Reasoning with Knights and Knaves: Towards an understanding of reasoning about truth and falsity

Floris van Vugt
floris.van.vugt@ens.fr

January 31, 2009

1 Introduction

At the end of the eighties a type of logical puzzle called *knight and knave* problems(Smullyan, 1987) enters the scene of psychology of reasoning(Rips, 1989). They are staged on an imaginary island where only two kinds of people live: knights, who always tell the truth, and knaves, who always lie. It is furthermore assumed that they have complete and correct knowledge of each other's being knight or knave. A puzzle then consists of a number of utterances from a few of these inhabitants. The task for the reader is to decide of each character whether he or she is a knight or a knave.

To illustrate, consider the example in table 1.

As with all puzzles, there are multiple ways to arrive at the solution. One

Table 1: A sample problem from (Johnson-Laird & Byrne, 1990)

- A: A and B are knaves.
- B: A is a knave.

could start out making the assumption that A is a knight. This means both A and B must be knaves, but that is contrary to our assumption. If however one would have assumed A a knave, then he must be lying, i.e. not both A and B are knaves. Since he himself is a knave by assumption, that leaves as only possibility that B is knight. And indeed what B says is true. Since one of our two assumptions led to a contradiction, we conclude the other is correct and that A is a knave and B knight.

1.1 Problem solvability

First of all, as none of the existing literature has made explicit, the above puzzle is rather unique in that it has one and exactly one solution. As a matter of fact the *solvable* puzzles can be said to reside on a thin line in between what one could call *paradoxical* problems on the one hand and *underspecified* problems on the other.

1.1.1 Paradoxical problems

I will call a problem *paradoxical* if any attribution of knight– and knave–status to the speakers leads to a contradiction. For example,

- A: B is a knight.
- B: A is a knave.

Clearly if A is a knight, then B must be a knight, but at the same time B must then be lying in saying that A is a knave, so there is a contradiction. If A is a knave then he must lie and therefore B is a knave, however what B says is suddenly true.

An even more primitive example of such a paradoxical phrase is the direct translation of the Liar sentence(Kripke, 1975)(Tarski, 1983) which says of itself that it is false,

- A: A is a knave.

1.1.2 Underspecified problems

A problem is *underspecified* if there are multiple attributions of knight– and knave–status to the speakers that are consistent. For example,

- A: B is a knight.
- B: A is a knight.

If A is a knight, then she must be telling the truth, hence B is a knight also, which is consistent with what he says. If A is a knave however, then she must be lying and B therefore is a knave, indeed it is a lie that A is a knight.

The problem is that this sequence cannot be solved on the basis of the utterances we have.

Again, the most primitive form of such sentences is found in Truth–teller sentences(Kripke, 1975)(Tarski, 1983) and its equivalent on the island of knights and knaves would be

- A: A is a knight

1.2 Outline

The purpose of this paper is to provide a very modest overview of the discussion that took place from the early nineties onwards between the main players in the psychology of reasoning and which revolved around the knight–and–knave problems. Also, my aim will be to provide my personal reflections on the arguments presented.

2 Rips and the mental rules approach

Rips(Rips, 1989) was the first to suggest these brain teasers as an object of study for the psychology of reasoning. His motivation is that so far the field has focused on a very narrow body of reasoning tasks such as Aristotelian syllogisms and Wason's selection task. However, one could assert that his more or less hidden agenda was to address the question to what extent psychological theories should

Table 2: Rips' knight–knave–specific rules

1. if says(x, p), knight(x) then p
2. if says(x, p), knave(x) then $\neg p$
3. if $\neg$knave(x) then knight(x)
4. if $\neg$knight(x) then knave(x)

appeal to semantic concepts such as truth and falsity(Rips, 1986). Rips' hope is to demonstrate that his inference rule–based model sufficiently explains how a subject handles these problems without explicitly requiring a notion of truth or falsity on the level of the theory. This would in a broader context serve as an argument that in cognitive science such semantic concepts are superfluous.

2.1 The approach of mental deduction rules

On the basis of an informal observation of subjects solving these puzzles, Rips suggests the following model for their reasoning.

