On the Theory of Knights and Knaves

Words. And some formulas. • 22 Sep 2024

I. There is an island of knights who always say the truth and of knaves who never do. There is a country of day-knights who speak truthfully only at daytime and of night-knights whose speech at the day is a lie. There is an island of Democrats and Republicans. And they all came into being only to let use solve riddles.

II. But solving riddles is not the only thing one can do on this strange archipelago. One can also search for the general laws that govern the life on the islands. I will now use propositional logic for both: solving riddles in a systematical way and finding some of the general laws. Both activities are aspects of the same question.

III. There are many variants of these riddles. In their simplest form, we read a dialogue between persons who could be knights or knaves, without knowing who is which, and we must find out who has said the truth or at least which of their statements are true. Sometimes only partial solutions are possible.

Then there are the day-knights and night-knights that Raymond Smullyan describes in his book *"To Mock a Mockingbird"*. Day-knights speak the truth at daytime and lie when you meet them at night, while the night-knights do the opposite. As Smullyan tells us, these knights live in a subterranean place where you cannot see the sun and clocks are forbidden – only they know how late it is. He does not tell us where their country is located, but I am certain it is on an island.

A newer variant is Ben Orlin's "Island of Democrats and Republicans" where the inhabitants tell the truth to members of their own party and lie to their opponents. As usual in such riddles, everyone belongs to one of the two parties.

IV. In a simple knights-and-knaves riddle, a person could say, "Among me and my wife, one is a knight and one is a knave", and the wife add, "He is the knight". Since each of them could be a knave, we can take none of their statements at face value.

To translate this into logic, we introduce variables for the unknown facts. Since the unknown facts are here whether husband and wife are knights, we introduce the variables t_H and t_W to express that. This are logical variables and can take only the values **true** and **false**. The t in t_H stands for *truth-sayer*.

I use t_X and not something like k_X for the fact that person X is a knight because otherwise we might also think that k_X means that X is a knave. In fact, "X is a knave" is translated to $\neg t_X$ and literally means that X is not a knight: The symbol $\neg$ stands for negation.

V. Returning to the riddle, we can now translate the "dialog" between husband and wife into formulas as

$$\text{Husband: } t_H \leftrightarrow \neg t_W$$
$$\text{Wife: } t_H$$

The husband's statement uses the symbol $\leftrightarrow$ for logical equivalence. So the statement is true if and only if either the two terms at its side are both true or both false, and this happens only if one of the two persons is a knight and the other one a knave.

VI. There is some variation in the symbols that are used in propositional logic. I will use here $\wedge$ for "and", $\vee$ for "or" and $\rightarrow$ for the logical implication. For more information see the list of logical connectives in Wikipedia.

VII. On the island of knights and knaves, a statement is true exactly then when the person that makes it is a knight. This can be translated to the rule (from Raymond Smullyan's book "Forever Undecided", Chapter 7): *When person A makes a statement S, its meaning is $t_A \leftrightarrow S$.*

VIII. There are similar rules for the other islands:

- On the island of of day-knights and night-knights, a person says the truth if they are either a day-knight and it is day, or not a day-knight and it is not day. We can therefore define the variables d_A for "A is a day-knight" and D for "it is now day" and have the rule, *When person A makes a statement S, its meaning is $(d_A \leftrightarrow D) \leftrightarrow S$.*
- On the island of Democrats and Republicans, the variable d_A takes the meaning, "Person A is a Democrat", and the translation of a statement depends on the person to which it is directed: *When person A makes a statement S to person B, its meaning is $(d_A \leftrightarrow d_B) \leftrightarrow S$.*

IX. All these rules have the same structure. A statement S is translated to a formula $c \leftrightarrow S$, where c is the *local truth condition* for that island. On the island of knights and knaves, the truth condition is t_A, on the island of day-knights and night-knights it is $d_A \leftrightarrow D$, and so on.

