The Possible Asylum

Bart Bogaerts

July 3, 2024

Abstract

Last year, Jeremy Avigad and coauthors published a paper in this journal entitled "The Impossible Asylum" showing, using automated reasoning tools, that the last puzzle of Raymond Smullyan's article "The Asylum of Doctor Tarr and Professor Fether" is inconsistent. We here argue to the contrary: the puzzle is consistent after all. Our solution was also found using automated reasoning and is small enough to be presented here. In fact, the smallest asylum satisfying the original puzzle is only inhabited by four persons.

So how can we explain this discrepancy, where on the one hand we have a formal proof of inconsistency and other hand a satisfying assignment? Well, this is due to an ambiguity in the natural language statement of the original puzzle. We will argue that with a more natural reading of the original puzzle, its logical representation changes ever-so-slightly and becomes consistent.

1 The Problem and Solution.

We refer the reader to the work of Jeremy Avigad et al. [1] (from now on referred to as ABBHN) for a full exposition of the problem and its history and jump straight to the point. Raymond Smullyan's original puzzles [2] are about an investigative journalist visiting several asylums of which each inhabitant is either a doctor or a patient, and each inhabitant is either sane or insane. (In)sanity has a very strong interpretation in these puzzles: a sane person believes a statement if and only if that statement is true; an insane person believes it if and only if it is false. In this context, his last problem is the following.

> "The last asylum that Craig visited he found to be the most bizarre of all. This asylum was run by two doctors named Doctor Tarr and Professor Fether. There were also other doctors on the staff. Now, an inhabitant was called peculiar if he believed that he was a patient. An inhabitant was called special if all patients believed he was peculiar and no doctor believed he was peculiar. Inspector Craig found out that the following condition holds.

Condition C: Each inhabitant has a best friend in the asylum. Moreover, given any two inhabitants—A, B—if A believes that B is special, then A's best friend believes that B is a patient.

Shortly after this discovery, Inspector Craig had private interviews with Dr. Tarr and Professor Fether. Here is the interview with Doctor Tarr:

Craig: *Tell me, Doctor Tarr, are all the doctors in this asylum sane?*

Tarr: *Of course they are!*

Craig: *What about the patients; are they all insane?*

Tarr: *At least one of them is.*

The second answer struck Craig as a surprisingly modest claim! Of course, if all the patients are insane, then it certainly is true that at least one is, but why was Dr. Tarr so cautious? Anyway Craig then had his interview with Professor Fether, which went as follows:

Craig: *Dr. Tarr said that at least one patient here is insane. Surely, that is true, isn't it?*

Professor Fether: *Of course it is true! All the patients in this asylum are insane!*

What kind of asylum do you think we are running?

Craig: *What about the doctors; are they all sane?*

Professor Fether: *At least one of them is.*

Craig: *What about Dr. Tarr; is he sane?*

Professor Fether: *Of course he is! How dare you ask me such a question?*

At this point, Craig realized the full horror of the situation! What was it?" — [2]

Before describing how we found it, we present a state of affairs that is consistent with this story. In our solution, there are four inhabitants: *Tarr*, *Fether*, and *Alice* are doctors, while *Bob* is the sole patient. This is not a coincidence: with our interpretation of condition C, all solutions have exactly one patient. As already observed by Smullyan [3] (and in Line with Edgar Allan Poe's short story "The System of Doctor Tarr and Professor Fether"), all doctors are insane, while all patients (just one in this case) are sane. It follows that everyone is peculiar. *Bob*, the sole sane inhabitant of this horrible asylum correctly believes everyone to be peculiar, and all the (insane) doctors (incorrectly) believe no-one to be peculiar. As a result, everyone is special. Our proposed solution is also consistent with the interviews: it can be verified that all of Dr. Tarr and Professor Fether's statements are false in this instance. The only thing that is then left to verify is that condition C is indeed satisfied. To argue this, we need to know of course who everyone's best friend is. In the solution we found, *Bob*'s best friend is *Fether* and *Bob* is the best friend of each of the doctors. For verifying that conditions C indeed holds for each two inhabitants (A and B), we consider two cases:

- If A is a doctor, then A does not believe B to be special (since A is insane but each inhabitant, and in particular B, is special). So the condition is vacuously satisfied.
- If A is a patient, then A can only be *Bob* and A correctly believes B to be special. In this case, B is one of the doctors. We should verify that *Fether* (*Bob*'s best friend) believes that B is a patient. Since Fether is insane and B is not a patient, this is indeed the case.

