

A simple solution to the hardest logic puzzle ever

BRIAN RABERN & LANDON RABERN

We present the simplest solution ever to ‘the hardest logic puzzle ever’. We then modify the puzzle to make it even harder and give a simple solution to the modified puzzle. The final sections investigate exploding god-heads and a two-question solution to the original puzzle.

1. The simplest solution to the ‘hard’ puzzle

The puzzle. Three gods A, B, and C are called, in some order, ‘True’, ‘False’, and ‘Random’. True always speaks truly, False always speaks falsely, but whether Random speaks truly or falsely is a completely random matter. Your task is to determine the identities of A, B, and C by asking three yes-no questions; each question must be put to exactly one god. The gods understand English, but will answer all questions in their own language, in which the words for ‘yes’ and ‘no’ are ‘da’ and ‘ja’, in some order. You do not know which word means which.¹

Boolos 1996 provides the following guidelines:

- (B1) It could be that some god gets asked more than one question (and hence that some god is not asked any question at all).
- (B2) What the second question is, and to which god it is put, may depend on the answer to the first question. (And of course similarly for the third question.)
- (B3) Whether Random speaks truly or not should be thought of as depending on the flip of a coin hidden in his brain: if the coin comes down heads, he speaks truly; if tails, falsely.

¹ Boolos 1996: 62. The so-called ‘hardest logic puzzle ever’ is coined as such by George Boolos. Boolos credits the logician Raymond Smullyan as the originator of the puzzle and the computer scientist John McCarthy with adding the difficulty of not knowing what ‘da’ and ‘ja’ mean. Related puzzles can, however, be found throughout Smullyan’s writings, e.g. in Smullyan 1978: 149–56, he describes a Haitian island where half the inhabitants are zombies (who always lie) and half are humans (who always tell the truth) and explains that ‘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, they reply “Bal” or “Da” – one of which means *yes* and the other *no*. The trouble is that we do not know which of “Bal” or “Da” means *yes* and which means *no*’. In fact, Smullyan solves his own puzzle 162 by using an instance of the embedded question lemma*, so he had already introduced the essential ingredient needed for a simple solution to the hardest logic puzzle ever. (For another related puzzle see Smullyan 1997: 114.)

(B4) Random will answer ‘da’ or ‘ja’ when asked any yes-no question.

Before continuing with this article the reader may wish to pause and attempt a solution.

To solve this puzzle we introduce a function from questions to questions and prove a lemma, which trivializes the puzzle.² Let E be the function that takes a question q to the question ‘If I asked you ‘ q ’ in your current mental state would you say ‘ja’?’¹

Embedded question lemma. When any god g is asked $E(q)$, a response of ‘ja’ indicates that the correct answer to q is affirmative and a response of ‘da’ indicates that the correct answer to q is negative.

Proof. If g is either True or False, the result follows, since both a double positive and a double negative make a positive. Hence we may assume that g is Random. According to (B3), when we pose $E(q)$ to Random the hidden coin in his brain is flipped. If the coin comes down heads, Random’s mental state is that of a truth-teller; if tails, Random’s mental state is that of a liar. In either case, the result again follows, since both a double positive and a double negative make a positive.

With the embedded question lemma in our arsenal the ‘hard’ puzzle is no more difficult than the following trivial puzzle.

The trivial puzzle. Three gods A, B, and C are called, in some order, ‘Zephyr’, ‘Eurus’, and ‘Aeolus’. The gods always speak truly. Your task is to determine the identities of A, B, and C by asking three yes-no questions; each question must be put to exactly one god. The gods understand English and will answer in English.

2. Random troubles and the Random modification

One virtue of logical argumentation is that there is not a gap between what one means and what one says or what one says and what one means. The puzzle was presented precisely as above and we have provided the simplest solution to the puzzle as presented. Nevertheless, the spirit of the original Smullyan-puzzle has certainly been lost. Most commentators on the puzzle have assumed that Random answers randomly and that therefore nothing can be gleaned from his answers; but that is not how Random works.

Notice what happens if we ask Random: ‘Are you going to answer this question with a lie?’ If his brain-coin lands heads, he must answer negatively (since it is not true that he will lie) and if his brain-coin lands tails

² Note that throughout this article we are limiting our focus to polar questions (i.e. yes-no questions), e.g. the functions that we introduce, E and E^* , only take yes-no questions as argument.

he also must answer negatively (since while it is true that his answer is a lie, he is lying so he will not answer affirmatively). In what sense is this random? He always has to answer this question negatively!³

This predictability that has been built into Random (apparently unintentionally) is precisely what we have exploited to trivialize the puzzle. To make Random truly random, we replace (B3) with the following (and make the necessary modification to the original puzzle):

(B3*) Whether Random answers 'ja' or 'da' should be thought of as depending on the flip of a coin hidden in his brain: if the coin comes down heads, he answers 'ja'; if tails, he answers 'da'.

