In defense of the two question solution to *the hardest logic puzzle ever*[*]

Brian Rabern and Landon Rabern

Abstract

In Rabern and Rabern (2008) we presented a two question solution to 'the hardest logic puzzle ever' (as presented in Boolos (1996)), which relied on self-referential questions. In this note we respond to several worries related to this solution. We clarify our claim that some yes-no questions cannot be answered by the gods and thus that asking such questions of the gods will result in *head explosion*. We argue that the inclusion of exploding head possibilities is neither cheating nor ad hoc but is instead forced upon us by principles related to Tarski's theorem. We also respond to concerns that have been raised about our use of self-referential questions in support of the two question solution. In particular, we address the worry that there is a revenge problem lurking, which is analogous to revenge problems that arise for purported solutions to the liar paradox. And we make some further observations about the relationship between self-referential questions, truth-telling gods and the semantic paradoxes. In the appendix we give a two question solution to the modified puzzle (where Random randomly answers 'ja' or 'da').

1 Background

In Rabern and Rabern (2008) we presented several simple solutions to 'the hardest logic puzzle ever', both for the puzzle actually presented in Boolos (1996) and for

[*][Editorial Comment, added 2021] This paper was written in April 2009, and was posted online shortly thereafter with no plans for publication. Overtime it was lost and forgotten about, and became buried deep within an old dropbox folder. After the death of the second author in October 2020 the first author was looking back at unfinished joint projects and came across this draft. There are still no plans for publication and the first author has made no effort to revise or improve the draft. Doing so properly would require incorporation of the many papers published in the meantime that touch on the themes in this draft (see footnote 1). The only changes made to the original have been fixing a few typos and updating a few references. The original acknowledgements read as follows: Thanks to John Cusbert, Peter Fritz, Wolfgang Schwarz, Rachel Anderson, Jonathan Farrell, Stefan Wintein and the members of $P(i_{55})$.

the puzzle that was clearly intended.[1] The only difference in the puzzles being the way in which the god named 'Random' answers questions. In what we called *the original puzzle* Random randomly either tells the truth or lies, whereas in what we called *the modified puzzle* Random randomly either responds 'ja' or 'da'.

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, and Random *provides randomized responses.* 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 thier own language, in which the words for 'yes' and 'no' are 'da' and 'ja', in some order. You don't know which word means which.[2]

This puzzle is neutral between the original and modified puzzles but if it is supplemented with (B3) below it yields the original puzzle and if it is supplemented with (B3*) below it yields the modified puzzle.

> (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.

> (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'.

The original puzzle becomes trivial in light of the fact that there is a trick to make it irrelevant whether the god you address lies, tells the truth or randomly either lies or tells the truth. Moreover, this trick makes it irrelevant whether 'ja' means *yes* or *no.* The trick is encoded in *the embedded question lemma*, which states that when any god g is asked ⌜If I asked you 'q' (in your current mental state), would you say 'ja'?⌝, 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.[3] This lemma also provides a simple solution to the modified puzzle (it is not quite as straightforward since Random is completely unpredictable).[4]

[1] Many papers have been published that touch on similar themes, e.g. Uzquiano (2010), Wintein (2011a), Wintein (2011b), Wheeler and Barahona (2012), Wintein (2012a), Wintein (2012b), Rosenhouse (2014), and Buchanan and Conway (2017), among others.

[2] Boolos (1996), p. 62.

[3] See Rabern and Rabern (2008), p. 106

[4] The solution to the modified puzzle has the following form: Ask B, "If I asked you 'Is A Random?' would you say 'ja'?", by elimination reasoning we can determine that a particular

This much we think is uncontroversial. The worries arise with respect to our two question solution to the puzzle, which relied on self-referential questions and the phenomenon that we called "exploding heads". We first argued that some yes-no questions cannot be answered by the gods and thus that asking such questions of the gods will result in *head explosion*. Consider, what happens when we ask a truth-teller "Are you going to answer 'no' to this question?". If he says 'no' he contradicts himself and if he says 'yes' he contradicts himself. Thus, he will be unable to respond with a truthful answer.