So how can we explain this solution in light of the existing proofs of inconsistency of the puzzle [1]? The answer is that we have a different interpretation of condition C. The puzzle states

> "Moreover, given any two inhabitants—A, B— [...]".

ABBHN translate this into first-order logic as a statement of the form

$$\forall A, B : \varphi,$$

where φ expresses the part of condition left out in the quote above. We, on the other hand, would argue the correct translation is

$$\forall A, B : A \neq B \Rightarrow \varphi.$$

In other words, the point in which we disagree is whether or not the original statement implies that A and B are distinct individuals. We believe the word "two" in the puzzle does imply this. When people, in natural language, speak about **two** inhabitants, they typically mean two distinct ones and most people would agree that *Tarr*, *Tarr* does not constitute *two* inhabitants. It is one inhabitant, just named twice. If Smullyan would have written "For every individual A and every individual B", it would be debatable whether or not he intended to mean that they are distinct; but the explicit use of the word "two" in his statement, to us clearly suggests that A and B cannot be the same.

It is common for students who learn logic to assume that different variables "automatically" refer to different objects. For instance, the formula

$$\exists x, y : P(x) \wedge P(y)$$

will easily be read as "there are at least two Ps". This is, however, not the meaning of this first-order logic formula: this formula is satisfied in all structures in which there is at least *one* object satisfying P (since x and y can just take the same value). When modelling natural language statements in logic, this is something to be wary about.

This common mistake, however, is not what happened in the work of ABBHN, who are experienced practitioners of logical methods. In their paper, not only in the logical parts, but also in their human-readable proofs (extracted from the automatically generated proofs), they repeatedly make use of the fact that condition C can be instantiated with A and B referring to the same person.

The only reasonable explanation for this discrepancy is thus that we disagree on the meaning of "two inhabitants—A,B—".

So which of the two[1] readings of this natural language sentence is *correct*? Unfortunately, that is something that only the late Raymond Smullyan could have told us. While we are of the opinion that the natural reading of the English text is that "two Xs" means "two distinct Xs", it deserves to be mentioned that[2] there is good evidence that Smullyan actually had ABBHN's interpertation in mind. First of all, in his solution, Smullyan applies condition C to A and their best friend B without explicitly stating that A is not their own best friend, which provides some support for ABBHN's reading. Secondly, the 2023 endnotes of this very journal [4] mention that Dr. Pieter Audenaert has contacted the journal after reading ABBHN's article with the information that Raymond Smullyan knew at least since 2001 that the puzzle was inconsistent (citing personal communications).

2 How We Found the Solution.

To find this solution, we modelled the puzzle in the IDP language [5]. This language is a rich extension of typed first-order logic, supporting features such as inductive definitions [6] and aggregates [7], but none of these features were really necessary for modelling this puzzle. Our representation closely follows the first-order representation of ABBHN, it also makes use of the trick to model "A believes φ" as

$$Sane(A) \Leftrightarrow \varphi,$$

using the very strong definition of (in)sanity assumed in the puzzle. Indeed, this formula is true if A is sane and φ holds or if A is insane and φ does not hold. The minor points in which it differs are

- We make use of explicit predicate symbols *Insane* and *Patient* for reasons of clarity. This possibility was also mentioned by ABBHN.
- For all definitions in the puzzle, we make use of definitions in the IDP language. For instance, peculiarity is defined by

  $$\{\forall x : Peculiar(x) \leftarrow (Sane(x) \Leftrightarrow Patient(x))\}.$$

  This definition should be read as "Peculiar is defined as follows: x is peculiar if x is sane if and only if x is a patient" or, using the trick mentioned above, "Peculiar is defined as follows: x is peculiar if x believes they are a patient". In other words, the definitional arrow $\leftarrow$ corresponds to the "if" that is often found in mathematical definitions, and not to the "if" often found in assertions, which then translates to material implication $\Leftarrow$.

