Logic and **L**ogical **P**hilosophy
Volume 4 (1996), 97–103

Adam Kolany

A GENERAL METHOD OF SOLVING SMULLYAN'S PUZZLES

Abstract. In this paper we present a general method of solving Smullyan's puzzles. We do this by showing how a puzzle is translated into Classical Propositional Calculus.

CONTENTS

Received March 6, 1996. Revised November 19, 1996

I. Introduction

Raymond M. Smullyan, apart from his purely logical work, produced a certain number of booklets on logical puzzles. In almost all of them, one can distinguish two main groups of puzzles. The first group is about the notion of truth, beliefs, lying, etc. We find here so called "Knight and Knave"-puzzles — amazing stories with knights, knaves, vampires, ghosts, zombies, and many other creatures, in which we are expected to behave appropriately in order to achieve desired goals. We must guess who is who there, and we must carefully answer questions in order to avoid Dracula's rage. The other group of Smullyan's puzzles deals with self-reference, recurrence and Gödel's Incompleteness Theorem. Here we search for properties of the more and more complicated machines of McCulloch and attempt to discover the secrets of Monte Carlo Castle.

This paper concerns the first group of puzzles. Although only two of Smullyan's booklets are cited, the methods presented can be applied to the others, as well.

We will deal with three types of puzzles which we will briefly call:

- Guess, who I am!
- I've forgotten what you said!
- What am I to say? (What shall I ask for?)

There is also a type of puzzle which Smullyan calls "metapuzzles", and "I've forgotten what you said!"-puzzles are, in fact, examples of these. Although the methods presented here would help solve metapuzzles, we will not attempt to do so.

II. Puzzles

Below are examples of each of considered types of puzzles.

1. Knights and Knaves ([1], p. 20): *There is a wide variety of puzzles about an island in which certain inhabitants called "knights" always tell the truth, and others called "knaves" always lie. It is assumed that every inhabitant of the island is either a knight or a knave.*

Example ([1], p. 20/27; you meet two inhabitants A and B of the Island of Knights and Knaves):

Suppose A *says, "I am a knave, but* B *isn't". What are* A *and* B*?*

2. Still on the same Island.
Example ([1], p. 22, 36):

[...] *I came across two of the inhabitants resting under a tree. I asked one of them, "Is either of you a knight?" He responded, and I knew the answer to my question.*

What is the person to whom I addressed the question — is he a knight or a knave; And what is the other one? I can assure you, I have given you enough information to solve this problem.

3. On the Island of Zombies ([1], p. 149): *On a certain island near Haiti, half of the inhabitants have been bewitched by voodoo magic and turned into zombies. The zombies of this island do not behave according to the conventional concept: they are not silent or deathlike — they move about and talk in as lively a fashion as do the humans. It's just that the zombies of this island always lie and the humans of this island always tell the truth.*
So far, this sounds like another knight-knave situation in a different dress, doesn't it? But it isn't! The situation is enormously complicated by the fact that although all the natives understand English perfectly, an ancient taboo of the island forbids them ever to use non-native words in their speech. Hence whenever you ask them a yes-no question, theyreply Bal *or* Da *— one of which means* Yes *and another* No*. The trouble is that we do not know which of* Bal *or* Da *means* Yes *and which means* No*.*

Example ([1], p. 150/160):

Suppose you are not interested in what Bal *means, but only in whether the speaker is a zombie. How can you find this out in only one question? (Again he will answer* Bal *or* Da*.)*

III. The Method

The general method of solving the puzzles we present in this paper is the following. Given a puzzle, one has to design a consequence operator Cn, which is a conservative extension of the consequence operator of Classical Propositional Calculus, $\mathsf{Cn}_{\mathsf{CPC}}$, in its detachmental version, for the appropri-

ate language. Then we inspect the set of theses of Cn in order to determine whether it contains certain formulas (mostly propositional constants). In the case of "I've forgotten what you said", we have to choose one of $X_1, \ldots, X_n$, so that $\mathsf{Cn}(X_j)$ has desired properties.

