Accepted for *Proceedings of the International Ludwig Wittgenstein Symposium 2021*

Does Artificial Intelligence Use Private Language?

Ryan Michael Miller[1][0000-0003-0268-2570]

[1] Philosophy Department, University of Geneva, Genève 4, Switzerland
Ryan.Miller@unige.ch

Abstract. Wittgenstein's Private Language Argument holds that language requires rule-following, rule following requires the possibility of error, error is precluded in pure introspection, and inner mental life is known only by pure introspection, thus language cannot exist entirely within inner mental life. Fodor defends his *Language of Thought* program against the Private Language Argument with a dilemma: either privacy is so narrow that internal mental life can be known outside of introspection, or so broad that computer language serves as a counter-example. I suggest that the developing field of artificial intelligence (deep learning neural networks) tends to vitiate Fodor's defense and hence vindicate the Private Language Argument. The first horn of Fodor's dilemma requires language to encompass genuinely internal mental life, i.e. non-projected intentional states, which are not exhibited in classical machine learning but only by deep learning neural networks (artificial intelligence). Such networks act as black boxes, however, whose state cannot be understood by tracking the changes in their supervenience bases without shared context, and that shared context introduces the possibility of error. The language of artificial intelligence is not private.

Keywords: Private Language Argument, Language of Thought, Artificial Intelligence, Machine Learning

1 Introduction

The Private Language Argument (PLA) found at *Philosophical Investigations* §243-271 is perhaps Ludwig Wittgenstein [1]'s most lasting contribution to philosophy. Maximalist readers of the argument like Saul Kripke [2] and John McDowell [3] hold that the PLA requires a wide-ranging reappraisal of philosophy of mind. Nonetheless, other philosophers of mind like Jerry Fodor [4] are very dismissive of the generality and force of the PLA. While Fodor [4]'s remarks are brief—and I will argue later based on a misreading of Wittgenstein—Fodor [5] offers a more interesting rejoinder based on the existence of computer languages. I show that Fodor's argument requires a setting of artificial intelligence (AI)—i.e., deep learning neural networks—rather than ordinary software, but that what we now know about AI fails to validate Fodor's argument. The language of AI is not private, and so Fodor's counter-example to the PLA fails. The plan of the paper is as follows. First, introduce a formal version of the much-contested Private Language Argument, show why Fodor's *Language of Thought* program is in tension with it, and review Fodor's responses to the PLA. Second, investigate Fodor [5]'s computation counter-example to the PLA in detail, and show that it requires an AI setting, rather than Fodor's more general framing. Finally, show that Fodor's counter-example fails given what we now know about AI, leaving the PLA untouched.

2 Whither Private Language?

Wittgenstein's argument against the possibility of private language in the *Philosophical Investigations* is much contested, with critics failing to agree even on which section of the text is supposed to present the argument, and Kripke [2]'s clear and influential presentation of the Private Language Argument derided as a mind-meld called "Kripkenstein" rather than accurate exegesis of Wittgenstein [6]. Given the length constraints here and the focus on AI rather than human language, I will avoid detailed exegesis and instead rely on Harris [7]'s more modest reconstruction, which goes as follows:

P1 (LANGUAGE): language → rule-following (language is used only if rules are followed)

P2 (NORMATIVITY): rule-following → ◊ systematic error (rule-following implies the possibility of systematic error)

P3 (SUBJECTIVITY): ◊ systematic error → ¬ rules known wholly by introspection (the possibility of systematic error precludes epistemic access outside pure introspection)

P4 (PRIVACY): ¬ rules known by pure introspection → ¬ wholly grounded by internal mental life (epistemic access outside pure introspection precludes internal mental grounds, i.e. privacy)

C (PLA): language → ¬ grounded wholly by internal mental life (language is never private)

The argument is deductively valid by conditionalization, so disputing the conclusion requires disputing the truth of one or more of the premises.

Fodor [4,5]'s influential *Language of Thought* program is in tension with the PLA because Fodor argues as follows:

P1 (COGSCI): cognitive science [is true] (i.e., its leading theories about the human mind are correct)

P2 (COGCOMP): cognitive science → computation (cognitive science tells us that the human mind is computational in nature)

P3 (COMPLANG): computation → $\text{language}_{\text{thought}}$ (computation requires a language of thought)

C (LOT): [there is a] $\text{language}_{\text{thought}}$ (there is a language of thought)

Here the conclusion follows by two applications of *modus ponens*. But Fodor also holds:

P4 ($^{\text{NOT}}\text{NORMATIVITY}_{\text{THOUGHT}}$): $\text{language}_{\text{thought}}$ → ¬ ◊ systematic error (a language of thought is incapable of systematic error)

If Fodor were to grant NORMATIVITY and LANGUAGE then by two applications of *modus tollens* his "language of thought" would cease to be a language at all, in contradiction to the spirit of COMPLANG. Fodor therefore takes his *Language of Thought* program as a *reductio ad absurdum* of one of the first two premises of the Private Language Argument. In *LOT2: The Language of Thought Revisited* [4], Fodor rejects LANGUAGE.[1] Here Fodor follows Rush Rhees [8]'s reconstruction of the PLA, where the requirement of rule-following for language use is taken to follow from the further hidden premises:

(LANGPURP): language → [learned ∨ communicative] (language is always either learned or used for communication)

(RULEREQ): [learned ∨ communicative] → rule-following (learning and use for communications both require following rules)

Fodor takes the language of thought as a counter-example to LANGPURP, since it is supposed to be innate and internal, leaving LANGUAGE unmotivated. The trouble with Fodor's defense here is twofold. First, Rhees gives no textual evidence for his contention that LANGUAGE is motivated by LANGPURP and RULEREQ, and many influential interpreters of the PLA like McDowell [3] seem to hold it on different grounds that would apply to a language of thought. Second, Fodor insists in both [4] and [5] that the language of thought always refers determinately, e.g. the concept *RABBIT* always determinately refers to (all and only) rabbits. This kind of determinate reference seems suspiciously like rule-following, since if *RABBIT* sometimes referred to dogs, then

[1] In another place in [4], Fodor instead distinguishes LANGUAGE, offering that while language *use* might imply rule-following, the language of thought "though it is a system of representations, isn't a system of representations that anybody uses, correctly or otherwise. One doesn't use thoughts, one just has them. Having thoughts isn't something that you do; it's something that happens to you." The mere having of thoughts, however, is also rule-following in Wittgenstein's sense because according to Fodor those thoughts are supposed to have determinate reference. That determination of reference motivates Wittgenstein's argument, not an action theory principle of intentionality or instrumentality. In [5] Fodor admitted as much, granting that for Wittgenstein a private language is merely a "language for the applicability of whose terms there exist no public criteria" and consequently allowing that "though nothing requires that the language of thought should be construed as a sense datum language, it may seem, nevertheless, to fall in the scope of Wittgenstein's argument and thus to be in peril of that argument being a good one."

RABBIT would fail to follow the rule that concepts in the language of thought always have determinate reference. The PLA therefore deserves a more serious rejoinder than the mere "snark" and brief side notes Fodor [4] grants it. Interestingly, however, Fodor [5] makes a more sustained and serious argument against NORMATIVITY, rather than LANGUAGE, using a detailed analysis of the nature of computation, which he ignores in the later volume. It is to this earlier argument against the claim that rule-following implies the possibility of systematic error that I now turn.

3 Fodor's Dilemma

While Fodor's argument in both *Language of Thought* volumes is about human concept use, his belief that human thinking is computational leads him to an analogy with computer languages. Here Fodor [5] offers a constructive dilemma for the partisan of the Private Language Argument: computers use language for computation (i.e., software programs are executed in programming languages) so that language use either is or is not grounded in the internal mental life of the computer. In the first horn of the dilemma, grounding computer language use in the internal mental life of the computer leads to a *reductio ad absurdum* argument against NORMATIVITY. In the second horn of the dilemma, grounding computer language use outside the internal mental life of the computer leads to an argument that no mental phenomena are ever private, whether linguistic or otherwise, so the Private Language Argument is trivially satisfied and utterly irrelevant. I review the prospects for each horn of the dilemma in turn.

For the first horn of the dilemma Fodor—unlike in his later volume—begins with a justification for the claim that computer language use *is* rule-following:

P1 (LANGUAGE$_{\text{COMP}}$): language$_{\text{computer}}$ → rule-following

Computer languages follow rules, he says, because their "use comports with the conditions specified in the representation in an appropriate meta-language" [5]. We might think of this as the process of formal software verification, where a program is checked for correctness by a provable relationship between its inputs and outputs in a formal meta-language used by the verifier [9]. To this universally instantiated version of LANGUAGE, Fodor adds premises 2-4 of Wittgenstein's original Private Language Argument:

P2 (NORMATIVITY): rule-following → ◊ systematic error

P3 (SUBJECTIVITY): ◊ systematic error → ¬ rules known wholly by introspection

P4 (PRIVACY): ¬ rules known by pure introspection → ¬ wholly grounded by internal mental life