This will be useful when we speak about observations that are true on all islands; I can describe them in the form that is true on the island of knights and knaves and you can then translate them for other islands by replacing the t_A term at the right with the local truth conditions of these islands.

X. Let us return to the husband-and-wife riddle.

We can now translate the statements of the husband and the wife, and their meanings are $t_H \leftrightarrow (t_H \leftrightarrow \neg t_W)$ and $t_W \leftrightarrow t_H$, respectively.

To find a solution, we need to know that *logical equivalence is associative*: $x \leftrightarrow (y \leftrightarrow z)$ is equivalent to $(x \leftrightarrow y) \leftrightarrow z$.

Because of this, the statement of the husband is equivalent to $(t_H \leftrightarrow t_H) \leftrightarrow \neg t_W$, which is equivalent to $\mathsf{true} \leftrightarrow \neg t_W$ and then to $\neg t_W$. So the wife is a knave and therefore, because of the wife's translated statement $t_W \leftrightarrow t_H$, the husband is a knave too and the riddle is solved.

XI. Before we proceed further, a remark about notation is in order.

There is a convention in mathematics that relational operators like $=$ or $\leq$ can be *chained*: We can for example write $x \leq y \leq z$, and it must be understood as $x \leq y \wedge y \leq z$.

Chained operators are good for long computations, and I want to use them for example instead of "is equivalent to" in the computation of the last section. But I will also need the unchained version of these operators.

Therefore I introduce the following convention: *The operators $\Leftarrow$, $\Rightarrow$ and $\Leftrightarrow$ are the chained versions of $\leftarrow$, $\rightarrow$ and $\leftrightarrow$. All other logical operators are not chained. The chained operators have also lower precedence, so that for example $x \rightarrow y \Rightarrow z$ means $(x \rightarrow y) \Rightarrow z$.*

XII. With this convention, the computation of the previous paragraph becomes

$$\begin{aligned} t_H \leftrightarrow (t_H \leftrightarrow \neg t_W) &\Leftrightarrow (t_H \leftrightarrow t_H) \leftrightarrow \neg t_W \\ &\Leftrightarrow \mathsf{true} \leftrightarrow \neg t_W \\ &\Leftrightarrow \neg t_W\,. \end{aligned}$$

XIII. The proof of the associativity of $\leftrightarrow$ is still missing. One could do that by comparing the values of $(x \leftrightarrow y) \leftrightarrow z$ and $x \leftrightarrow (y \leftrightarrow z)$ for all values of x, y and z, but that is boring. A more insightful way to do it is to note that both terms, $(x \leftrightarrow y) \leftrightarrow z$ and $x \leftrightarrow (y \leftrightarrow z)$, changes their truth value when one of the variables changes its value. So if the terms have same truth value for one setting of x, y and z, they will agree for all settings of these variables. We can therefore set all variables to true, see that $(\mathsf{true} \leftrightarrow \mathsf{true}) \leftrightarrow \mathsf{true}$ is the same as $\mathsf{true} \leftrightarrow (\mathsf{true} \leftrightarrow \mathsf{true})$, and conclude that the terms $(x \leftrightarrow y) \leftrightarrow z$ and $x \leftrightarrow (y \leftrightarrow z)$ also agree for all other settings of x, y and z.

Associativity of $\leftrightarrow$ also means that we can now remove the braces from terms that consist of a longer sequence of $\leftrightarrow$ symbols. I will however leave them in the formulas when it makes the meaning clearer.

XIV. The associativity of the $\leftrightarrow$ operator leads to an interesting "inversion" phenomenon. On the island of day-knights, when a person says, "I am a day-knight", it means that it is now day, and when they say, "It is now day", it means that they are a day-knight.