In most cases the extension Cn will be an extension by definitions, augmented by axioms describing the very situation we are considering at the moment (for instance, a translation of the dialog, we participated in, or additional facts we know about individuals taking part in the situation). The language we will work with is the language we will denote $\mathsf{L}_{\mathsf{Smull}}$ and which is an extension of the usual language of the propositional calculus constructed by adding to the latter an infinite number of propositional constants T_X, F_X, S_X, read "X is a truth-teller", "X is a liar", „X is sane", respectively, $X = \mathsf{A}, \mathsf{B}, \ldots$, an infinite number of modal connectives $X \lhd \ldots$ and $X(\ldots)$, which are read "X says that…" and "X is convinced that…", resp., $X = \mathsf{A}, \mathsf{B}, \ldots$. Moreover, we assume that $\mathsf{L}_{\mathsf{Smull}}$ contains a modal connective $\mathcal{B}(\ldots)$ meaning "The right answer to … is $\mathtt{Bal}$" and an infinite number of modal connectives $X \lhd \mathcal{B}(\ldots)$ meaning "X answers $\mathtt{Bal}$ to the question on …" (see [1], p. 149). Instead of $\mathcal{B}(\mathbf{truth})$ we shall simply write $\mathcal{B}$. Hence $\mathcal{B}$ means that "$\mathtt{Bal}$ means *Yes*". We admit also that $\mathsf{L}_{\mathsf{Smull}}$ contains other symbols if it is needed.

IV. Applying the method

In this section we define some sets of formulas of the language $\mathsf{L}_{\mathsf{Smull}}$ which will serve as axioms for extending $\mathsf{Cn}_{\mathsf{CPC}}$ to desired consequence operators.

Let $\mathsf{Cn}_\lambda(X) = \mathsf{Cn}_{\mathsf{CPC}}(X \cup X_\lambda)$, $\lambda \in \{\mathsf{KK}, \mathsf{vampire}, \mathsf{zombie}, \mathsf{normals}\}$, where

$$
\begin{aligned}
X_{\mathsf{KK}} = {} & \{T_X \equiv \neg F_X \,:\, X = \mathsf{A}, \mathsf{B}, \mathsf{C}, \ldots\} \cup \\
& \cup \{X \lhd \alpha \equiv (T_X \equiv \alpha) \,:\, X = \mathsf{A}, \mathsf{B}, \ldots,\ \alpha \in \mathsf{Frm}(\mathsf{L}_{\mathsf{Smull}})\},
\end{aligned}
$$

$$
\begin{aligned}
X_{\mathsf{vampire}} = {} & \{T_X \equiv \neg F_X : X = \mathsf{A}, \mathsf{B}, \mathsf{C}, \ldots\} \cup \\
& \cup \{X \lhd \alpha \equiv (T_X \equiv X(\alpha)) \,:\, X = \mathsf{A}, \mathsf{B}, \ldots,\ \alpha \in \mathsf{Frm}(\mathsf{L}_{\mathsf{Smull}})\} \cup \\
& \cup \{X(\alpha) \equiv (S_X \equiv \alpha) \,:\, \alpha \in \mathsf{Frm}(\mathsf{L}_{\mathsf{Smull}}),\ \mathsf{X} = \mathsf{A}, \mathsf{B}, \ldots\},
\end{aligned}
$$