> "But they are infallible gods! They have but one recourse – their heads explode" (Rabern and Rabern (2008), p. 109)

So we took this to prove that there are not two but three possible reactions that a truth-teller must have upon being posed a yes-no question: 'yes', 'no', and *explode*. If so, then three possibilities can be distinguished with one yes-no question. As Buchanan and Conway (2017) recently put it:

> "As they wiped the ichor from their faces, I've no doubt Rabern, and the equally merry Rabern, reflected heartlessly that with this third outcome to one question, it may be possible to identify the three gods with just two questions in Boolos' original world, since $3 \times 2 = 6$." (p. 174)

For example, the following puzzle has a solution:

The puzzle of enlightenment. On your journey to the monastery you come upon a junction with three roads leading out: the left road, the middle road and the right road. At the junction is a monk who will truthfully answer any yes-no question if he can do so without contradicting himself; otherwise, he will sit and meditate for eternity. The monk will only answer one question per traveler. What question shall you ask of the monk to find out which road leads to the monastery?

We recommend that the reader pause here and attempt a solution before preceding with this article.[5]

There are three ways the world might be and our task is to concoct one question that narrows down the space of epistemic possibilities to one. So asking questions

god X is not random. Then we ask X, "If I asked you 'Is X True?' would you say 'ja'?", this determines X's identity. Finally, ask X, "If I asked you 'Is B Random?' would you say 'ja'?", this determines the identities of the rest. See Rabern and Rabern (2008), pp. 107-108

[5] *Hint*: A monk saw a turtle in the monastery garden and asked his mentor, "All beings cover their bones with flesh and skin, so why does this being cover its flesh and skin with bones?". The mentor took off one of his sandals and covered the turtle with it.

like 'Is the middle road the monastery road?' or 'Is the left or right road the road to the monastery?', are no help since they are only guaranteed to narrow down the space of possibilities to two. What we need is a question that has the potential to put the monk into a meditative trance; a question such that the monk will answer 'yes' if the road is to the right, 'no' if the road is to the left, and will go into meditation if the monastery road is down the middle. The following is just such a question.[6]

> ξ: (you are going to answer 'no' to ξ AND the monastery road is the middle road) OR (the monastery road is the right road)?[7]

We can prove that (i) a response of 'yes' indicates that the monastery road is the right road, (ii) a response of 'no' indicates that the monastery road is the left road, and (iii) monk meditation indicates that the monastery road is the middle road.

Proof. (i) Assume the monk responds 'yes' and the monastery road is not the right road. The the monk affirms the first disjunct, and thus, affirms that he answers 'no' to ξ. Thus, he has contradicted himself. (ii) Assume that the monk responds 'no' and the monastery road is not the left road. Then the monk denies both disjuncts. The denial of the second indicates that the monastery road is not the right road, and is thus the middle road. The denial of the first conjunct indicates that the monk does not answer 'no' to ξ or that the monastery road is not the middle road. Contradiction. (iii) Assume the monk sits and meditates and the monastery road is not the middle road. The monastery road is not the right road either; for otherwise the monk would answer 'yes'. Hence, since the monastery road is neither the middle road nor the right road, the monk will deny both disjuncts and thus will answer 'no' to ξ. This final contradiction completes the proof. □

The fact that we can distinguish three possibilities with one question combined with the embedding trick mentioned above (which renders the truth-teller/liar distinction and the language of the gods irrelevant) led us to the two question solution to 'the hardest logic puzzle ever' (both the original puzzle and the modified puzzle; see appendix A).

[6] Alternatively, one could ask ζ: If, if you answer 'no' to ζ, then the temple road is the middle road, then the temple road is the right road?

[7] Naming the question makes things more clear, but it is not necessary. Consider "You are going to answer 'no' to this question and the monastery road is the middle road or the monastery road is the right road"?