[1] The two distinct readings, to avoid all potential confusion.
[2] As pointed out during the reviewing process of this paper.

In the specific case of a definition with only a single case and without recursion (as is the case here), this definition can be replaced by a logical equivalence $\Leftrightarrow$, in this case

$$\forall x : Peculiar(x) \Leftrightarrow (Sane(x) \Leftrightarrow Patient(x)),$$

but in general, IDP also supports, next to standard first-order sentences, *inductive definitions*. For instance, the definition

$$\left\{ \begin{array}{l} \forall x\ y : T(x,y) \leftarrow E(x,y). \\ \forall x\ y : T(x,y) \leftarrow \exists z : E(x,z) \wedge T(z,y) \end{array} \right\}$$

expresses that T is the transitive closure of E, which cannot be expressed in standard first-order logic. We believe it is clarifying from the modelling perspective to make definitions explicit.

- We have not included the constraint $\exists x : Sane(x)$ in the specification of the puzzle since it does not appear explicitly in the original puzzle [2]. It is, however, included explicitly in a modified version that appeared later in a puzzle book [8].

The IDP system is a *knowledge base system* [9], meaning it can, on a single (logical) theory (or "knowledge base") execute multiple forms of inference, including automated theorem proving [10], model expansion [11], and methods for *temporal reasoning* [12]. For solving and analyzing the puzzle at hand, we used two types of inference.

Firstly, we made use of IDP's *theorem proving* capabilities in order to confirm some claims made by Smullyan himself when discussing the solution to his original problem [3]. We also used this to confirm that ABBHN's assumption that there is at least one sane person (condition 6) is indeed entailed. However, it deserves to be mentioned here that this is non-trivial and is in fact *not* entailed from the pre-interview knowledge alone. Finally, we used this capability to confirm ABBHN's claim of inconsistency in case the stronger condition C is imposed. None of these results are new. In fact, the technology that is used under the hood is the same as what ABBHN built on: for theorem proving, IDP translates its input into the TPTP format [13] and then calls an out of the box theorem prover that supports this input format (the default solver used is SPASS [14], but this can be configured). We included these checks here as a sanity check and to illustrate the flexibility the knowledge base system paradigm offers.

Secondly, we made use of IDP's *model expansion* inference [11]. This inference method takes as input a logical theory and a structure with a finite domain interpreting some, but not necessarily all of the symbols. The output is a structure that expands the input and that satisfies the input theory. In our case, we presented it with a structure with an increasing domain size. For a domain size of four, the system was able to produce a model, namely the solution discussed above. Internally, to execute model expansion, IDP translates its input into (an extension of) propositional logic and then calls an (extended) SAT solver [15].

Our solution can be run and modified at `http://dtai.cs.kuleuven.be/krr/idp-ide/?src=1070d22152ca7f9d4990dda70f84e72c` or downloaded from Zenodo [16].

3 Conclusion

This paper once more highlights the importance of formal specifications by identifying an ambiguity in natural language. The ambiguity here even arises in the very controlled setting of logical puzzles, where usually not a lot of room for interpretation exists. Additionally, the paper also points out that the conclusions reached by automated reasoning tools are only as good as the formal specifications they start from. Neither of these observations is novel, or surprising, but they are nonetheless good to keep in mind.

References

[1] Avigad J, Baek S, Bentkamp A, Heule M, Nawrocki W. An Impossible Asylum. Am Math Mon. 2023;130(5):446-53. Available from: `https://doi.org/10.1080/00029890.2023.2176668`.