The model considers mental deduction rules a psychological primitive and they are used to calculate conclusions from a limited number of assumptions. The core propositional rules are adapted from a generic model(Rips, 1983)(Rips, 1989) that is capable of performing elementary inferences. This general propositional model is then supplemented with knight–knave–specific deduction rules, which represent the content of the instructions that one gives to the subject, e.g. that knights always speak the truth and knaves always lie. These specific rules are listed in table 2. [1]

Given these rules, the strategy for solving the puzzles can be represented as the following computer program.

1. It begins by assuming that the first speaker is a knight.

[1] In what follows, says(x, p) will represent that x utters "p," and $\neg p$ is the negation of p, i.e. NOT p.

2. From this assumption and using the generic and specific deduction rules, the program derives as many conclusions as possible. This phase stops when either of the following obtain:

    (a) The set of assumptions and conclusions is inconsistent. In this case the assumption that the first speaker is a knight is abandoned and replaced by the assumption that he is a knave.

    (b) No more rules apply, i.e. none of th deduction rules yields a conclusion that was not already found. In this case the program proceeds with the assumption that the second speaker is a knight.

3. The program continues like this until it has found all consistent sets of assumptions about the knight– and knave–status of the individuals.

The essential statement of the model is that the total number of applications of the rules needed to arrive at a conclusion is a measure for the complexity of the problem. This means that it predicts that problems that require a larger number of steps will take subjects longer and they will make more errors on them. In order to eliminate the influence of irrelevant factors, Rips forms pairs of problems that contain the same number of speakers and clauses, i.e. atomic propositions, but require different numbers of steps to solve them. By comparing subject performance within each pair only, Rips thus cancels out influence of processes other than actually solving the puzzle, such as reading the problem statement.

2.1.1 Suppositional reasoning

It is interesting to note that the program is *suppositional*, that is, it starts out by making an assumption, an *Ansatz*. Rips decided this was the most authentic procedure since he observed in an informal experiment where subjects were instructed to think aloud while solving the problems that they all started by making such an assumption and seeing where that reasoning led. However, when this procedure was reproduced in follow–up studies reference to the amount

Table 3: A model structure in the rule–based derivation.

A	B
knight	?

of suppositions is at least ambiguous. For instance (Elqayam, 2003) explains subjects make a supposition about the status of the first speaker and derive its consequences, and then make the contrary assumption, and "[t]hey thus proceed"(p.268). This means, they continue to make the supposition the second speaker is a knight, and then that he is a knave, and then the same for the third and so on. This is the only correct way to interpret Rips' explanation of the procedure(Rips, 1989)(p.91). Moreover it is not dificult to see that many problems would not even be solvable without making such multiple assumptions.

What is remarkable is that in each consistent set(Rips, 1989)(p.91), *ever* speaker thus is once the object of an assumption, not just the first speaker. This, in my point of view calls the question to what extent the subject in Rips' model is not actually constructing *models* and using the natural deduction rules to verify that they are consistent. Rips' derivation requires the subject to have at every point in time least some sort of a structure which can be visualised as a table and which keeps track of what speakers have what status.

For instance, in the problem mentioned in the introduction, the subject, after the assumption that A is a knight he needs a structure like in table 3 to represent that he has made one assumption about A and none yet about B.

That is to say, I argue that if Rips' mental rules are a psychological primitive, at least some sort of mental model is also.

2.1.2 Origin of the knight–knave rules

The model proceeds by repeated *rule application.* A first question could be why subjects would have precisely these rules in mind. Particularly puzzling is the absence of *backward inference rules* like

if says(x, p) and p then knight(x)

The reason they are not in the model is that Rips' bases his model on the subjects' thinking aloud when solving the problems and he never observed them using such *backward inference rules*(Rips, 1989)(p.89).

Moreover Rips' model does not need these rules, since the procedure of reasoning outlined above will simulate its behaviour. For instance, if $\text{says}(x, p)$ and p, then the program will begin by assuming $\text{knight}(x)$, which is precisely the conclusion we wanted. If it would consider the opposite, i.e. $\text{knave}(x)$, then it runs into a contradiction upon application of the knave–rule 2 of table 2 because it yields $\neg p$.

However, personally I found myself using this rule directly in a number of problems. The response that Rips could give here is that the number of inference steps his model yields is nevertheless a *measure* of how long it takes a subject to solve the problem even though the subject might occasionally optimise his strategy since such heuristics can probably be applied across the problem types equally.