2 A revenge problem?

Some readers have voiced objections to our solution. The self-referential questions like "Are you going to answer 'no' to this question?" bear some resemblance to the *Liar sentence*, i.e. the question asks of itself if it will be answered negatively and the liar sentence says of itself that its truth-value is negative.[8] Consider the liar sentence.

> λ: λ is not true.

The liar sentence cannot be assigned either true or false consistently. Assume that λ is true. If λ is true, then what it *says* must be the case; but what it says is that it is not true. Contradiction. So, λ is not true. But that is what λ says is the case, thus λ is true. Again contradiction.

There have been numerous attempts to deal with this paradox. Some extreme and arguably wrong-headed approaches either ban self-reference altogether or claim that the liar sentence is *nonsensical*.[9] Some, perhaps, more promising approaches suggest that 'true' is context-sensitive, indexical or that its meaning is constantly under some sort of "revision". And others claim that none of the purported solutions succeed, thus we must concluded that the liar sentence is both true *and* false. But for our purposes we need only focus on one very natural strategy, which is to abandon bivalence (in the sense that every declarative sentence is either *true* or *false*) and adopt a three-valued logic. There are many ways to carry this out—one type of approach is found in Kripke (1975).

The basic idea is this: The liar sentence is neither true nor false but has some third truth-value, *neuter*. But there is a complication here because now it looks like one could reason as follows:

> (1) λ is neuter.
> (2) If λ is neuter, then λ is not true.
> (3) Thus, λ is not true.
> (4) λ *says* that λ is not true.
> (5) Thus, λ is true.

[8] For a paradigmatic example of self-reference consult Rabern, Rabern and Macauley (2013), footnote 8.

[9] Self-reference or circular-reference is not sufficient for paradox, so banning it altogether is a drastic response to a more delicate problem; and self-reference or circular-reference is not even *necessary* for paradox, so the ban is ultimately ineffective (see (Yablo 1993)). The claim the liar-sentences are nonsensical is committed to the absurd consequence that the sentence 'The only sentence on the white board in 4103 Coombs is false' is meaningless, if it is written on the white board in 4103 Coombs (cf. (Kripke 1975)).

We seem to be stuck with the same paradox even when we add another truth-value. But a common way to deal with the problem is to insist that 'not' and 'true' are *weak* in the sense that they map neuter onto neuter. If ϕ is neuter, then the negation of ϕ is neuter; and a sentence which predicates 'true' of ϕ is also neuter. Thus, the inference from (1) and (2) to (3) does not go through, since premise (2) assumes that 'not' and 'true' are *strong*. Hence, λ can be consistently assigned 'neuter'.

Notice that this approach is committed to the claim that *strong negation* cannot be expressed in the object language. If, in the object language, we can express the fact that λ has a truth-value *other than* 'true', then we can reinstate the paradox. Let, 'untrue' expresses the property of *having a truth-value other than 'true'*. And let our new liar sentence, $\lambda*$, say of itself that it is *untrue*. If $\lambda*$ is neuter, then it is untrue...etc.

This type of problem is a *revenge problem* for purported solutions to the liar paradox.[10] Matti Eklund nicely sums up the basic form of these problems as follows:

> The standard form of the revenge problem is this: the expressive resources of our language allow us to exhaustively and exclusively divide sentences into the true ones and the rest. If our language has sufficient expressive resources to state an exhaustive and exclusive division of all sentences into the true ones and the rest, paradox can be reinstated. Just let our new liar sentence say of itself that it belongs to the rest.[11]

It may seem that we are going to be faced with a sort of revenge problem with respect to our treatment of the truth-telling gods. When the god is posed the question "Are you going to answer 'no' to this question?", we suggest that his head explodes. And we use this assumption to construct our solution. But, the worry goes, haven't we basically assumed that there is a response that the god can give, which isn't susceptible to revenge-like problems? (Moreover, by analogy, haven't we assumed that there is a solution to the liar paradox, which isn't susceptible to revenge problems?)

