Learning as a Criterion: Why the Chinese Room Targets the Wrong Machine

Mehmet Mirioğlu

Abstract. Searle's Chinese Room argument concludes that no symbol-manipulating system could understand, because understanding requires intentionality and intentionality is biological. This paper argues that the argument fails because it targets a model Turing had already rejected. Section 7 of Turing's 1950 paper explicitly proposes a learning machine that builds its responses through experience rather than executing a fixed mapping. This paper takes learning as a necessary criterion for understanding and argues that once this criterion is introduced, the Chinese Room's central intuition does not transfer. The operator follows a rulebook he did not write; a learning system constructs its own responses from exposure, and in this respect it is no different from a human acquiring a language. A child learning Chinese does not begin with a rulebook either. The fading qualia argument further undermines the biological necessity claim: if neurons are replaced one by one with functionally equivalent components and behavior remains unchanged, the attribution of intentionality cannot be withheld on the basis of substrate alone.

1. INTRODUCTION

Turing's question "Can machines think?" was replaced with a behavioral test[1]. A machine passes the test if a human interrogator, communicating only through written exchange, cannot reliably distinguish it from a human. Turing's view was that "thinking" and "understanding" carry too much conceptual baggage to serve as clear criteria, and that behavioral indistinguishability is a more tractable target. However, Searle challenged this move. His Chinese Room thought experiment places a human operator inside the machine. The operator receives Chinese symbols as input, consults a rulebook that maps inputs to outputs, and returns the appropriate Chinese symbols[2]. From the outside, the performance is indistinguishable from that of a Chinese speaker. But the operator understands nothing. Searle's conclusion is that behavioral equivalence cannot be sufficient for understanding, because symbol manipulation of this kind carries no semantic content. The symbols, as he puts it, have syntax but no semantics.

This paper does not dispute the thought experiment on its own terms. What it disputes is the claim that the thought experiment addresses Turing's proposal. The system Searle describes is static: its responses are fixed in advance by a rulebook. Turing did not defend such a system. Section 7 of his 1950 paper proposes a learning machine, one whose responses are not pre-programmed but built through a process of education and experience. This distinction matters because the Chinese Room's central intuition depends on the static assumption. The paper argues that once learning is introduced as a criterion, the argument does not apply in the same way, and that this has implications for how the debate about machine understanding should be framed.

2. THE CHINESE ROOM AND ITS ASSUMPTION

Searle's argument rests on a specific picture of what a symbol-manipulating system does. The operator in the room has a rulebook that covers every possible Chinese input. He follows the

[1] Turing, 1950, p. 433.

[2] Searle, 1980, p. 417.

rules, returns the correct output, and understands nothing. Searle generalizes from this: any system that works by manipulating formal symbols on the basis of their shape, rather than their meaning, cannot understand what it processes. Understanding requires intentionality, and intentionality is a biological phenomenon that cannot be produced by formal symbol manipulation alone[3].The argument has been challenged in several ways. Copeland points out that the inference is logically invalid: from the fact that the operator does not understand Chinese, it does not follow that the system of which he is a part does not understand Chinese[4]. On the other hand, Dennett argues that the thought experiment relies on the reader's failure to imagine the actual complexity of a system capable of passing the Turing Test[5]. A system that could sustain the relevant conversation would require world knowledge, self-knowledge, and the capacity to track its interlocutor's responses in ways that go well beyond rule-following.These objections target different weaknesses in the argument. But neither questions the assumption that the system is static. The present argument targets that assumption, and in doing so it supports the direction both objections point toward. If the system learns its responses rather than executing a pre-written mapping, Copeland's claim that the system may understand becomes less speculative: there is now a mechanism that could ground it. And Dennett's intuition that complexity matters gets a more concrete basis: a static system remains static regardless of its size, and it is learning, not size alone, that produces the kind of complexity Dennett has in mind.

3. TURING'S LEARNING MACHINE

Turing raised a version of the static system problem himself. In section 7 of his 1950 paper, he asks why one should try to simulate an adult mind directly. His answer is that this approach is misconceived. Instead, he proposes simulating a child's mind and subjecting it to an appropriate process of education[6]. The resulting system would not have its responses fixed in advance. It would develop them through experience, in a way that the programmer could not fully predict. Thanks to this, Turing describes a learning machine whose rules change as a result of experience, and he notes explicitly that an important feature of such a machine is that its teacher will often be largely ignorant of what is going on inside it[7]. This is the opposite of the Chinese Room operator, who follows explicit instructions for every input. The operator is not ignorant of what is going on inside him; the rulebook tells him exactly what to do. A learning machine, by contrast, constructs its own responses in ways that cannot be fully anticipated by whoever designed it. Modern machine learning systems are closer to Turing's description than to Searle's target. A system trained on large amounts of data does not have a pre-written mapping from inputs to outputs. It constructs responses through a process of pattern extraction and generalization. Its behavior is not pre-determined by a rulebook; it emerges from training. The Chinese Room operator has no analogue in such a system. There is no human following instructions; there is a system that has learned, from exposure, how to respond.