Rather disconcerting is that Rips later in the article finds himself forced to add additional rules to his model to allow it to solve a wider variety of knight–knave–problems. Indeed one might long for a proof that the given model is capable of solving all problems, unless of course one wants to allow for the possibility that there are formally solvable problems that no human can solve.

2.1.3 Determinism of rule application

Another interesting note that the literature has not picked up is that Rips' procedure relies on the decidability of the deduction rules. This is crucial in the second step of the program, where as many conclusions as possible are drawn from the previously discovered or assumed facts. If this procedure would not eventually terminate in a state where the application of any rule no longer leads to a conclusion that was not already drawn, the subject would continue to derive new conclusions without ever passing to different assumptions. That is, the program would not be guaranteed to halt.

It seems plausible that the rules Rips proposes have this decidability property, even though he does not explicitly prove it. The reason is that he formulates uniquely elimination rules which only reduce the already finite complexity of phrases in the set of conclusions and therefore the procedure will eventually run out(Gentzen, 1969).

What remains remarkable, however, is that subjects would have precisely such a set of decidable rules in their minds. There are numerous examples in logic where different axiomatisations which are equivalent in terms of conclusions that are derivable from them, nevertheless differ in their decidability. One of these examples is Lambek's application of Gentzen's sequent calculus to phrase structures(Lambek, 1958) where he starts with the most intuitive set of rules, which is not decidable, and then needs several important revisions before arriving at a decidable set with the same proof–theoretic power.

If Rips is right that reasoning works with mental deduction rules, we are then also faced with the question why, of all possible axiomatisations, we have a decidable one in our heads.

2.1.4 Optimalisation in problem solving

Rips devotes a single sentence to report that his program also uses an optimalisation heuristic: "After each step, the program revises the ordering of its rules so that rules that have successfully applied will be tried first of the next round."(Rips, 1989)(p.91). But this raises the question how Rips' model accounts for this structure in subjects' performance if he wants to do away with any metalogical reasoning. For to be able to decide on the "success" of a rule it seems one needs a certain representation of one's one reasoning in the previous step.

However, if again we grant Rips that such optimalisations can be performed equally well across the experimental conditions their effect (i.e. the lowering of the number of steps required for solution) will be overall and therefore cancel out when comparing subject's performance in different problems.

2.2 Experimental confirmation of mental rules–account

In a first experiment Rips registers only the accuracy of the subject's responses. First of all he observed widespread incapacity to solve the problems: 10 out of 34 subjects gave up on the experiment within 15 minutes, and among the subjects that completed the test solved on average only about 20% of the problems was correctly answered. Second, among the pairs of problems matched for number of clauses and speakers more errors are made on the more difficult ones.

In a second experiment he measured the time it takes subjects to solve two–speaker three–clause problems. Again the main finding is that in spite of high error rates, subjects take longer to solve problems that take model more inference steps to solve.

3 Critique of mental rules and introduction of mental models

3.1 Evans

The first critique of Rips' study comes from (Evans, 1990).

First of all, Evans argues, the knight and knave problems are not meaningful in the real–world context, where one hardly encounters people who either always lie or speak the truth. This means in particular that it is doubtful to what extent subjects' performance in the experiment reflects reasoning as it is employed *in vivo*: "[w]e must recognise that almost all real–world cognition occurs in the presence of meaningful context"(Evans, 1990)(p.86–87).

Secondly, Evans feels the procedure Rips proposes for solving these riddles is unjustifiably deterministic in the sense that it eventually always finds the correct answer. The observations with towering high error rates contain only a fraction of correct responses, so in the best possible scenario Rips' model can be applied to this fraction only. The errors themselves can hardly be accounted for by a model that has no way to "generate" these errors itself.

Finally, though he himself places less emphasis on it, he nevertheless raises the interesting remark that Rips' model takes as a starting point the puzzle encoded in a logical format, e.g. says$(x, p \wedge q)$ rather than "X says that p and q." Although Evans, nor any other author that I know of, for that matter, does not develop this further, it does point into a seemingly trivial but essential nuance that might not be clear from the problem description: the scope of conjunction. For instance, the natural language version of our example could also have been transcribed as says$(x, p) \wedge$ says(x, q). This distinction is crucial for it turns the problem into a completely different one and I would even go as far as to argue that at least part of the errors can be attributed to this kind of misunderstanding. For instance, the problem in table 1 becomes paradoxical as soon as we would take A as uttering two assertions which therefore need to both be true or both be false. The first of which, $\neg A$ would render the puzzle paradoxical.