In this vein, Wintein (2011b) asks: "what would happen if we asked True: 'is the case that your answer to this question is 'no' or that you explode on this question?' Such strengthened Liar objections show that [Rabern and Rabern]'s proof is suspect, to say the least, and that an explanation of the assumptions involved is needed" (p.612).

[10]The discussion of revenge problems in this section closely follows Eklund (2008) and various lecture notes from Eklund's seminar on paradoxes.

[11]Eklund (2008).

More generally one may wonder what happens when the god is asked "Are you going to respond with something *other than* 'yes' to this question?". If the god explodes, then this indicates that the truthful answer to the question was 'yes', thus suffering a head explosion is in some sense "inconsistent" with telling the truth. Hence adding the head explosion option only seems to be a temporary fix. Just as one can always concoct a revenge sentence for attempted solutions to the liar paradox, one can always concoct a revenge question for attempted repairs of the truth-telling god. And so the truth-telling god who answers 'yes', 'no' or explodes is completely unstable. Since our two question solution relied on the responses of such a god, our purported two question solution seems sketchy at best.[12]

3 Proof of the non-existence of a god

In reply, we would first like to make one thing very clear: *god does not exist.* At least, a god who always tells the truth does not exist. And we can easily prove it. To be precise we will prove that a god with the following property does not and cannot exist; *being such that one answers all yes-no questions truthfully with either 'yes' or 'no'.* Call such a god an *absolute truth-telling god.*

Non-existence theorem. An absolute truth-telling god is logically impossible.

Proof. Assume (to reach a contradiction) that an absolute truth-telling god g exists. Ask g the following yes-no question: 'Are you going to answer 'no' to this question?'[13] If g responds 'yes', then she affirms that she answers 'no' but she did not answer 'no'. Thus, she did not tell the truth. If instead g responds 'no', then she denies that she answers 'no' but she did answer 'no'. Either way we get a contradiction. Thus, an absolute truth-telling god is logically impossible.[14] □

[12]Concerns of this general sort have been raised on a few weblogs where we in turn gave proto-versions of this paper, e.g. in the comments thread on Kenny Easwaran's *LiveJournal* (December 2007) and a post at *XOR's Hammer* (August 2008) among others. And a similar objection has recently been raised in Wintein (manuscript) (cf. Wintein (2011b)).

[13]Here we assume that it is logically possible to ask the absolute truth-teller any question. One could, in principle, turn the argument into a *reductio* of this assumption. The move here bears some resemblance to an interesting reaction to the grandfather paradox – backwards time travel is possible its just that it is impossible to pull the trigger when one does so (i.e. the timeline is fixed). Arguing that it is logically impossible to ask the absolute truth-teller certain questions ultimately falls into the category of restricting the set of allowable questions, which we discuss and reject later.

[14]Note that this theorem can be proven even without the use of a self-referential question. Ask g, "If I repeatedly asked you 'Are you going to answer 'no' to all my future questions?' would you always say 'no'?". If she says 'yes', then she affirms that in the future she will always

Notice that this is essentially *Tarski's undefinability theorem*, i.e. there is no truth predicate for a language, which has the expressive resources to talk about it's own sentences (while employing classical logic), that satisfies the T-schema. For a predicate to satisfy the T-schema a valid schema must result when the predicate is substituted for 'Tr' in $\ulcorner Tr(\psi) \equiv s \urcorner$ (where instances of this schema are obtained by substituting sentences for 's' and substituting names of the corresponding sentences for ψ). See (Tarski 1935).

The analog for the absolute truth-telling god is this: there is no truth-telling god that answers yes-no questions from a language, which has the expressive resources to talk about its own sentences (while employing classical logic), that satisfies the g_A-schema where for a god to satisfy the g_A-schema is for a valid schema to result when a functor associated with the god's answering abilities is substituted for 'g_A' in $\ulcorner (g_A(\psi) = 1) \equiv s \urcorner$ (where instances of this schema are obtained by substituting sentences for 's' and substituting names of the corresponding sentences for ψ). Intuitively, we can think of g_A, which denotes the divine answering function, as taking a declarative sentence ψ as argument interpreting it as a yes-no question and outputting either 'yes' or 'no' depending on the truth-value of ψ. The theorem proves that there is no such function.