Next comes the premise from the first horn of the constructive dilemma, that computer language use *is*, in contradiction to the conclusion of the Private Language Argument, grounded wholly by the internal mental life of the computer:

P5 ($^{\text{NOT}}$PLA$_{\text{COMP}}$): language$_{\text{computer}}$ → grounded wholly by internal mental life

Now by twice over conditionalization and *modus tollens*,

C1 ($^{\text{NOT}}$NORMATIVITY$_{\text{COMP}}$): language$_{\text{computer}}$ → ¬ ◊ systematic error

C2 ($^{\text{NOT}}$LANGUAGE$_{\text{COMP}}$): language$_{\text{computer}}$ → ¬ rule-following

$\perp$

Since the conclusion $^{\text{NOT}}$LANGUAGE$_{\text{COMP}}$ contradicts the first premise LANGUAGE$_{\text{COMP}}$, we have a *reductio ad absurdum* of the set of premises. As the first premise $^{\text{NOT}}$LANGUAGE$_{\text{COMP}}$ is taken as an empirical fact and the fifth premise $^{\text{NOT}}$PLA$_{\text{COMP}}$ is merely assumed by constructive dilemma, one of the core premises of the Private Language Argument (NORMATIVITY, SUBJECTIVITY, or PRIVACY) must go, and Fodor thinks that NORMATIVITY is the most dubious. In any case, the result of taking the first horn of Fodor's dilemma is that the PLA is unsound, and thus need not worry advocates of the *Language of Thought* program.

Naturally, defenders of the Private Language Argument will wish to avoid this result by exploring the other horn of the constructive dilemma offered by Fodor, replacing $^{\text{NOT}}$PLA$_{\text{COMP}}$ with:

(PLA$_{\text{COMP}}$): language$_{\text{computer}}$ → ¬ grounded wholly by internal mental life (there can be epistemic access to computer language use outside of the internal mental life of the computer, i.e. publicly)

Fodor, however, presses defenders of the PLA by asking what could motivate PLA$_{\text{COMP}}$. After all—again following Rhees [8]'s analysis—Wittgenstein's examples of the public nature of language come from LANGPURP: language is public when it is learned or used to communicate. Computer languages, though—like the language of thought—are neither learned by the computer nor used by the computer to communicate its output. Rather, they are fundamentally computational. If computer languages are not learned or communicative, where would public epistemic access to their use arise? Fodor suggests that the only plausible answer is that computer language use, i.e. software state, supervenes on the hardware state, which is externally accessible by snooping on voltage changes in its circuits. This method of external snooping is the basis for microarchitectural timing attacks that reveal cryptographic secrets and other private data from computer systems without access to the input or output stream of the software handling the sensitive information [10]. Thus everything privately grounded in the internal mental life of the computer, i.e. the state of its running software, is also publicly grounded outside the internal mental life of the computer, i.e. in the state of its hardware, because the hardware state is the ground of the software state. Computer language use is thus public for reasons independent of the NORMATIVITY premise which Fodor finds problematic.

The problem for the Private Language Argument, according to Fodor, is that this argument generalizes, and thus proves too much. According to the widely-held thesis of physicalism [11–15], *all* mentality supervenes on physical "hardware" and is thus public, in a way totally independent of any theses about language. The PLA is thus *trivially* true and so irrelevant as a criticism of the *Language of Thought* program. On the first horn of Fodor's dilemma, computers are private like minds, so if computers can use language and follow rules without the possibility of error, so can humans. On the second horn of Fodor's dilemma, minds are public like computers, so if computers can use language and follow rules without the possibility of error, so can humans. Either way, the mere existence of computers executing programming languages is supposed to disarm the Private Language Argument and make way for the *Language of Thought* program.

For Fodor's constructive dilemma to work, though, the thesis PLA$_{\text{COMP}}$ must have some propositional content which can be meaningfully affirmed or negated. Otherwise the premise is meaningless and the argument is unsound in any logic which takes content seriously, e.g. [16,17]. Fodor merely takes it as obvious that computers use language and assumes that they do so in a way that involves internal mental life. *Pace* Fodor, however, such internal mentality is not trivially achievable for computer systems. Mark Ressler [18] argues that software states must be both *intentional* and *non-projected* in order to meet this criterion. Intentionality is required because unless software states are *about* something, then they cannot support reference, let alone the determinate reference that Fodor thinks is characteristic of computer language use. Non-projection is required because those intentional states have to be relevantly *internal*. If the states are merely the result of projection by an outside agent, they will be public from their inception.