3. The simplest solution to the modified puzzle

The modified puzzle. Three gods A, B, and C are called, in some order, 'True', 'False', and 'Random'. True always speaks truly, False always speaks falsely, but *whether Random answers 'ja' or 'da' is a completely random matter*. Your task is to determine the identities of A, B, and C by asking three yes-no questions; each question must be put to exactly one god. The gods understand English, but will answer all questions in their own language, in which the words for 'yes' and 'no' are 'da' and 'ja', in some order. You do not know which word means which.

To solve the modified puzzle we introduce another function from questions to questions and prove two lemmas. Let E^* be the function that takes a question q to the question 'If I asked you ' q ' would you say 'ja'?'.⁴

*Embedded question lemma**. When either True or False are asked $E^*(q)$, a response of 'ja' indicates that the correct answer to q is affirmative and a response of 'da' indicates that the correct answer to q is negative.

Proof. Both a double positive and a double negative make a positive.

Identification lemma. If it has been determined that a particular god is not Random and two questions remain, then every god's identity can be determined.

³ Young notes in the appendix 'Some random observations' to his unpublished manuscript that if we ask Random 'Is it true that (you are lying iff Dushanbe is in Kirghizia)?' Random will answer negatively when and only when it is true that Dushanbe is in Kirghizia and will answer affirmatively when and only when Dushanbe is not in Kirghizia. But he does not note how this trivializes the puzzle. Since Dushanbe is in Tajikistan, not in Kirghizia, Random will always answer the above question affirmatively and it is in virtue of this unintentional predictability built into Random that we can get useful information out of him (see the embedded question lemma).

⁴ Questions of similar flavour were used in Roberts 2001.

Proof. Without loss of generality, assume that we address god A. We ask, 'Are you going to answer this question with the word that means *no* in your language?'. If his head explodes, then we know he is True and we are done. Thus we may assume that his head does not explode. We ask 'Are you going to answer this question with the word that means *yes* in your language?'. If his head explodes, then we know he is False. If his head does not explode, then we know he is Random.

We can now attain another simple solution of the modified puzzle using this lemma; to wit: first determine the identity of A via the exploding identity lemma and then depending on A's identity, ask B either: 'Are you going to answer this question with the word that means *no* in your language?' or 'Are you going to answer this question with the word that means *yes* in your language?'.

5. Boolos's original puzzle in two questions?

Prima facie, it would seem that we could prove that it requires at least three questions to determine the identities of all the gods – there are six possible ways for the gods to be arranged and each yes-no question distinguishes at most two possibilities, so we need at least $\log_2(6)$, i.e. 3, questions to determine their identities. However, the assumption that each yes-no question distinguishes at most two possibilities is in error. It is possible to distinguish three possibilities with one question if we ask a question that has the possibility of exploding a god-head. To illustrate we solve the trivial puzzle in two questions.

The trivial puzzle. Three gods A, B, and C are called, in some order, 'Zephyr', 'Eurus', and 'Aeolus'. The gods always speak truly. Your task is to determine the identities of A, B, and C by asking three yes-no questions; each question must be put to exactly one god. The gods understand English and will answer in English.

Tempered liar lemma. If we ask A 'Is it the case that: [(you are going to answer "no" to this question) AND (B is Zephyr)] OR (B is Eurus)?', a response of 'yes' indicates that B is Eurus, a response of 'no' indicates that B is Aeolus, and an exploding head indicates that B is Zephyr. Hence we can determine the identity of B in one question.

Proof. Assume A responds 'yes' and B is not Eurus. Then A has answered 'yes' to the question 'Is it the case that you are going to answer "no" to this question?'. This is impossible since A tells the truth.

Assume A responds 'no' and B is not Aeolus. Then A has answered 'no' to both the question 'Is it the case that: you are going to answer "no" to this question AND B is Zephyr?' and the question 'Is it the case that B is Eurus?'. The denial of the latter indicates that B is not Eurus and is thus

Zephyr. The denial of the former indicates either that A did not answer ‘no’ or that B is not Zephyr. Contradiction.

Assume A’s head explodes and B is not Zephyr. Then B is not Eurus either; for otherwise A would answer ‘yes’. Hence, since B is neither Zephyr nor Eurus, A would deny both sides of the disjunction and hence he would answer ‘no’ to the entire question. This final contradiction completes the proof.