3.2 Johnson–Laird and Byrne

The next substantial criticism comes from a hardly surprising corner(Johnson-Laird & Byrne, 1990).

3.2.1 Criticism of the mental deduction rules

The main problem Johnson–Laird and Byrne identify in Rips' approach is again the deterministic nature of the procedure he describes. On the one hand it seems unrealistic to assume that subjects come to the task with a ready–made solution procedure as effective as Rips' model, and on the other hand it seems that even if they had, the procedure is so powerful that it would place unrealistic demands on their computational facilities.

Essentially, this problem lies in the need to follow up on disjunctive sets of models. For instance, if one speaker asserts $p \wedge q$ and the program arrived at the point of assuming that speaker a knight, it will then have to follow–up on each of the situations $\{p, \neg q\}$, $\{\neg p, q\}$ and $\{\neg p, \neg q\}$ and especially when it would need to compute additional disjunctive situations concerning other speakers in each case, the number of cases to be considered would grow exponentially, placing impossible demands on subject's memory.

However, as Rips argues in his defense(Rips, 1990), Johnson–Laird and Byrne seem to have misrepresented his position although they claim to use a simple "notational variant"(p.73). Though Rips does not explain this further, most probably he refers to the fact that his program will never consider such disjunctive cases separately but simply derive whatever conclusion is possible from the statement of the disjunction. In the example of $\neg(p \wedge q)$ the identities in his natural deduction model lead to conclude $\neg p \vee \neg q$ and then leave it at that.

A point Rips himself did not raise but which seems equally valid, is that even if subjects were in some way required to compute these disjunctive cases, then perhaps these "impossible demands" are precisely an explanation of the high error rates. To develop this further, one would need to determine which of the presented problems required following–up on disjunctive sets and see whether they yielded higher error rates and reaction times.

3.2.2 Model approach: developing strategies

Their point of view is that reasoning is based on *mental models*, or "internal model[s] of the state of affairs that the premises describe."(Johnson-Laird & Byrne, 1991)(p.35). Instead of deriving conclusions using rules without necessarily knowing what sorts of situations, or extensions, these conclusions refer to, Johnson–Laird and Byrne propose that reasoning is the construction and manipulation of mental representations that are more or less explicit. Broadly speaking, when a subject performs a modus ponens, he or she starts with a mental representation in which both premises are verified and then tries to create a model in which these remain true but the conclusion is false. Once he or she realises this cannot be done, the modus ponens is accepted as logically valid.

They feel it unreasonable to assume that subjects already have a ready–made procedure for solving the puzzles and rather develop ways, called *strategies*, to solve them as they observe themselves working.

They suggest to account for the data observed by Rips as the workings of four such mental strategies that are much like heuristics and which result from subjects' observing themselves performing the task: "With experience of the puzzles, they are likely to develop more systematic strategies."(Johnson-Laird & Byrne, 1990)(p.72). This is the kind of meta–cognitive capacity they feel Rips tried to evade in his model.

The proposed strategies are the following:

1. *Simple chain.* This strategy is to, like in Rips' model, follow all the consequences of assuming the first speaker to be a knight, with one difference: once one is required to look into disjunctive consequences, that is, precisely the case described before, which they identified as problematic in terms of cognitive complexity.

2. *Circular.* Once a speaker utters something that is self–referential, such as: "I am a knave and B is a knave," then the strategy is to follow up only on the immediate consequences, i.e. those that require a single reasoning step, since those often already rule out one of the cases. Thus the strategy is to not pursue the consequences of the consequences. In our example, assuming that the speaker is a knight can in such a way be rejected instantly.

3. *Hypothesise–and–match.* This strategy involves matching other speaker's utterances to previous conclusions. For instance, consider the following example:

   - A: A and B are knights.
   - B: A is a knave.

   The point is that as soon as one concludes that A cannot be a knight, and therefore must be a knave, then we can match this conclusion with B's assertion. Since they are the same thing, B must be a knight.

Interestingly, this deduction is precisely the inverse rule mentioned in section 2.1.2.