$$
\begin{aligned}
X_{\mathsf{zombie}} = {} & \{T_X \equiv \neg F_X \,:\, X = \mathsf{A}, \mathsf{B}, \mathsf{C}, \ldots\} \cup \{X \lhd \mathcal{B}(\alpha) \equiv \\
& (T_A \equiv X(\mathcal{B}(\alpha))) \,:\, X = \mathsf{A}, \mathsf{B}, \ldots,\ \alpha \in \mathsf{Frm}(\mathsf{L}_{\mathsf{Smull}})\} \cup \\
& \cup \{X(\alpha) \equiv (T_A \equiv \alpha) \,:\, \alpha \in \mathsf{Frm}(\mathsf{L}_{\mathsf{Smull}}),\ X = \mathsf{A}, \mathsf{B}, \ldots\} \cup \\
& \cup \{\mathcal{B}(\alpha) \equiv (\mathcal{B} \equiv \alpha) \,:\, \alpha \in \mathsf{Frm}(\mathsf{L}_{\mathsf{Smull}})\},
\end{aligned}
$$

$$
\begin{aligned}
X_{\mathsf{normals}} = {} & \{\neg T_X \vee \neg F_X \,:\, X = \mathsf{A}, \mathsf{B}, \mathsf{C}, \ldots\} \cup \\
& \cup \{X \lhd \alpha \equiv (T_X \to \alpha) \wedge (F_X \to \neg\alpha) \,:\, X = \mathsf{A}, \mathsf{B}, \ldots\}.
\end{aligned}
$$

Now solving a puzzle reduces to considering what belongs to $\mathsf{Cn}_\lambda(X_{\mathsf{puzzle}})$, for appropriate $\lambda \in \{\mathsf{KK}, \mathsf{vampire}, \mathsf{normals}, \mathsf{zombie}\}$ and $X_{\mathsf{puzzle}} \subseteq \mathsf{Frm}(\mathsf{L}_{\mathsf{Smull}})$.

(1) Given an inhabitant A of the Island of Knights and Knaves, let us assume that $\mathsf{A} \lhd \alpha$, then $T_{\mathsf{A}} \to \alpha$ and $F_{\mathsf{A}} \to \neg\alpha$. Since F_{A} is equivalent to $\neg T_{\mathsf{A}}$, here, we obtain that $\mathsf{A} \lhd \alpha \to (T_{\mathsf{A}} \equiv \alpha)$. On the other hand, let us assume that $T_{\mathsf{A}} \equiv \alpha$. If A is a knight, α must be true, by our assumption, hence $\mathsf{A} \lhd \alpha$. If A is a knave, in turn, α is false and, again, it must be that $\mathsf{A} \lhd \alpha$. This proves that in order to solve the puzzle, we have to decide, which of T_{A}, F_{A} and T_{B}, F_{B} is in $\mathsf{Cn}_{\mathsf{KK}}(X_{\mathsf{puzzle}})$, where $X_{\mathsf{puzzle}} = \{\mathsf{A} \lhd (F_{\mathsf{A}} \wedge \neg F_{\mathsf{B}})\}$. Since we have $T_{\mathsf{A}} \equiv (\neg T_{\mathsf{A}} \wedge T_{\mathsf{B}}) \in \mathsf{Cn}_{\mathsf{KK}}(X_{\mathsf{puzzle}})$, and, since the latter formula is equivalent to $\neg T_{\mathsf{A}} \wedge \neg T_{\mathsf{B}}$, we conclude that both chaps are knaves.

(2) Here we must decide which of $\mathsf{A} \lhd \alpha_j$, $j = 1, 2$, took place, where $\alpha_1 = (T_{\mathsf{A}} \vee T_{\mathsf{B}})$ and $\alpha_2 = \neg T_{\mathsf{A}} \wedge \neg T_{\mathsf{B}}$. Since the hero of the Smullyan's story was able to deduce the right answer, we choose that of α_i, $i = 1, 2$, which satisfies $\alpha_j \in \mathsf{Cn}_{\mathsf{KK}}(\{\mathsf{A} \lhd \alpha_i\})$, for some $j = 1, 2$, (in fact, for exactly one j). Since, $T_{\mathsf{A}} \vee (\neg T_{\mathsf{A}} \wedge \neg T_{\mathsf{B}}) \in \mathsf{Cn}_{\mathsf{KK}}(\{\mathsf{A} \lhd \alpha_1\})$, and $\neg T_{\mathsf{A}} \wedge (T_{\mathsf{A}} \vee T_{\mathsf{B}}) \in \mathsf{Cn}_{\mathsf{KK}}(\{\mathsf{A} \lhd \alpha_2\})$, we conclude that α_2 has been answered and that A is a knave and B is a knight.

