Exploring the Possibility of a Private Language in a Thought Experiment: a ZKP Approach

J.L.

July 1, 2024

Abstract

This paper presents a theoretical examination of Ludwig Wittgenstein's Private Language Argument through the lens of a hypothetical artificial intelligence scenario. We propose a thought experiment where an AI, referred to as x, uses a private language to process instructions that, even after a reset of its memory, allows it to consistently execute tasks based on previously written notes in its private language. The implications of this experiment challenge traditional notions of language as inherently public and communicable, suggesting that under certain conditions, an operational private language might be possible within artificial systems.

1 Introduction

Ludwig Wittgenstein famously argued in his later works that the notion of a private language, a language that only one individual can understand, is conceptually incoherent because language inherently requires public criteria for its words. This paper revisits Wittgensteins Private Language Argument (PLA) by proposing a thought experiment involving an artificial intelligence that supposedly operates with a private language.

2 Background

Wittgensteins Private Language Argument suggests that no language can be inherently private because the meanings of words and symbols need external verification that transcends individual experience. This principle has been widely accepted in philosophical circles; however, the rise of artificial intelligence and advanced cryptographic systems, such as Zero-Knowledge Proofs (ZKPs), pose new questions about the bounds of this argument.

3 The Thought Experiment

The proposed thought experiment involves an AI named x, which can record instructions in a private language following the observation of a randomized event (a coin flip). After recording these instructions and undergoing a memory reset, the AI is still able to follow the instructions accurately, suggesting a consistent internal rule or grammar that is not accessible or understandable to anyone other than x.

3.1 Experimental Setup

The experimental setup is designed to test the operational capability of a private language developed by an artificial intelligence, x (the prover), which we hypothesize to be comprehensible solely by x itself. The procedure of the experiment is conducted in independent trials, detailed as follows:

- **Observation Phase:** In each trial, indexed by i, a random event, a coin flip, is generated, resulting in an outcome o_i which can be either head (H) or tail (T). This outcome is shown to the prover x and the verifier y.
- **Recording Phase:** Upon observing o_i, x is tasked to write down a note of instructions n_i in its private language. These instructions are intended to direct x to later replicate the outcome o_i by manipulating a coin on a table. The verifier y is unable to read, and not allowed to read n_i (for example, encrypt n_i using methods such and one-time pad).
- **Reset Phase:** After x records the instructions, its memory is reset to ensure that it retains no memory of o_i or n_i from the previous phase. The reset does not erase the private language ability of x. Likewise, the coin used for the initial flip is also reset to a random state on the table, removing any potential biases or marks.
- **Execution Phase:** x is then presented with the note n_i, which contains the private instructions that no one else can understand. x tries to execute the instructions in n_i.
- **Outcome Verification:** Whether x successfully executed the instructions, or misread the instructions, or completely failed to follow the instructions, the coin on the table would have a final state H or T, denoted as e_i, and it is observed and recorded by the y, whom can simply check whether $e_i = o_i$.
- **Repetition and Statistical Analysis:** This process is repeated across multiple independent trials m times. The consistency of x's ability to match

e_i with o_i is statistically analyzed. If x can successfully manipulate the coin to match the original flip outcome with high reliability, specifically with a confidence level of $1 - \frac{1}{2^m}$, it substantiates the claim that x operates using a coherent, yet private language.

The implications of x's consistent performance under these conditions could profoundly challenge existing philosophical notions about the nature of language and cognition, particularly in the context of Wittgenstein's Private Language Argument—a private language can exists, and is verifiable.

3.2 Implications of Results

If x can successfully follow its own instructions across multiple independent trials, it suggests the existence of a systematic, consistent private language. This operational definition of a private language challenges Wittgenstein's argument by demonstrating a functional, albeit non-human, language system. Note, by observing the output e_i from x, y gained no additional information about the private language, because $e_i = o_i$ is the information that y already knew since the start of the trial. Zero knowledge about the private language is leaked in this process, thus the private language will remain private and undecryptable forever.

4 Discussion

While the experiment suggests an AI can develop a form of private language, it raises questions about the definition of language itself. Is a private sequence of 1's and 0's equivalent to a language? Moreover, this thought experiment underscores the distinction between human and artificial cognition, opening up discussions on the limits and capabilities of AI in understanding and processing information.

5 Conclusion

This thought experiment not only challenges Wittgenstein's Private Language Argument but also enhances our understanding of how artificial intelligence might navigate and develop unique systems of communication. The implications for both philosophy of language and artificial intelligence are profound, suggesting new areas of study in the intersection of technology and humanistic inquiry.