4. *Same–assertion–and–match.* In the case where two speakers make the same assertion, any other speaker who attributes a different status to them is necessarily lying.

   - A: C is a knave.
   - B: C is a knave.
   - C: A is a knight and B is a knave.

In a post–hoc analysis of Rips' very own data, they then proceed to show that problems which can be solved using these four strategies yield significantly more correct answers than those who cannot.

3.2.3 Reflection on mental models account

I would like to remark that the *simple chain* and *circular* strategies (and possibly the other two as well) only serve to eliminate parts of the "tree" of cases to be considered for a complete solution. As such, they are what in information science would be called a *heuristic*, they cut down parts of the search tree but they do not alter significantly the nature of the problem solution.

Secondly, there appears to be no unsystematic theory that unites them and therefore they can be said to be *ad hoc* in the sense that it would not be a surprise if one would come up with another strategy or maybe conclude that one of them is not applied after all. The problem about this is that the model has too many free parameters and therefore escapes scientific testing, rendering it pseudoscientific in a Popperian sense. Equivalently, it is very doubtful what the strategies really *explain* in subject's performance.

3.2.4 Rips' response to mental strategies

Rips responds in considerable detail(Rips, 1990) to the criticism outlined before.

Hardly surprisingly, one of his first remarks is that there is not much that the *mental models* contribute to Johnson–Laird and Byrne's approach to the problem. The strategies could have been formulated equally easy in a mental deduction rule framework, as in one based on mental models. Therefore, first of all, they do not particularly confirm the mental models account as such.

Furthermore, Rips remarks that in their post–hoc analysis of his data, of the four strategies, the *circular* was not included in the test for any puzzle in which it applied could have also been solved by the simple chain. Similarly, the *same–assertion–and–match* strategy because it applied in too little cases to allow statistical comparison. Then, if one matches the problems for number of clauses and speakers of the remaining two only *hypothesise–and–match* significantly explain the difference in scores. In other words, there is only experimental evidence for one of the four strategies.

However, in a study focussing more broadly on strategies in reasoning, Byrne and Handley(Byrne & Handley, 1997), mounting experiments of their own, find further evidence for reasoning strategies, taking away much of the power of this objection of Rips'.

Finally, in a remarkably lucid passage that, unfortunately to my knowledge has not been followed up in the literature, Rips also clarifies his position concerning the rejection of the use of meta–logical notions in psychological theories. Although Johnson–Laird and Byrne and Evans for that matter have taken him to reject using the notion of truth altogether, he argues only against appealing to expert theories of truth to explain subject's behaviour. Rips feels a theory can call on stage the subject's representation of truth, but it should not go further than that by using some independent theory of truth that logicians provide us with to explain how subjects behave: "Although cognitive psychologists can investigate people's beliefs about truth... it is quite another thing for cognitive psychologists to explain behaviour by appeal to the nature of truth itself."(Rips, 1990)(p.296–297)

Table 4: Example utterance from (Elqayam, 2003)(p.280)

- I am a knave or I am a knight

4 Elqayam and the norm in knight and knave puzzles

At this point in time the discussion between Rips, Evans and Johnson–Laird and Byrne falls quiet. More than a decade later, new light is shed on the discussion by Shira Elqayam(Elqayam, 2003).

4.1 Truth–value gaps

Among her most profound comments is that so far all studies into the knight and knave puzzles have assumed that there is a single "correct" answer. However, Elqayam observes the knight–knave puzzles presented to the subjects contained instances of the Liar and Truth–teller sentences.

These sentences are the starting point of Kripke's theory of truth, because they show that one cannot define a truth–predicate such that "p is true" is true if and only if "p."(Kripke, 1975) From there onwards several solutions are proposed, most of them introducing a third, "undefined" truth value in addition to "true" or "false," or, equivalently, a true predicate simply not applying to a certain number of sentences, like the liar.

Consider for instance the utterance in table 4.

Since the island is supposed to contain only knights and knaves, one can consider this phrase a tautology. On the other hand, Elqayam argues it can equally well be considered false since neither of the subphrases is necessarily true and some authors in philosophical logic classify such phrases as false. Finally, as long as the knight– or knave–status of the speaker has not been determined we can consider the two subphrases as undefined, i.e. the third truth value, and hence also their disjunction. Thus, depending on the norm we apply one can justifiedly consider a phrase either true, false, or neither.