A moral here is that if one were to infer from the premise that the truthful answer to a question is 'yes', to the conclusion that the truth-teller *must* answer 'yes', one would be neglecting a valuable Tarskian lesson.

4 No revenge

Then what sort of god must the puzzle be talking about? We can safely assume that any god under consideration will only make true statements. What other properties would we like a 'truthy' puzzle-god to have? Surely, we want her to tell the truth when doing so won't lead her to contradict herself.

There are two ways to ensure this result: either restrict the set of allowable questions or expand the set of possible responses. We could stipulate that no self-

answer 'no' to "Are you going to answer 'no' to all my future questions?". But then she commits herself to denying that she answers 'no' when she does answer 'no'. Thus, she will not tell the truth. If instead she responds 'no', then she denies that in the future she will always answer 'no', and thus that at some point she will answer 'yes'. But if she does so, then she will affirm that from that point on she will always answer 'no' to "Are you going to answer 'no' to all my future questions?". And she will again commit herself to denying that she answers 'no' when she does answer 'no'. And she will again not tell the truth. Thus, this proves (without self-referential questions) that an absolute truth-telling god is logically impossible. Alternatively, ask g, "Would you answer 'no' to all of the Yabloean questions?", where a *Yabloean question* is one of the form q_k: $\ulcorner$ Would you answer 'no' to all q_i for $i > k$?$\urcorner$ (where k is any natural number). See Yablo (1993).

referential questions are allowable or that no counterfactual questions are allowable or that no questions which ask about responses (e.g. questions that contain phrases like "answer 'yes'" or "answer 'no'") are allowable, etc.

We think such moves are unmotivated. First, restrictions of this sort tend to go against that spirit of the original Smullyan puzzles where one asks questions like "Would the other guard tell me that this door leads to the castle?" or "If I asked you 'Are you a knave?' would you say 'yes'?". See, for example, Smullyan (1978), Smullyan (1998), among many others. Secondly, such restrictions rule out too much. Self-referential questions like "Is this a self-referential question?" are perfectly answerable and can even be used to gain valuable information from the gods.

Moreover, it is difficult, if not impossible, to settle on a principled restriction on the set of allowable questions, without it reducing to "No questions that cause trouble". When we take into account Kripke's lessons about so-called *empirical liars*, the task of finding a principled restriction looks hopeless. Consider the question "Would you answer 'no' to the only question written on the white board in room 4103 Coombs?". Whether we can ask that question of the gods will depend on whether or not it is tokened on the white board in room 4103 Coombs. It is for analogous reasons that Kripke states "it would be fruitless to look for an intrinsic criterion that will enable us to sieve out as meaningless, or ill-formed those sentences which lead to paradox...There can be no syntactic or semantic 'sieve' that will winnow out the 'bad' cases while preserving the 'good' ones" (Kripke (1975), 692).

We think it is more natural and more interesting to put no restrictions on the set of allowable questions but to patch up the the truth-telling god so that she can coherently deal with all questions. What shall we have the god do if she can't answer 'yes' or 'no' without contradicting herself? She could

(1) sit there forever not responding,
(2) say "I can't answer your question without contradicting myself",
(3) suffer a head explosion,
(4) say "The truthful answer to your question is neither 'yes' nor 'no'",
...

Note an important distinction between the first three responses and response (4) – in response (4) she makes a claim about *the truthful answer* to the question, in the others she does not. There is an important distinction between providing a truthful response to a question and doing something in reaction to a question, e.g. making a true assertion after or about a question, being embarrassed by a question, looking confused as the result of a question, putting your shoe on a turtle

after a question, etc. Response (4) is in the first camp while the others are in the second. Responses (1), (2), and (3) are essentially the same, but we prefer not to use (1) as it brings in temporal complications that are irrelevant to the problem at hand.[15]

Also, we prefer (3) over (2) for dramatic effect.