Now to solve the trivial puzzle in two questions, just use the tempered liar lemma to determine B’s identity in one question and then for some god that B is not, ask B if C is this god.⁷

As noted in the first section of this article, the embedded question lemma reduces finding a three-question solution to Boolos’s original puzzle to finding a three-question solution to the trivial puzzle. It seems reasonable that a similar relationship would hold for two-question solutions as well. This is indeed the case; however, care must be taken when embedding questions that contain indexicals or demonstratives, i.e. the complex demonstrative ‘this question’ refers to the innermost quotational block in which it is contained. We require a term that refers to the outermost quotational block (i.e. the outermost question type) in which it is contained. This can be achieved by introducing a name.⁸

Let the following question be named ‘Query-1’:

E(‘Is it the case that: [(in your current mental state you would always answer “da” to Query-1) AND (B is True)] OR (B is False)?’)

*Tempered liar lemma**. If we ask A Query-1, a response of ‘ja’ indicates that B is False, a response of ‘da’ indicates that B is Random, and an exploding head indicates that B is True. Hence we can determine the identity of B in one question.

Proof. Assume A responds ‘ja’ and B is not False. Then, by the embedded question lemma, the correct answer to the question ‘Is it the case that:

⁷ In the case that B is Zephyr, we are not able to ask A any more questions, since asking A the first question caused his head to explode.

⁸ There are other ways to achieve this as well. One could use a definite description, e.g. ‘the question in which this question is embedded’ or one could introduce a new indexical that functions to always refer to the outermost question (or sentence) type in which it is embedded, e.g. ‘this-question^C’. Using this new indexical we could prove a related lemma that would also provide a two-question solution to Boolos’s original puzzle: If we ask A the question *E*(‘Is it the case that: [(in your current mental state you would always answer “da” to this-question^C) AND (B is True)] OR (B is False)?’), a response of ‘ja’ indicates that B is False, a response of ‘da’ indicates that B is Random, and an exploding head indicates that B is True. The proof follows the same reasoning as the proof of the tempered liar lemma*.

[(in your current mental state you would always answer “da” to Query-1) AND (B is True)] OR (B is False)?’ is affirmative. Since B is not False, the correct answer to ‘Is it the case that in your current mental state you would always answer “da” to Query-1?’ is affirmative, but A answered ‘ja’ to Query-1. Contradiction.

Assume A responds ‘ja’ and B is not Random. Then, by the embedded question lemma, the correct answer to both the question ‘Is it the case that: in your current mental state you would always answer “da” to Query-1 AND B is True?’ and the question ‘Is it the case that B is False?’ is negative. The denial of the latter indicates that B is not False and is thus True. The denial of the former indicates that either A did not answer ‘da’ or that B is not True. Contradiction.

Assume A’s head explodes and B is not True. Then B is not False either; for otherwise A would answer ‘ja’. Hence, since B is neither True nor False, A would deny both sides of the disjunction and hence would answer ‘da’ to the entire question. This final contradiction completes the proof.

Now to solve Boolos’s original puzzle in two questions, just use the tempered liar lemma* to determine B’s identity in one question and then for some god that B is not, ask B if C is this god (in an embedded question of course).⁹

*University of California
Santa Barbara, CA 93106-309, USA
brian.rabern@gmail.com*

*Department of Mathematics
University of California
Santa Barbara, CA 93106-3090, USA
landon.rabern@gmail.com*

⁹ This paper would not have been possible without both the loving support and immense tolerance of our partners Rhiannon Rabern and Jen Sorkin and the inquisitive looks of Adisyn and Olivia Rabern. Many thanks to Tim Roberts for his encouragement and comments on earlier drafts of this paper. We would also like to thank Ian Nance for introducing us to the puzzle and for his helpful conversations about the complexities of the gods. Thanks also to Jason Sundram and Richard Chappell for looking over earlier drafts of the paper and Peterson Tretheway, who helped initiate rigorous investigations over wine, coffee and coherent one relator groups.

References

- Boolos, G. 1996. The hardest logic puzzle ever. *The Harvard Review of Philosophy* 6: 62–65. Repr. in his *Logic, Logic, and Logic*, 406–10. 1998. Cambridge, Mass.: Harvard University Press.
- Smullyan, R. 1978. *What is the Name of This Book?* Englewood Cliffs, NJ.: Prentice Hall.
- Smullyan R. 1997. *The Riddle of Scheherazade*. New York: A. A. Knopf, Inc.
- Roberts T. 2001. Some thoughts about the hardest logic puzzle ever. *Journal of Philosophical Logic* 30: 609–12.
- Young C. unpublished. How I could have solved the hardest logic puzzle ever.