MATHEMATICAL SHORTCOMINGS IN A SIMULATED UNIVERSE

SAMUEL ALEXANDER

Abstract. I present an argument that for any computer-simulated civilization we design, the mathematical knowledge recorded by that civilization has one of two limitations. It is untrustworthy, or it is weaker than our own mathematical knowledge. This is paradoxical because it seems that nothing prevents us from building in all sorts of advantages for the inhabitants of said simulation.

Published in *The Reasoner* **12** (9), pp. 71–72, 2018.

Imagine we were to program a simulation of a human-like civilization. Imagine the simulation were sophisticated enough that its inhabitants began stating mathematical axioms. Imagine we were to contrive that a monolith descends to earth inside the simulation. A monolith equipped with a keyboard and with instructions. The instructions implore them to use the keyboard to record mathematical axioms into the monolith. They are instructed to always continue recording the strongest mathematical axioms they can think of (in, say, the language of set theory). And they are warned never to record any mathematical falsehood into the monolith.

Having designed this simulation, we would know its source-code. From that source-code, we could infer a code for the list L of all axioms which are ever recorded into the monolith.

Could we trust the statements in L to be true? A priori, the answer depends on the simulation. At one end of the spectrum, the inhabitants might be afraid of the monolith and never record anything: their assertions would vacuously be true. On the other extreme, they might immediately record "$1 = 0$", just to see what happens.

If we could trust the truth of the inhabitants' statements, then we would necessarily be smarter than those inhabitants. We would be smarter than them because we would know a mathematical truth Φ which dominates all the axioms they ever record: namely,

$$\Phi = \text{“all statements in } L \text{ are true”}$$

(there are technical problems with this, which I address later in this paper).

Here's a paradox. Suppose we engineer the simulation in such a way that its inhabitants have vast motivation and resources to do mathematics. Let your imagination run wild (limited only by Turing computability).

- We could program the inhabitants to have huge brains.
- We could endow their world with vast energy reserves.
- We could program the world to constantly encourage the recording of strong mathematical axioms. For example, we could embed encouraging messages into the simulation, like in [2].
- We could engineer the simulation to discourage the recording of falsehoods. For example, maybe the monolith is preceded by fake monoliths which punish people for asserting pre-selected falsehoods (we would have to pre-select some falsehoods because the full set of falsehoods is non-computable).

In spite of all such contrivances, the inhabitants' recorded mathematical knowledge would fall short in at least one of two ways: it is untrustworthy, or it is weaker than our own.

E-mail address: samuelallenalexander@gmail.com.

This seems paradoxical because we ourselves do not have such huge brains, such vast energy reserves, or such great motivation.

Now I will address three technical problems.

- (Impracticality) Running the simulation might be prohibitively expensive. This is irrelevant: we do not need to actually run the simulation to reason about what would happen if we did run it. We only need to know its source-code. (Since it is a computer simulation, it is deterministic.)
- (Undefinability of truth) Even if we know Φ, we might not be able to *state* it in the language of set theory (which we would need to do in order to meaningfully contrast Φ against the statements in L). We can avoid this problem by using codes for computable ordinals. Computable ordinals have the property that there is a computable function f such that, given a code n for a computably enumerable list of codes of computable ordinals $\alpha_1, \alpha_2, \ldots$, $f(n)$ is a code for a computable ordinal α greater than all the α_i. Having the source-code for the simulation, we could infer a code n for the list of all numbers k such that the statement "k is a code of a computable ordinal" is implied by the inhabitants' axioms. If we know the inhabitants do not assert false axioms, we can infer $f(n)$ is a code of a computable ordinal. Thus, we know a computable ordinal bigger than all the computable ordinals the inhabitants know.
- (Self-reference) Perhaps, by diagonalization and self-reference, L already contains a statement equivalent to Φ. This objection can be defeated by the same computable ordinal trick used above.

(The above computable ordinal trick is reminiscent of how ordinals were used in [1] to shed light on the Lucas-Penrose argument.)

The more incentives and resources we give them to record strong axioms, the more likely the inhabitants will accidentally record a falsehood. The more we discourage them from recording falsehoods, the more afraid they will be to record strong axioms. Our argument suggests this dilemma is irreconcilable. We are incapable of designing computable incentives and resources strong enough to simulate a civilization smarter than us while balancing those with computable disincentives strong enough to ensure its trustworthiness.

The above reasoning makes us wary of consensus in our own world. We acknowledge that false axioms are occasionally proposed, and this is okay because over time they are corrected. But suppose we decide that once per millenium, we shall single out one new axiom we're really really sure of, never to be revised. Knowing the risk, we take many precautions to avoid choosing an error. Whatever incentives we impose on ourselves to avoid error, without crippling the strength of the chosen axioms, such safeguards apparently must fail: if they could succeed for us, then why not for the people we simulate?

References

[1] I.J. Good, 1969. "Gödel's Theorem is a Red Herring." The British Journal for the Philosophy of Science **19** (4), 357–358.

[2] Steven Hsu and Anthony Zee, 2006. "Message in the Sky." Modern Physics Letters A **21** (19), 1495–1500. https://arxiv.org/pdf/physics/0510102.pdf