This is because if person A says, "I am a day-knight", it means

$$\begin{aligned}(d_A \leftrightarrow D) \leftrightarrow d_A &\Leftrightarrow d_A \leftrightarrow d_A \leftrightarrow D \\ &\Leftrightarrow D,\end{aligned}$$

while "It is now day" means

$$(d_A \leftrightarrow D) \leftrightarrow D \Leftrightarrow d_A\,.$$

In the same way we can show that on the island of Democrats and Republicans, when A says to B, "I am a Democrat", it means, "You are a Democrat", and vice versa. And similar phenomena occur when speaking about night-knights and Republicans.

XV. In the last calculations, I have used a theorem that is worth spelling out explicitly: *If a chain of logical equivalences contains the same term twice, then both occurrences can be removed and the value of the chain stays the same.*

This theorem, which I will call the "*pair rule*", follows from the associativity and commutativity of the $\leftrightarrow$ operator: An expression like

$$A \leftrightarrow x \leftrightarrow B \leftrightarrow x \leftrightarrow C$$

can be brought into the form

$$(x \leftrightarrow x) \leftrightarrow (A \leftrightarrow B \leftrightarrow C)$$

in which some of the braces are restored for clarity. The term $(x \leftrightarrow x)$ is always true, and $\mathsf{true} \leftrightarrow (\dots)$ is equivalent to $(\dots)$ for any term between the braces, therefore the whole formula is equivalent to

$$A \leftrightarrow B \leftrightarrow C\,.$$

The terms A, B and C can also be themselves longer chains of equivalences.

XVI. Longer chains of equivalences especially arise when the inhabitants of the islands ask each other questions. Person A may ask person B whether a statement Q is true, and B will answer with "Yes" or "No". If they answer with "Yes", they actually make the statement Q, and if they answer with "No", they state $\neg Q$. But B may be a liar, at least in this situation, and so we must transform the answer in the usual way:

On the island of knights and knaves, if A asks "Q?" and B answers "yes", this means $t_B \leftrightarrow Q$; an answer "No" means $t_B \leftrightarrow \neg Q$. On other islands, t_B must be replaced with the local truth condition.

XVII. Sometimes we need to make B's answer a variable that is true when B says "yes" and otherwise is false.

Then we can say: If A asks "Q?" and B answers with a, this means $t_B \leftrightarrow a \leftrightarrow Q$ on the island of knights and knaves. On other islands, t_B must be replaced with the local truth condition, which might make the formula a bit complex: With Democrats and Republicans, the interpretation of B's answer is $d_B \leftrightarrow d_A \leftrightarrow a \leftrightarrow Q$, which is already a longish chain.

XVIII. So far we have only listened to questions – but how do we ask them? Smullyan has a simple rule for asking the the right question even if you do not know whether the person you ask says the truth: *If you want to know whether a statement S is true, ask, "Are you one of the people who can say that S is true?"*

He calls this rule *Nelson Goodman's principle*, after the philosopher Nelson Goodman, who is best known for his "bleen-grue" paradox.

XIX. Let us now unpack Nelson Goodman's principle. It relies on the fact that the people on the islands always follow their rules. They may lie under certain conditions, but under these conditions they always lie. On the island of knights and knaves, an inhabitant A can say a true statement S if and only if they are a knight, that is, if $t_A \leftrightarrow S$ is true. In other words, the people who can say that S is true are exactly those for which the meaning of the statement when they say it, in our case $t_A \leftrightarrow S$, is true.

Now it is clear what happens: In order to know whether S is true, we ask "$t_A \leftrightarrow S$" from person A, and A replies with the statement a. This answer then has the meaning

$$t_A \leftrightarrow (a \leftrightarrow (t_A \leftrightarrow S)),$$

which by associativity and the pair rule is equivalent to

$$a \leftrightarrow S\,.$$

We have our answer!

XX. Nelson Goodman's principle is especially useful when we can simplify Goodman's convoluted phrase. As an example, assume that on the island of Democrats and Republicans, A wants to ask B whether B is in the same party as a third person, C. Goodman's question is then, "Are you one of the people who can say to me that C is in the same party as you?", or

$$(d_A \leftrightarrow d_B) \leftrightarrow (d_B \leftrightarrow d_C)\,.$$

Simplified with the pair rule, this becomes

$$d_A \leftrightarrow d_C$$

and A only needs to ask, "Is C a member of my party?" to get the answer they want.