Fodor is right that computer language is not acquired or communicative, but it is generally public in a properly linguistic way that has nothing to do with hardware: it is *programmed*. Ordinarily, the mental life of a computer is merely the state resulting from its programming—the intentionality is shared with the programmer. This explicit shared intentionality is now called "symbolic" or "old-fashioned" artificial intelligence [19]. Computers using such systems fail to support Fodor's dilemma because their language is not grounded wholly by their internal mental life—they validate PLA$_{\text{COMP}}$—but for a substantive reason not shared with human intelligences, which are not explicitly programmed. If such computer systems were the only examples available, the second horn of Fodor's dilemma would lose its bite. Only a computer with non-projected intentional states can support Fodor's dilemma, but those are the domain of deep learning neural networks, the modern approach to artificial intelligence. As the name suggests, such computer systems learn their concepts rather than having them supplied by the programmer [20], so their mental life can be meaningfully internal. It is to such deep learning neural networks—artificial intelligence properly so-called—that our analysis of Fodor's argument must now turn.

4 Artificial Intelligence and Private Language

Deep learning neural networks are frequently referred to as "black boxes." A face-classifier, for instance, learns to identify faces by building hierarchies of features from its training data, but those features do not necessarily correspond to any human concept like *NOSE*, *EAR*, or *EYE* [21,22]. The intermediate elements of the trained model are thus intentional—they are about facial features—but they are also non-projected, and so not trivially public.[2] If such an artificial intelligence were to validate PLA$_{\text{COMP}}$, it would presumably have to be for the reason

[2] The intermediate elements may also be learned from another AI, as in [23,24], but this validates LANGPURP, so Fodor would classify it as non-trivially public.

Fodor gives: because its software state could be known by inspecting the hardware state on which it supervenes. Precisely because deep learning neural networks are black boxes, however, they fail to validate PLA_{COMP}. Inspecting the hardware state of such an AI cannot reveal its language use since the snooper has no conceptual access to the state thus revealed, Absent such conceptual access, hardware state is just hardware state,[3] with no evident intentionality. The snooper cannot say what the revealed hardware state is *about* without having a concept for that state. Thus a pure black box AI, while capable of non-projected intentionality and hence internal mental life, does not validate PLA_{COMP} via a trivializing supervenience of software on hardware.

Tracking the state of a deep learning neural network, such that its language use could be publicly analyzed and correlated with its hardware state, requires introducing shared context [22,25]. For example, an AI could be rewarded for correctly identifying noses and ears, or for matching pictures of the same person's eyes rather than only their entire face. The AI could also be required to output not only the face classification, but also the feature classification in human concepts. In other words, whatever state the snooper wants to track at the hardware level must be tagged at the software level, so that the snooper and AI are operating on a shared conceptual basis. Once those states have been tagged, the snooper can know what concept the AI is representing with a particular hardware state, and so validate PLA_{COMP} via a trivializing supervenience of software on hardware.

The trouble for Fodor is that once shared context has been reintroduced to an artificial intelligence, the possibility of systematic error is similarly reintroduced. The whole point of Fodor's dilemma was to cast doubt on NORMATIVITY given Fodor's commitment to $^{\text{NOT}}\text{NORMATIVITY}_{\text{THOUGHT}}$. Computers' ability to follow rules and use public language without the possibility of systematic error was supposed to do the work in this argument. AIs with public concepts, however, are vulnerable to adversarial inputs [26,27]. For example, a face classifying AI could be given input photos with slight mis-coloration, and then systematically mis-classify the locations of the eyes and ears of the people pictured in the photos. Because the AI's concept *EYE* is supposed to determinately refer to eyes and not ears, we can say that the AI systematically errs when given this adversarial input. Artificial intelligences sophisticated enough to validate PLA_{COMP} thus also validate NORMATIVITY.

5 Conclusion

Fodor cannot win with his constructive dilemma against the Private Language Argument based on computational examples. Computational systems simple enough to plausibly follow rules without the possibility of error use languages which are public in virtue of the intentionality of their programmers. Artificially intelligent computational systems which are complex enough to exhibit their own non-projected intentionality, meanwhile, introduce the possibility of systematic error in their attempts at rule-following. Shared concepts can be used correctly or incorrectly and subject to adversarial attacks creating systematic errors. Either way, computer language use is public for non-trivial reasons, so Fodor's dilemma fails to rebut Wittgenstein's argument against the possibility of private language. The *Language of Thought* program must reckon with how its concepts are shared and how systematic errors are possible.

6 References

[1] Wittgenstein L. Philosophical investigations. 3rd ed. Oxford: Basil Blackwell; 1968.

