Solution to the xkcd Blue Eyes puzzle

Jakub Kdzioka

December 14, 2021

Abstract

In a puzzle published by Randall Munroe [2], perfect logicians forbidden from communicating are stranded on an island, and may only leave once they have figured out their own eye color. We present a method of modeling the behavior of perfect logicians and formalize a solution of the puzzle.

Contents

1 Introduction

The original problem statement [2] explains the puzzle well:

> A group of people with assorted eye colors live on an island. They are all perfect logicians – if a conclusion can be logically deduced, they will do it instantly. No one knows the color of their eyes. Every night at midnight, a ferry stops at the island. Any islanders who have figured out the color of their own eyes then leave the island, and the rest stay. Everyone can see everyone else at all times and keeps a count of the number of people they see with each eye color (excluding themselves), but they cannot otherwise communicate. Everyone on the island knows all the rules in this paragraph.
>
> On this island there are 100 blue-eyed people, 100 brown-eyed people, and the Guru (she happens to have green eyes). So any given blue-eyed person can see 100 people with brown eyes and 99 people with blue eyes (and one with green), but that does not tell him his own eye color; as far as he knows the totals could be 101 brown and 99 blue. Or 100 brown, 99 blue, and he could have red eyes.
>
> The Guru is allowed to speak once (let's say at noon), on one day in all their endless years on the island. Standing before the islanders, she says the following:
>
> "I can see someone who has blue eyes."
>
> Who leaves the island, and on what night?

It might seem weird that the Guru's declaration gives anyone any new information. For an informal discussion, see [1, Section 1.1].

2 Modeling the world

We begin by fixing two type variables: *'color* and *'person*. The puzzle doesn't specify how many eye colors are possible, but four are mentioned. Crucially, we must assume they are distinct. We specify the existence of colors other than blue and brown, even though we don't mention them later, because when blue and brown are the only possible colors, the puzzle has a different solution — the brown-eyed logicians may leave one day after the blue-eyed ones.

We refrain from specifying the exact population of the island, choosing to only assume it is finite and denote a specific person as the Guru.

We could also model the Guru as an outside entity instead of a participant. This doesn't change the answer and results in a slightly simpler proof, but is less faithful to the problem statement.

```
context
  fixes blue brown green red :: 'color
  assumes colors-distinct: distinct [blue, brown, green, red]

  fixes guru :: 'person
  assumes finite (UNIV :: 'person set)
begin
```

It's slightly tricky to formalize the behavior of perfect logicians. The representation we use is centered around the type of a *world*, which describes the entire state of the environment. In our case, it's a function *'person* ⇒ *'color* that assigns an eye color to everyone.[1]

The only condition known to everyone and not dependent on the observer is Guru's declaration:

```
definition valid :: ('person ⇒ 'color) ⇒ bool where
  valid w ⟷ (∃ p. p ≠ guru ∧ w p = blue)
```

We then define the function *possible n p w w'*, which returns *True* if on day *n* the potential world *w'* is plausible from the perspective of person *p*, based on the observations they made in the actual world *w*.

Then, *leaves n p w* is *True* if *p* is able to unambiguously deduce the color of their own eyes, i.e. if it is the same in all possible worlds. Note that if *p* actually left many moons ago, this function still returns *True*.

```
fun leaves :: nat ⇒ 'person ⇒ ('person ⇒ 'color) ⇒ bool
  and possible :: nat ⇒ 'person ⇒ ('person ⇒ 'color) ⇒ ('person ⇒ 'color) ⇒ bool
  where
    leaves n p w = (∀ w'. possible n p w w' ⟶ w' p = w p) |
    possible n p w w' ⟷ valid w ∧ valid w'
    ∧ (∀ p' ≠ p. w p' = w' p')
    ∧ (∀ n' < n. ∀ p'. leaves n' p' w = leaves n' p' w')
```

Naturally, the act of someone leaving can be observed by others, thus the two definitions are mutually recursive. As such, we need to instruct the simplifier to not unfold these definitions endlessly.

```
declare possible.simps[simp del] leaves.simps[simp del]
```

A world is possible if

1. The Guru's declaration holds.
2. The eye color of everyone but the observer matches.
3. The same people left on each of the previous days.

[1] We would introduce a type synonym, but at the time of writing Isabelle doesn't support including type variables fixed by a locale in a type synonym.