Definition 1. *An ambitious truth-telling god is one which when posed any yes-no question*

> *(i) answers 'yes' if she can do so without contradicting herself and answering 'yes' is truthful, otherwise*
> *(ii) answers 'no' if she can do so without contradicting herself and answering 'no' is truthful, otherwise*
> *(iii) she tries so hard to tell the truth that her head explodes.*

Given these conditions it follows that when any yes-no question q is asked of an ambitious god she will either respond 'yes' or 'no' or she will *explode*. Thus, ambitious gods are by design not subject to revenge problems.

To make the contrast clear, let's consider the sort of truth-telling god that results from taking option (4) above.

Definition 2. *An arrogant truth-telling god is one which when posed any yes-no question*

> *(i) answers 'yes' if she can do so without contradicting herself and answering 'yes' is truthful, otherwise*
> *(ii) answers 'no' if she can do so without contradicting herself and answering 'no' is truthful, otherwise*
> *(iii) she audaciously responds "The truthful answer to your question is neither 'yes' nor 'no'".*

Let g be an arrogant god. Ask her "Are you going to respond 'no' to this question?". If she says 'yes' or 'no', then she contradicts herself. So she says, "The truthful answer to your question is neither 'yes' nor 'no'". Thus, the truthful answer is 'no', since she did not answer 'no'. But g asserted that the truthful answer was neither 'yes' nor 'no' and in particular, that it was not 'no'. Thus, the arrogant god has contradicted herself. The arrogant god is vulnerable to revenge.

[15]See Ellis (2008), "We should suppose that the gods may take an arbitrarily long time to answer questions. This precludes us from asking questions which at least one of the gods cannot answer: if we were to ask it to him, we'd never know if he couldn't answer or if he just hadn't answered yet (cf. recursive enumerability). In this way we avoid the possibility of Rabern and Rabern's 'exploding head' answers."

Now let's try the same reasoning with an ambitious god. Let g be an ambitious god. Ask her "Are you going to respond 'no' to this question?". If she says 'yes' or 'no', then she contradicts herself. So, she has a head explosion. Thus, the truthful answer is 'no', since she did not answer 'no'. But we cannot derive a contradiction from this. All that follows from the the fact that the truthful answer is 'no' is that she answers 'no' if she can do so without contradicting herself or she explodes.

5 Gods, questions, and paradoxes

The relationship the semantic paradoxes and self-referential questions or truth-telling gods raises interesting questions. Providing a solution to the liar paradox is like providing a consistent response for the truth-teller, which isn't vulnerable to a revenge problem. What answer should the god give? It seems that any attempt at repairing the god such that she always answers with the correct response is open to a revenge problem. As we see it, either the puzzle concerns the ambitiuos god as we have defined it or there must be restrictions on which yes-no questions one can ask of the god. What the analog of this principle is for the liar paradox isn't clear. Does it motivate a kind of quetism about the liar paradox? When confronted with the liar sentence is the best we can do just shrug our shoulders? This is not very satisfying. Or is it that you should either opt for quietism or give up on the idea that language is maximally expressive, i.e. give up universality?

The sorites paradox was first given in the form of a series of yes-no questions and much discussed by the Stoics.[16] Chrysippus believed that for every sentence ϕ there there was one correct answer to the question 'ϕ?', namley 'yes' if 'ϕ' is true and 'no' if ϕ is false. Thus, for every sequence of sentences $\phi_1, \phi_2, ..., \phi_n$ there is one sequence of correct answers to the questions 'ϕ_1?','ϕ_2?',. . . ,'ϕ_n?', each member of which is either 'yes' or 'no'. But of course a sorites series of questions causes trouble for this Stoic view. In a series of questions of the form $\ulcorner$Are i grains a heap?$\urcorner$, where $i \in \mathbb{N}$, if there is a point in the series where the correct answer is 'no' and a point in the series at which the correct answer is 'yes', then there must be a unique point at which the answers switch from 'no' to 'yes'. But it seems implausible that one grain of sand can make a difference between being a heap and not being a heap. To this problem Chrysippus recommended that at some point in the sorites interrogation one should fall silent. Would Chrysippus recommend the same thing for the inquisitive version of the liar paradox, i.e. "Are you going to respond 'no' to this question?"?[17]

[16] Williamson (1994), pp. 12-22.