This directly undermines the definition of a "correct" answer and thus might provide an essential clue as to the nature of the large number of "errors" observed. She argues that this absence of an objective norm could be remedied by allowing subjects when they classify speakers as either knight or knave the option that they "do not know."

4.2 Reflection on truth–gaps

I think Elqayam's observation of the implicit assumption of a logical norm in computing the response "correctness" is invaluable, and deserves as much credit as David Hume whom Immanuel Kant thanked for rousing him from his dogmatic slumber.

Before introducing my criticm, I would like to point out that Elqayam is precisely embarking in the analyses that Rips warned against(Rips, 1990) that is, she appeals to expert theories of truth to explain subject's behaviour. I would agree with Rips in the sense that it is important not to take the truth theories as restricting the possibilities of reasoning. Instead, Elqayam's analysis appears to me valid in that it hypothesises what notion of truth the subject uses when solving the task.

4.2.1 Truth–value gaps violate an instruction

The instruction given to the subjects stating that each inhabitant of the island is either knight or knave, is equivalent to the law of excluded middle. Therefore, the explicit instruction to the subject is to operate in bivalent logic. That is, as soon as a subject would consider that what a certain speaker has said is neither true nor false, he has violated the aspect of the puzzle that every inhabitant is either knight or knave and therefore in a way he or she is no longer solving the puzzle that was originally given. Thus, although Elqayam might offer a valid explanation of the "errors" observed in Rips' original experiment, it is not an example of the subject "justifiedly" using a different norm, which is what she argues.

At this point it is interesting to notice firstly the parallel with children solving the Tower of Hanoi problem, where often they are observed impose themselves additional constraints[2]. The difference here is that if subjects consider truth–value gaps in knight–knave puzzles they not elaborated the puzzle but they simply ignored one of its essential instructions: the law of excluded middle.

4.2.2 Paradoxality

In response to Elqayam's observation, it is good to remind ourselves that none of the speakers refers *only* to himself in their utterances. In those cases the problem could have also been formulated by eliminating that utterance, because at best they are redundant by not adding anything to the problem and at worst they cause the problem to be "underspecified," to use the distinction I introduced before. For instance, the puzzle of table 4 was never part of a problem presented to the subjects. This means that in particular, sentences such as the Liar and the Truth–teller, which are so far the only compelling reasons for us to abandon a bivalent truth assignment, do not occur.

It seems that Elqayam has confounded self–referentiality with paradoxality. This has been recently a greatly investigated topic in logic. Broadly speaking, Yablo showed an example of a paradox without self–reference(Yablo, 1993) and, conversely, Leitgeb argues in a recent paper that many sentences that refer to themselves can be considered not paradoxical(Leitgeb, 2005). The example Elqayam gives herself also falls in this latter category. In conjunction, these results show that paradoxicality and self–referentiality are far from being the same thing. Ironically, Elqayam seems to have applied a high–level version of the *circularity* strategy of Johnson–Laird and Byrne, suspecting paradox as soon as a speaker refers to himself.

[2] They take many more steps to solve the Tower of Hanoi problem since they do not allow themselves to move a stone two piles away.

4.2.3 Paradox by circularity and paradox by excluded–middle

Let us then turn to sentences which *could* and *did* occur in the problems presented to the subjects and look a bit closer at why they would contain a truth–value gap. For instance,

- A: I am a knight or B is a knave.

Elqayam would consider the first part ("I am a knight") as undetermined, in analogy to the liar sentence, which is undetermined. The reason is most likely that she feels there is a certain circularity analogous to the liar sentence, where, if we want to know whether it is true or false, we first need to know whether it itself is true or false, thus begging the question.