(3) Reasoning in a quite similar way as in **(1)**, we conclude that $\mathsf{A} \lhd \mathcal{B}(\alpha)$ is equivalent to $T_{\mathsf{A}} \equiv \mathcal{B}(\alpha)$. Now, in order to encrypt what $\mathcal{B}(\alpha)$ should mean, we notice that $\mathcal{B}(\alpha)$ is the same as saying that either $\mathtt{Bal}$ means Yes and α is true, or $\mathtt{Bal}$ means No and α is false. This however means that $\mathcal{B}(\alpha)$ is equivalent to $\mathcal{B} \equiv \alpha$. Now, in order to solve the puzzle, we have to find a sentence α with the property $(\mathsf{A} \lhd \mathcal{B}(\alpha)) \equiv F_{\mathsf{A}} \in \mathsf{Cn}_{\mathsf{zombie}}(\emptyset)$. Since $((\mathsf{A} \lhd \mathcal{B}(\alpha)) \equiv F_{\mathsf{A}}) \equiv (T_{\mathsf{A}} \equiv \alpha \equiv \mathcal{B} \equiv \neg T_{\mathsf{A}}) \in \mathsf{Cn}_{\mathsf{zombie}}(\emptyset)$, we obtain that $\alpha = \neg\mathcal{B}$ would fit our expectations. I.e. we could ask him whether $\mathtt{Bal}$ means Yes. If he answers $\mathtt{Bal}$, he is a zombie.

In order to illustrate the use of $\mathsf{Cn}_{\mathsf{vampire}}$ and $\mathsf{Cn}_{\mathsf{normals}}$, let us solve other two puzzles of the types "Guess, who I am!" and "What am I to say?", respectively.

- **In Transylvania:** ([2], p. 45) [...] *Transylvania is inhabited by both vampires and humans; the vampires always lie and the humans always tell the truth. However, half of the inhabitants, both human and vampire, are insane and totally deluded in their beliefs* [...] *— all true propositions they believe false and all false propositions they believe true. The other half inhabitants are completely sane and totally accurate in their*

judgements [...] — *all true statements they know to be true and all false statements they know to be false.*
[...] *Sane humans and insane vampires both make only true statements; insane humans and sane vampires make only false statements* [...]

EXAMPLE ([2], p. 47/2; Here are the statements of two brothers, both named "Bela". One of them is a human, the other one is a vampire):

[...]
Bela *the Elder: I am human.*
Bela *the Younger: I am human.*
Bela *the Elder: My brother is sane.*
Which of them is a vampire?

SOLUTION: Taking into account the presence of insane individuals, we must be more careful. Now, X says that α is equivalent to $T_X \equiv X(\alpha)$, where $X(\alpha) \equiv (S_X \equiv \alpha)$. While the latter equivalence seem quite obvious, the first needs some explanation. This, however, follows from the simple observation that being a truthteller means making statements one is convinced of — not which are really true.