[17] Williamson (1994) uses the story of Chrysippus and the Stoics to motivate his version of Epistemicism. Perhaps similar motivation could be adapted to the liar paradox.

Question. For which questions will an ambitious truth-teller explode and what relation does this question bear to Kripke's sieve?

A A two question solution to the modified puzzle

In Rabern and Rabern (2008) we only provided a two question solution to Boolos' original puzzle, where Random randomly tells the truth or lies. Some people have thought that the modified puzzle cannot likewise be solved in two questions. But it can be done.

Since the god languages and the distinction between truth-tellers and liars are made irrelevant by using embedded questions, we assume that there are two English-speaking ambitious truth-tellers and one god that answers 'yes' or 'no' randomly.

Generic tempered liar lemma. Let g be an ambitious truth-teller. Let p_1, p_2, and p_3 express any three propositions exactly one of which is true. If we ask g "Is it the case that: [(you are going to answer 'no' to this question) AND p_1] OR p_2?", a response of 'yes' indicates that p_2 is true, a response of 'no' indicates that p_3 is true, and an exploding head indicates that p_1 is true.

Proof. Assume g says 'yes' and p_2 is false. Then g has said 'yes' to the question "Is it the case that you are going to answer 'no' to this question?". This is impossible since g tells the truth. Assume g says 'no' and p_3 is false. Then g has said 'no' to both the question "Is it the case that: [(you are going to answer 'no' to this question) AND P_1]?" and the question "Is it the case that p_2?". The denial of the latter indicates that p_2 is False and thus p_1 is True. The denial of the former indicates either that g did not answer 'no' or that p_1 is false. Contradiction. Assume g's head explodes and p_1 is false. Then p_2 is false also, for otherwise g could have said 'yes'. Hence, since both p_1 and p_2 are false, g could have said 'no' to both sides of the disjunction and hence she could have said 'no' to the entire disjunction. This final contradiction completes the proof. □

The core puzzle. Three gods A, B, and C are called, in some order, 'Jarrah', 'Wongoola', and 'Yapunyah'. Wongoola and Yapunyah will truthfully answer any yes-no question if they can do so without contradicting themself; otherwise their heads will explode. Jarrah randomly responds 'yes' or 'no' to any yes-no question. Your task is to determine the identities of A, B, and C by asking *two* yes-no questions; each question must be put to exactly one god. The gods understand English and will answer in English.

For ease of exposition, we can set up a question schema using the generic tempered liar lemma: For any god y, let Q_y be the generic tempered liar question with $p_1 = \ulcorner y$ is Jarrah$\urcorner$, $p_2 = \ulcorner y$ is Wongoola$\urcorner$, and $p_3 = \ulcorner y$ is Yapunyah$\urcorner$. *First question*: ask Q_B of A. If A's head explodes, then A is not random and we conclude that B is Jarrah. Thus, C is a truth-teller. Ask C "Are you Wongoola?" to determine the identities of the rest. If A answers 'yes', then either A is random or B is Wongoola, A is Yapunyah and C is Jarrah. Similarly, if A answers 'no', then either A is random or B is Yapunyah, A is Wongoola and C is Jarrah. Thus, if A answers either 'yes' or 'no', it follows that B is *not* random. *Second question*: ask Q_C of B to determine C's identity. If C is Jarrah, then A is not random and the identities are determined by A's response. If C is Wongoola, then A must be random (Jarrah) and thus B is Yapunyah. If C is Yapunyah, then again A must be random (Jarrah) and thus B is Wongoola.[18]

References

Boolos, G.: 1996, The hardest logic puzzle ever, *The Harvard Review of Philosophy* **6**, 62–65. (Reprinted in: George Boolos, "Logic, Logic, and Logic", 1998, pp. 406-410, Cambridge, Mass.: Harvard University Press.).