4. WHY LEARNING?

The Chinese Room operator does two things: he receives symbols and he follows rules to produce output. Searle's point is that neither of these steps produces understanding. This much is convincing. But the argument leaves a question unanswered: what would be sufficient for understanding? Searle's answer is intentionality, and intentionality, he claims, is a

[3] Searle, 1980, p. 422.
[4] Copeland, 2015, pp. 142-147.
[5] Dennett, 1991, pp. 435-440.
[6] Turing, 1950, pp. 456-457
[7] Turing, 1950, p. 458

biological phenomenon. Biology is where the argument stops.

This is not a satisfying stopping point. To say that intentionality is biological is to say that biological systems have it. It does not say what biological systems do that produces it. And what biological systems do, in the relevant sense, is learn. A child acquiring Chinese does not begin with a rulebook. She encounters sounds and symbols repeatedly, in contexts that connect them to objects, actions, and responses from other people. The meaning of those symbols is not handed to her; it emerges from sustained exposure. By the time she understands Chinese, she has built, through experience, the very connections that the Chinese Room operator was given in advance and never built at all. This matters because it locates the relevant difference between the operator and a Chinese speaker. Searle's own description of the case implies that the operator fails because he does not know what the symbols refer to. But a Chinese speaker knows what the symbols refer to because she learned it. The difference is not substrate. It is the presence or absence of a learning process. If that is what distinguishes understanding from non-understanding in the human case, then learning is not an arbitrary criterion. It is the criterion that Searle's own contrast case relies on.

Searle's response to this is that intentionality requires causal powers that only biological systems possess[8]. Learning may be necessary, but biology does additional work that no formal system can replicate. This is where fading qualia becomes relevant[9]. Suppose neurons are replaced one by one with functionally equivalent artificial components. At each step, the system produces the same outputs and responds to the same inputs in the same ways. If intentionality is lost at some point in this process, there is no way to detect it, because the behavior remains unchanged. But intentionality is something we attribute to other systems on the basis of their behavior. We do this for other people, for animals, and in weaker forms for simpler organisms. If the behavioral evidence remains constant across the substitution, withholding the attribution from the resulting system requires a principled reason. Searle does not provide one. He asserts that the biological substrate is necessary but does not explain what it contributes that the functionally equivalent replacement does not.

The Chinese Room is a compelling objection to naive functionalism: the view that any system with the right input-output relations thereby understands. But not all functionalist positions are naive in this sense. A position that takes learning to be a necessary condition for understanding is not committed to the claim that static symbol manipulation is sufficient. Searle's argument does not engage with this kind of position. It targets a version of the view that more sophisticated functionalists do not hold.

Once learning is introduced as a criterion, the role of the biological substrate becomes harder to specify. If what biology contributes is the capacity to learn, then a system that learns is doing what biology does in the relevant respect. If biology contributes something beyond learning, Searle owes an account of what that is. The Chinese Room does not provide one. It shows that the operator, who has not learned anything, does not understand. It does not show what would happen if he had.

5. CONCLUSION

Searle's Chinese Room is a compelling argument against a specific kind of system: one that follows a pre-written rulebook and never learns. Turing did not defend such a system. Section 7 of his 1950 paper explicitly proposes a learning machine, one whose responses are not fixed in advance but built through experience. Modern AI systems are closer to this description than to the static

[8] Searle, 1980, p. 422.
[9] Chalmers, 1996, pp. 254-259.

system the Chinese Room targets. Adding learning as a criterion to the Turing Test does not resolve all questions about machine understanding. What it does is show that Searle's specific argument does not apply to the systems Turing had in mind. The operator in the Chinese Room has no counterpart in a system that has learned its behavior from structured exposure to input. The intuition that drives the thought experiment, that a system following rules cannot understand, does not transfer straightforwardly to a system that built those rules itself through learning. Whether such a system understands in any philosophically robust sense remains open. Searle's argument, directed at the wrong target, does not close that question.

REFERENCES

Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Copeland, B. J. (2015). *Artificial Intelligence: A Philosophical Introduction*. Wiley.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335-346.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.