In modern logic and especially in recent days there has been considerable research into this idea, called *groundedness*(Leitgeb, 2005). The idea is that to determine the truth or falsity of certain sentences, like "It is true that snow is white," one needs to know the truth or falsity of "Snow is white" and that sentence itself does not depend on another sentence but on a state of affairs in the external world of which we are capable of verifying whether it is the case. Therefore, knowing this state of affairs we can fill in the truth value of "It is true that snow is white." This is why we tend to consider such sentences *grounded.*

However, in order to know the truth or falsity of a sentence like "This sentence is false." we would need to first know whether the sentence itself is true, for which we need to look at sentence itself again, and so on infinitely. This vicious circularity is why we call such sentences *ungrounded.*

And precisely here dawns a very important distinction between knight–knave puzzles and truth–predicate definition: in the latter case liar sentences are paradoxical because of *circularity* (for sentences become true or false by virtue of what they express being the case or not), in the former because of the knight–knave–island variant of the *excluded middle.*

If an inhabitant of the knight–knave island utters: "I am a knave," then in that will force us to abandon the assumption that all inhabitants are either knight or knave, if at all we want to evade contradiction. In that respect, even switching to trivalent logic would not help. But if an inhabitant utters: "What I now say is false," *that* will force us to abandon bivalent logic and with it also conclude that the one who utters it is neither knight nor knave.

Put in another way, we assume that each inhabitant is either a knight or a knave, even before he or she has said anything. The inhabitant does not *become* knight or knave by the uttering of a truth or a lie, he or she is assumed to have been so all along. It is only to *us*, listeners and explorers of the island, that their status turns from "indeterminate" *for us* to knight or to knave.

Thus, when Elqayam praises Rips for including a "do not know" option in his first experiment or other researchers(Schroyens *et al.* , 1999) for including even response patterns reminiscent of four–valued logic(Gupta & Belnap, 1993), that does not point subjects to three– or four–valued logic, but simply expresses their incapacity to tell.

4.2.4 Three–valued–logic and suppositional reasoning

The merit of Elqayam's proposal of the application of multivalued logic in the knight–knave puzzles has thus brought to light an essential difference between the knight–knave puzzles and the definition of a truth predicate in logic. The difference is that the island of knights and knaves, it seems, contains an additional layer where truth–value gaps can appear. For instance, if a person is neither knight or knave that would make for a "local" truth–value gap that violates the instruction that each person is either knight or knave. If a person utters a liar sentence, however, that makes for a "global" truth–value gap that violates bivalent logic.

Also, the distinction between paradox by circularity and paradox by excluded–middle helps to understand why the solution procedures proposed by all authors dealing with the knight–knave puzzles so far have always been *suppositional* (see section 2.1.1). That is, Rips already observed subjects need to start out by supposing a speaker to be either knight or knave and then deduce consequences. The point is that only making the supposition a speaker is a knight and then the supposition that the speaker is a knave will reveal the paradox by excluded–middle, whereas a paradox by circularity will yield a contradiction already by application of deduction rules. For instance, the liar sentence is shown to be paradoxical as soon as one substitutes it in the Tarski T–equivalence "p is true" iff p.

4.2.5 Bivalent logic

So where do we go then, if switching to trivalent logic does not help to explain the outcome of Rips' original experiment?

A clue might come from one of the most influential papers in contemporary logic(Leitgeb, 2005). Leitgeb proposes a definition of a predicate of truth which evades paradoxes while remaining in two–valued logic. This is achieved by applying the "naive" condition for truth predicates[3] only to grounded sentences. The unique feature of this approach to logical paradox that stays within bivalent logic and seems therefore the most appropriate candidate to handle knight–knave puzzles where the excluded–middle principle is an explicit constraint.

It would be interesting to use knight and knave puzzles to test whether subjects actually use such a conception of truth. Like Rips(Rips, 1990) emphasised, "[t]here is also no doubt that people have common–sense beliefs about truth and falsity, and it is of interest to document these notions and to compare them with expert theories." Perhaps, using the knight–knave paradigm, this question can actually be brought into the realm of experimental verification.

My very modest proposal is to eliminate the instruction that all inhabitants are either knight or knave. Thus, the only thing we instruct the subjects is that knights always tell the truth and knaves always lie.

Then consider the problems in table 5. The idea is that even though A utters an ungrounded sentence, B could be said to be a knight in virtue of knowing that snow is white or that a person cannot both be a knight and a knave.

[3] That is, the Tarski T–equivalence that a sentence "p is true" is true if and only if "p" is true

Table 5: Testing a subject's conception of truth

Problem I

- A: I am a knave.
- B: A is a knight or snow is white.
- Puzzle: What is B?

Problem II

- A: I am a knave.
- B: A is not both a knight and a knave.
- Puzzle: What is B?

If subjects turn out to be able to solve these two problems, one can conclude that the law of excluded middle is not inherent in their reasoning. For if it were, they would run aground upon hearing what A says. If subjects are not able to solve these problems that would corroborate Elqayam's point that subjects reason using a trivalent logic.