[18]Recently, in Ellis (2008), Tom Ellis gave a generalization of 'the hardest logic puzzle ever' where there are $2n + 1$ for any $n \geq 1$.

A harder puzzle. Before you sit $2n + 1$ gods. You know that at most n of them are modified random gods and the rest are either truth-tellers or liars. Each god speaks her own private language where 'ja' and 'da' mean *yes* and *no* in some order. There is a fork in the road, one path leads to a castle. Your task is to find the way to the castle using only $2n$ yes-no questions.

Using the embedded question lemma* this immediately reduces to the following puzzle. (As we noted in the earlier paper, the fact that the gods have their own private language is irrelevant to the embedded question lemma*.)

A harder puzzle (simplified). Before you sit $2n + 1$ gods. You know that at most n of them are modified random gods and the rest are truth-tellers. Each god speaks English. There is a fork in the road, one path leads to a castle. Your task is to find the way to the castle using only $2n$ yes-no questions.

As noted in Ellis (2008), if we can find a truth-teller in $2n - 1$ questions then we can solve the puzzle. Ellis leaves this problem to the reader as an exercise. We have a truly marvelous solution of this puzzle, which this footnote is too small to contain (we know from private communication with Ellis that his solution is different).

Question. Is it possible to determine the way to the castle in fewer than $2n$ questions?

Buchanan, A. G. and Conway, J. H.: 2017, An island tale for young anthropologists, *Raymond Smullyan on Self Reference*, Springer, pp. 165–180.

Eklund, M.: 2008, The Liar Paradox, Expressibility, Possible Languages, *Revenge of the Liar: New Essays on the Paradox* .

Ellis, T.: 2008, Even harder than *the hardest logic puzzle ever*. Online note. Statistical Laboratory, University of Cambridge.
URL: *http://www.srcf.ucam.org/ te233/maths/puzzles/evenharder.html*

Kripke, S.: 1975, Outline of a Theory of Truth, *Journal of Philosophy* **72**(19), 690–716.

Rabern, B. and Rabern, L.: 2008, A simple solution to the hardest logic puzzle ever, *Analysis* **68**(298), 105–112.

Rabern, L., Rabern, B. and Macauley, M.: 2013, Dangerous reference graphs and semantic paradoxes, *Journal of Philosophical Logic* **42**(5), 727–765.

Rosenhouse, J.: 2014, Knights, knaves, normals, and neutrals, *The College Mathematics Journal* **45**(4), 297–306.

Smullyan, R. M.: 1978, *What is the name of this book? The riddle of Dracula and other logical puzzles*, Dover Publications.

Smullyan, R. M.: 1998, *The riddle of Scheherazade and other amazing puzzles, ancient & modern*, Houghton Mifflin Harcourt.

Tarski, A.: 1935, The concept of truth in formalized languages, *Logic, semantics, metamathematics (1956)*, Oxford, Clarendon Press, pp. 152–278.

Uzquiano, G.: 2010, How to solve the hardest logic puzzle ever in two questions, *Analysis* **70**(1), 39–44.

Wheeler, G. and Barahona, P.: 2012, Why the hardest logic puzzle ever cannot be solved in less than three questions, *Journal of philosophical logic* **41**(2), 493–503.

Williamson, T.: 1994, *Vagueness*, Routledge.

Wintein, S.: 2011a, A framework for riddles about truth that do not involve self-reference, *Studia Logica* **98**(3), 445–482.

Wintein, S.: 2011b, On languages that contain their own ungroundedness predicate, *Logique et Analyse* pp. 599–615.

Wintein, S.: 2012a, Assertoric semantics and the computational power of self-referential truth, *Journal of philosophical logic* **41**(2), 317–345.

Wintein, S.: 2012b, On the behavior of true and false, *Minds and Machines* **22**(1), 1–24.

Wintein, S.: manuscript, Why Rabern and Rabern did not solve the hardest logic puzzle ever in two questions.

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