[2] Smullyan R. The Asylum of Doctor Tarr and Professor Fether. The Two-Year College Mathematics Journal. 1982;13(2):142-6.

[3] Smullyan R. The Asylum of Doctor Tarr and Professor Fether: Solutions. The Two-Year College Mathematics Journal. 1982;13(3):213-7.

[4] Editor's Endnotes. The American Mathematical Monthly. 2023;130(10):968-8. Available from: `https://doi.org/10.1080/00029890.2023.2277092`.

[5] De Cat B, Bogaerts B, Bruynooghe M, Janssens G, Denecker M. Predicate logic as a modeling language: the IDP system. In: Kifer M, Liu YA, editors. Declarative Logic Programming: Theory, Systems, and Applications. ACM / Morgan & Claypool; 2018. p. 279-323. Available from: `https://doi.org/10.1145/3191315.3191321`.

[6] Denecker M, Ternovska E. A logic of nonmonotone inductive definitions. ACM Trans Comput Log. 2008;9(2):14:1-14:52. Available from: `https://doi.org/10.1145/1342991.1342998`.

[7] Pelov N, Denecker M, Bruynooghe M. Well-founded and stable semantics of logic programs with aggregates. Theory Pract Log Program. 2007;7(3):301-53. Available from: `https://doi.org/10.1017/S1471068406002973`.

[8] Smullyan R. The Lady or the Tiger? and Other Logic Puzzles. Alfred A. Knopf; 1982.

[9] Denecker M, Vennekens J. Building a Knowledge Base System for an Integration of Logic Programming and Classical Logic. In: Garcia de la Banda M, Pontelli E, editors. Logic Programming, 24th International Conference, ICLP 2008, Udine, Italy, December 9-13 2008, Proceedings. vol. 5366 of Lecture Notes in Computer Science. Springer; 2008. p. 71-6. Available from: https://doi.org/10.1007/978-3-540-89982-2_12.

[10] Fitting M. First-Order Logic and Automated Theorem Proving, Second Edition. Graduate Texts in Computer Science. Springer; 1996. Available from: https://doi.org/10.1007/978-1-4612-2360-3.

[11] Mitchell DG, Ternovska E. A Framework for Representing and Solving NP Search Problems. In: Veloso MM, Kambhampati S, editors. Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA. AAAI Press / The MIT Press; 2005. p. 430-5. Available from: http://www.aaai.org/Library/AAAI/2005/aaai05-068.php.

[12] Bogaerts B, Jansen J, Bruynooghe M, De Cat B, Vennekens J, Denecker M. Simulating Dynamic Systems Using Linear Time Calculus Theories. Theory Pract Log Program. 2014;14(4-5):477-92. Available from: https://doi.org/10.1017/S1471068414000155.

[13] Sutcliffe G, Suttner C. The TPTP Problem Library for Automated Theorem Proving;. https://tptp.org/.

[14] Weidenbach C, Dimova D, Fietzke A, Kumar R, Suda M, Wischnewski P. SPASS Version 3.5. In: Schmidt RA, editor. Automated Deduction - CADE-22, 22nd International Conference on Automated Deduction, Montreal, Canada, August 2-7, 2009. Proceedings. vol. 5663 of Lecture Notes in Computer Science. Springer; 2009. p. 140-5. Available from: https://doi.org/10.1007/978-3-642-02959-2_10.

[15] De Cat B, Bogaerts B, Devriendt J, Denecker M. Model Expansion in the Presence of Function Symbols Using Constraint Programming. In: 25th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2013, Herndon, VA, USA, November 4-6, 2013. IEEE Computer Society; 2013. p. 1068-75. Available from: https://doi.org/10.1109/ICTAI.2013.159.

[16] Bogaerts B. IDP Specifications For Raymond Smullyan's Asylum Puzzle. Zenodo; 2024. Available from: https://doi.org/10.5281/zenodo.12608036.