Moreover, we require that the actual world w is *valid*, so that the relation is symmetric:

lemma *possible-sym*: *possible n p w w′ = possible n p w′ w*
⟨*proof*⟩

In fact, *possible n p* is an equivalence relation:

lemma *possible-refl*: *valid w ⟹ possible n p w w*
⟨*proof*⟩

lemma *possible-trans*: *possible n p w1 w2 ⟹ possible n p w2 w3 ⟹ possible n p w1 w3*
⟨*proof*⟩

3 Eye colors other than blue

Since there is no way to distinguish between the colors other than blue, only the blue-eyed people will ever leave. To formalize this notion, we define a function that takes a world and replaces the eye color of a specified person. The original color is specified too, so that the transformation composes nicely with the recursive hypothetical worlds of *local.possible*.

definition *try-swap* :: *'person ⇒ 'color ⇒ 'color ⇒ ('person ⇒ 'color) ⇒ ('person ⇒ 'color)* **where**
try-swap p c₁ c₂ w x = (if c₁ = blue ∨ c₂ = blue ∨ x ≠ p then w x else transpose c₁ c₂ (w x))

lemma *try-swap-valid*[*simp*]: *valid (try-swap p c₁ c₂ w) = valid w*
⟨*proof*⟩

lemma *try-swap-eq*[*simp*]: *try-swap p c₁ c₂ w x = try-swap p c₁ c₂ w′ x ⟷ w x = w′ x*
⟨*proof*⟩

lemma *try-swap-inv*[*simp*]: *try-swap p c₁ c₂ (try-swap p c₁ c₂ w) = w*
⟨*proof*⟩

lemma *leaves-try-swap*[*simp*]:
assumes *valid w*
shows *leaves n p (try-swap p′ c₁ c₂ w) = leaves n p w*
⟨*proof*⟩

This lets us prove that only blue-eyed people will ever leave the island.

proposition *only-blue-eyes-leave*:
assumes *leaves n p w* **and** *valid w*
shows *w p = blue*
⟨*proof*⟩

4 The blue-eyed logicians

We will now consider the behavior of the logicians with blue eyes. First, some simple lemmas. Reasoning about set cardinalities often requires considering infinite sets separately. Usefully, all sets of people are finite by assumption.

lemma *people-finite*[*simp*]: *finite (S::'person set)*
⟨*proof*⟩

Secondly, we prove a destruction rule for *local.possible*. It is strictly weaker than the definition, but thanks to the simpler form, it's easier to guide the automation with it.

lemma *possibleD-colors*:
assumes *possible n p w w′* **and** *p′ ≠ p*
shows *w′ p′ = w p′*

⟨*proof*⟩

A central concept in the reasoning is the set of blue-eyed people someone can see.

definition *blues-seen* :: (*'person* ⇒ *'color*) ⇒ *'person* ⇒ *'person set* **where**
blues-seen w p = {*p'*. *w p'* = *blue*} − {*p*}

lemma *blues-seen-others*:
assumes *w p'* = *blue* **and** *p* ≠ *p'*
shows *w p* = *blue* ⟹ *card* (*blues-seen w p*) = *card* (*blues-seen w p'*)
and *w p* ≠ *blue* ⟹ *card* (*blues-seen w p*) = *Suc* (*card* (*blues-seen w p'*))
⟨*proof*⟩

lemma *blues-seen-same*[*simp*]:
assumes *possible n p w w'*
shows *blues-seen w' p* = *blues-seen w p*
⟨*proof*⟩

lemma *possible-blues-seen*:
assumes *possible n p w w'*
assumes *w p'* = *blue* **and** *p* ≠ *p'*
shows *w' p* = *blue* ⟹ *card* (*blues-seen w p*) = *card* (*blues-seen w' p'*)
and *w' p* ≠ *blue* ⟹ *card* (*blues-seen w p*) = *Suc* (*card* (*blues-seen w' p'*))
⟨*proof*⟩

Finally, the crux of the solution. We proceed by strong induction.

lemma *blue-leaves*:
assumes *w p* = *blue* **and** *valid w*
and *guru*: *w guru* ≠ *blue*
shows *leaves n p w* ⟷ *n* ≥ *card* (*blues-seen w p*)
⟨*proof*⟩

This can be combined into a theorem that describes the behavior of the logicians based on the objective count of blue-eyed people, and not the count by a specific person. The xkcd puzzle is the instance where $n = 99$.

theorem *blue-eyes*:
assumes *card* {*p*. *w p* = *blue*} = *Suc n* **and** *valid w* **and** *w guru* ≠ *blue*
shows *leaves k p w* ⟷ *w p* = *blue* ∧ *k* ≥ *n*
⟨*proof*⟩

end

5 Future work

After completing this formalization, I have been made aware of epistemic logic. The *possible worlds* model in section 2 turns out to be quite similar to the usual semantics of this logic. It might be interesting to solve this puzzle within the axiom system of epistemic logic, without explicit reasoning about possible worlds.

References

[1] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

[2] Randall Munroe. Blue eyes — a logic puzzle. URL: https://xkcd.com/blue_eyes.html.