XXI. There are more logical connectives than just equivalence, and we must be able to handle them too. Unfortunately, the equivalence operator does not cooperate well with the other logical operator: There is for example no analogue to the distributive law for formulas with $\leftrightarrow$ and $\wedge$.

There is however a general transformation that is often useful. With it, we can (almost) eliminate a specific variable from a given formula. To demonstrate it, let $P(x)$ be a logical formula that depends on x and possibly other variables. Then $P(x)$ is equivalent to the expression

$$(\neg x \to P(\mathsf{false})) \wedge (x \to P(\mathsf{true}))\,.$$

This is simply a proof by cases: For $P(x)$ to be true, $P(\mathsf{false})$ must be true if x is false and $P(\mathsf{true})$ must be true if x is true.

XXII. This formula can be brought into a shorter form. We must only notice that its left term is equivalent to $\neg P(\mathsf{false}) \to x$: Then we can use the chained form of the $\to$ operator and the formula becomes

$$\neg P(\mathsf{false}) \Rightarrow x \Rightarrow P(\mathsf{true})\,.$$

I call this the *case expansion* of $P(x)$. It is especially useful when the variable x occurs multiple times in $P(x)$ because then $P(\mathsf{false})$ and $P(\mathsf{true})$ could become much smaller expressions.

XXIII. To actually apply this rule, we will also need something that is usually not given when a logical system is described, namely a list of simplification rules for terms where one operand of a logical connective is the constant true or false.

Simplification rules		
$\mathsf{true} \wedge x \Leftrightarrow x$	$\mathsf{true} \to x \Leftrightarrow x$	$\mathsf{true} \leftrightarrow x \Leftrightarrow x$
$\mathsf{false} \wedge x \Leftrightarrow \mathsf{false}$	$\mathsf{false} \to x \Leftrightarrow \mathsf{true}$	$\mathsf{false} \leftrightarrow x \Leftrightarrow \neg x$
$\mathsf{true} \vee x \Leftrightarrow \mathsf{true}$	$x \to \mathsf{true} \Leftrightarrow \mathsf{true}$	
$\mathsf{false} \vee x \Leftrightarrow x$	$x \to \mathsf{false} \Leftrightarrow \neg x$	

XXIV. With the case expansion we can especially get an interesting formula for a logical equivalence in which one side consists of a single variable: $P(x)$ is then an expression of the form $x \leftrightarrow Q(x)$, with $Q(x)$ an arbitrary formula. The case expansion of this $P(x)$ is

$$\neg(\mathsf{false} \leftrightarrow Q(\mathsf{false})) \Rightarrow x \Rightarrow (\mathsf{true} \leftrightarrow Q(\mathsf{true}))$$

and with help of the table it can be simplified to

$$Q(\mathsf{false}) \Rightarrow x \Rightarrow Q(\mathsf{true})\,.$$

XXV. As an illustration for this method, let us use it for the solution of another husband-and-wife riddle by Smullyan (“Forever Undecided”, Chapter 3 No. 3). In it, a man says, “If I am a knight, my wife is one too”. What can be made out of such a statement?

Well, its meaning is $t_H \leftrightarrow (t_H \to t_W)$ since it was made by the husband. The term $Q(x)$ is then $x \to t_W$, and so we have the expansion

$$\mathsf{false} \to t_W \Rightarrow t_H \Rightarrow \mathsf{true} \to t_W\,.$$

Now the term at the left is always true, while that at the right is equivalent to t_W, so that the expansion is equivalent to

$$\mathsf{true} \Rightarrow t_H \Rightarrow t_W\,.$$

And evaluating it from left to right, we see that husband and wife both must be knights.

(May be continued.)