[2] Kripke SA. Wittgenstein on Rules and Private Language. Harvard University Press; 1982.

[3] McDowell J. Mind and world: with a new introduction. 1st Harvard University Press Paperback Ed. Cambridge, MA: Harvard University Press; 1996.

[4] Fodor JA. LOT 2: The Language of Thought Revisited. Oxford: Oxford University Press; 2010.

[5] Fodor JA. The Language of Thought. New York: Thomas Crowell; 1975.

[6] Goldfarb W. Kripke on Wittgenstein on Rules. J Philos 1985;82:471–88. https://doi.org/10.2307/2026277.

[7] Harris R. The Private Language Argument Isn't as Difficult, Nor as Dubious as Some Make Out. Sorites 2007;18:98–108.

[8] Rhees R. Symposium: Can There Be a Private Language? Proc Aristot Soc Suppl Vol 1954;28:63–76.

[9] D'Silva V, Kroening D, Weissenbacher G. A Survey of Automated Techniques for Formal Software Verification. IEEE Trans Comput-Aided Des Integr Circuits Syst 2008;27:1165–78. https://doi.org/10.1109/TCAD.2008.923410.

[3] Such snooping may reveal elements of lower-level software state such as the instruction being executed, but these instructions are programmed, not learned, and so their intentionality is projected by the programmer—they are not part of the AI proper. Attacks at the model level require public concepts, as in [24].

[10] Ge Q, Yarom Y, Cock D, Heiser G. A survey of microarchitectural timing attacks and countermeasures on contemporary hardware. J Cryptogr Eng 2018;8:1–27. https://doi.org/10.1007/s13389-016-0141-6.

[11] Montero B, Papineau D. A defence of the via negativa argument for physicalism. Analysis 2005;65:233–7. https://doi.org/10.1111/j.1467-8284.2005.00556.x.

[12] Kim J. Physicalism, Or Something Near Enough. Princeton University Press; 2008.

[13] Stoljar D. Physicalism. Taylor & Francis; 2009.

[14] Ney A. Microphysical Causation and the Case for Physicalism. Anal Philos 2016;57:141–64. https://doi.org/10.1111/phib.12082.

[15] Spurrett D. Physicalism as an empirical hypothesis. Synthese 2017;194:3347–60. https://doi.org/10.1007/s11229-015-0986-8.

[16] Yablo S. Aboutness. Princeton University Press; 2014. https://doi.org/10.2307/j.ctt2tt8rv.

[17] Fine K. Angellic Content. J Philos Log 2016;45:199–226. https://doi.org/10.1007/s10992-015-9371-9.

[18] Ressler M. Connectionism and the Intentionality of the Programmer. Thesis. San Diego State University, 2003.

[19] Flasiński M. Symbolic Artificial Intelligence. In: Flasiński M, editor. Introd. Artif. Intell., Cham: Springer International Publishing; 2016, p. 15–22. https://doi.org/10.1007/978-3-319-40022-8_2.

[20] Garnelo M, Shanahan M. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. Curr Opin Behav Sci 2019;29:17–23. https://doi.org/10.1016/j.cobeha.2018.12.010.

[21] Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). Proc. 35th Int. Conf. Mach. Learn., PMLR; 2018, p. 2668–77.

[22] López-Rubio E. Throwing light on black boxes: emergence of visual categories from deep learning. Synthese 2021;198:10021–41. https://doi.org/10.1007/s11229-020-02700-5.

[23] Pal S, Gupta Y, Shukla A, Kanade A, Shevade S, Ganapathy V. ActiveThief: Model Extraction Using Active Learning and Unannotated Public Data. Proc AAAI Conf Artif Intell 2020;34:865–72. https://doi.org/10.1609/aaai.v34i01.5432.

[24] Miura T, Hasegawa S, Shibahara T. MEGEX: Data-Free Model Extraction Attack against Gradient-Based Explainable AI. ArXiv[Cs] 2021. https://doi.org/10.48550/arXiv.2107.08909.

[25] Kazhdan D, Dimanov B, Jamnik M, Liò P, Weller A. Now You See Me (CME): Concept-based Model Extraction. ACM Int. Conf. Inf. Knowl. Manag., 2020. https://doi.org/10.48550/arXiv.2010.13233.

[26] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks 2014. https://doi.org/10.48550/arXiv.1312.6199.

[27] von Eschenbach WJ. Transparency and the Black Box Problem: Why We Do Not Trust AI. Philos Technol 2021. https://doi.org/10.1007/s13347-021-00477-0.