Accordingly, our task is to conclude which elements of the set $\{\S_X : X = \mathsf{A}, \mathsf{B}, \S = S, T\} \cup \{\neg\S_X : X = \mathsf{A}, \mathsf{B}, \S = S, T\}$ are in $\mathsf{Cn}_{\mathsf{vampire}}(X_{\mathsf{puzzle}})$, where $X_{\mathsf{puzzle}} = \{\mathsf{A} \lhd T_{\mathsf{A}}, \mathsf{B} \lhd T_{\mathsf{B}}, \mathsf{A} \lhd S_{\mathsf{B}},\ T_{\mathsf{A}} \equiv \neg T_{\mathsf{B}}\}$. We have $T_{\mathsf{A}} \equiv \mathsf{A}(T_{\mathsf{A}})$, $T_{\mathsf{B}} \equiv \mathsf{B}(T_{\mathsf{B}})$, $T_{\mathsf{A}} \equiv \mathsf{A}(S_{\mathsf{A}})$, $T_{\mathsf{A}} \equiv \neg T_{\mathsf{B}} \in \mathsf{Cn}_{\mathsf{vampire}}(X_{\mathsf{puzzle}})$ and hence $T_{\mathsf{A}} \equiv S_{\mathsf{A}} \equiv T_{\mathsf{A}}$, $T_{\mathsf{B}} \equiv S_{\mathsf{B}} \equiv T_{\mathsf{B}}$, $T_{\mathsf{A}} \equiv S_{\mathsf{A}} \equiv S_{\mathsf{A}} \in \mathsf{Cn}_{\mathsf{vampire}}(X_{\mathsf{puzzle}})$, which implies that $S_{\mathsf{A}}, S_{\mathsf{B}}, T_{\mathsf{A}}, \neg T_{\mathsf{B}} \in \mathsf{Cn}_{\mathsf{vampire}}(X_{\mathsf{puzzle}})$. Thus we conclude that the both brothers were sane. The first of them is a human, the other one is a vampire.

- **Knights, Knaves and Normals** ([1], p. 23, 87): *An equally fascinating type of problem deals with three types of people: knights, who always tell the truth; knaves, who always lie; and normals, who sometimes lie and sometimes tell the truth.*
  [...] *a crime has been committed on the island, and for some strange reason it is suspected that you are the criminal. You are brought to court and tried. You are allowed to make only one statement on your behalf. Your purpose is to convince the jury that you are innocent.*

EXAMPLE ([1, p. 87/99):

[...] *Suppose it is known that the criminal is a knave. Suppose also that you are a knave (although the court doesn't know this) but that you are nevertheless innocent of this crime. You are allowed to make only one*

statement. Your purpose is not to convince the jury that you are not a knave, but only that you are innocent of this crime. What would you say?

Solution: Let G_{A} mean that A is guilty. Whatever consequence operator Cn is appropriate to the situation of the puzzle, it must satisfy $\mathsf{A} \lhd \alpha \to (T_{\mathsf{A}} \to \alpha) \wedge (F_{\mathsf{A}} \to \neg\alpha) \in \mathsf{Cn}(\emptyset)$. The aim of the puzzle is to find a sentence α with $\neg G_{\mathsf{A}} \in \mathsf{Cn}(\{\mathsf{A} \lhd \alpha, G_{\mathsf{A}} \to F_{\mathsf{A}}\})$. Because

$$\begin{aligned}\mathsf{Cn}(\{\mathsf{A} \lhd \alpha, G_{\mathsf{A}} \to F_{\mathsf{A}}\}) \supset \mathsf{Cn}(\{T_{\mathsf{A}} \to \alpha, F_{\mathsf{A}} \to \neg\alpha, G_{\mathsf{A}} \to F_{\mathsf{A}}\}) = \\ = \mathsf{Cn}(\{\neg T_{\mathsf{A}} \vee \alpha, \neg F_{\mathsf{A}} \vee \neg\alpha, \neg G_{\mathsf{A}} \vee F_{\mathsf{A}}\}) \ni \neg\alpha \vee \neg G_{\mathsf{A}}\end{aligned}$$

we see that $\alpha = G_{\mathsf{A}}$ fits our expectations and in order to be judged innocent, one must say, he is guilty!

References

[1] Smullyan R. M., *What is the Name of This Book?*, Prentice-Hall, inc, Englewood Cliffs, New Jersey, 1978.

[2] Smullyan R. M., *The Lady or the Tiger?*, Oxford University Press, 1991.

Adam Kolany
Institute of Mathematics
University of Silesia
Bankowa 14, 40-007 Katowice, Poland
e-mail: kolany@usctoux1.cto.us.edu.pl