I realise there are many problems with this task and it is quite beyond the scope of this paper to deal with them. My aim was mainly to point out the possibility that knight–knave puzzles can help to understand how subjects conceive truth, and perhaps in the future inspire a more thoughtful analysis.

5 Conclusion

We have seen almost two decades of research into how subjects reason to solve knight–knave brain–teasers. Rips proposed a model based on mental deduction rules in which we, as psychologists of reasoning, do not need to appeal to meta–cognition. The results were criticised by Evans and Johnson–Laird and Byrne who propose their own interpretation based on mental models and meta–logical reasoning strategies.

Elqayam, almost a decade later, calls into doubt the nature of the norm that the previous authors have presupposed to be the only meaningful norm in knight–knave puzzles. In particular, she argues the problems call for or at least justify the use of three–valued logic. My commentary is that knight–knave puzzles come with the explicit requirement of the excluded middle, which forced us to conclude that subjects who use three–valued logic are no longer solving the puzzle as it was proposed. This is perhaps the most truthful explanation of Rips' observation of high error rates.

On the other hand, perhaps more importantly, these considerations can lead us to view these puzzles in a different way: rather as a tool that might lead to discover what subjects' conceptions about truth and falsity are.

References

Byrne, Ruth M. J., & Handley, Simon J. 1997. Reasoning strategies for suppositional deductions. *Cognition*, **62**(1), 1 – 49.

Elqayam, Shira. 2003. Norm, error, and the structure of rationality: The case study of the knight-knave paradigm. *Semiotica*, **2003**(147), 265–289.

Evans, Jonathan St. B. T. 1990. Reasoning with knights and knaves: A discussion of rips. *Cognition*, **36**(1), 85 – 90.

Gentzen, Gerhard. 1969. *The collected papers of gerhard gentzen*. M. e. szabo edn. Studies in logic and the foundations of mathematics. Amsterdam: North-Holland Pub. Co.

Gupta, Anil, & Belnap, Nuel. 1993. *The revision theory of truth*. Cambridge, Massachusetts, and London, England: MIT Press.

Johnson-Laird, P. N., & Byrne, R. M. 1990. Meta-logical problems: knights, knaves, and rips. *Cognition*, **36**(1), 69–84; discussion 85–90.

Johnson-Laird, P.N., & Byrne, R.M.J. 1991. *Deduction.* Essays in cognitive psychology. Hove (UK) Hillsdale,NJ (USA): L. Erlbaum Associates.

Kripke, Saul. 1975. Outline of a theory of truth. *The journal of philosophy*, **72**(19), 690–716.

Lambek, J. 1958. The mathematics of sentence structure. *American mathematical monthly*, **65**, 154–170.

Leitgeb, Hannes. 2005. What truth depends on. *Journal of philosophical logic*, **34**, 155–192.

Rips, L. J. 1989. The psychology of knights and knaves. *Cognition*, **31**(2), 85–116.

Rips, L. J. 1990. Paralogical reasoning: Evans, johnson-laird, and byrne on liar and truth-teller puzzles. *Cognition*, **36**(3), 291–314.

Rips, Lance J. 1983. Cognitive processes in propositional reasoning. *Psychological review*, **90**(Jan), 38–71.

Rips, Lance J. 1986. *The representation of knowledge and belief.* Tucson: University of Arizona Press. Myles Brand and Robert M. Harnish. Chap. Mental Muddles.

Schroyens, W., Schaeken, W., & D'Ydewalle, G. 1999. Error and bias in meta-propositional reasoning: A case of the mental model theory. *Thinking and reasoning*, **5**(38), 29–66.

Smullyan, R.M. 1987. *What is the name of this book? the riddle of dracula and other logical puzzles.* NJ: Prentice-Hall: Englewood Cliffs.

Tarski, Alfred. 1983. The concept of truth in formalized languages. *In:* Corcoran, J. (ed), *Logic, semantics and metamathematics.* Indianapolis: Hackett Publishing Company. The English translation of Tarski's 1936 Der Wahrheitsbegriff in den formalisierten Sprachen.

Yablo, Stephen. 1993. Paradox without self-reference. *Analysis*, **53**(4), 251–252.