Berlin, 1989).

[16] D. Makinson, General patterns in nonmonotonic reasoning, in: *Nonmonotonic and Uncertainty Reasoning*, Handbook of Logic in Artificial Intelligence and Logic Programming, Vol. 3, eds. D.M. Gabbay et al. (Oxford University Press, Oxford, 1994) pp. 35–110.

[17] J.L. Pollock, Defeasible reasoning, Cognitive Sci. 11 (1987) 481–518.

[18] E. Raufaste and R.M. Da Silva Neves, Empirical evaluation of possibility theory in human radiological diagnosis, in: *Proceedings of the 13th Biennal Conference on Artificial Intelligence, ECAI'98*, ed. H. Prade (Wiley, London, 1998) pp. 124–128.

[19] S. Siegel and N.J Castellan, Jr., *Nonparametric Statistics for the Behavioral Sciences* (McGraw-Hill, New York, 1988).

[20] Y. Shoham, A semantical approach to nonmonotonic logics, in: *Proceedings Logics in Computer Science*, Ithaca, NY (1987) 275–279.

[21] R.M. Stevenson and D.E. Over, Deduction from uncertain premises, Quart. J. Experiment. Psychol. 48A (1985) 613–643.