What is Being Imitated? Revisiting the Idea of Intelligence in Turing Test

Deepa Shukla[1] and **Dr. George Kodimattam Joseph** [2]

Abstract. What did Turing mean when he suggested that the best strategy might be to "imitate something other than the behaviour of a man [6]"? Did he foresee that a perfect imitation of human intelligence would eventually feel like a parody? Or the fact that deception as a criterion of imitation is problematic? Despite convincing many users of their intelligence and even consciousness, contemporary AI systems miserably fail the imitation game. We argue that the very conditions of the imitation game are structurally ambiguous. While the game has often been interpreted as treating linguistic indistinguishability, which represents a machine's ability to imitate the conversational style of a normal human being, as a sufficient proxy for intelligence, it avoids any explicit definition of intelligence. To serve as an evaluation criterion, imitation still requires stable standards. In its current form, the test suffers from two major structural challenges. First, it lacks a determinate object of imitation, and secondly, it relies on inconsistent interrogators for its evaluation. These troubling constraints undermine stable standards for determining whether a machine has passed the Turing test. Contemporary AI systems expose these tensions with peculiar clarity. Even if we have systems that can imitate human conversational skills, they keep failing the Turing test. Thus, while the test remains philosophically influential even after seventy-five years of its introduction, methodologically it is unsettled. The question is no longer whether machines can pass the Turing test, but whether the Turing test itself provides a stable criterion for adjudicating the intelligence displayed by contemporary systems.

Keywords: intelligence, Turing test, imitation, artificial intelligence

1 INTRODUCTION

Alan Mathison Turing, in his seminal work 'Computing Machinery and Intelligence' [6], proposed the 'imitation game' as an attempt to test a machine's intelligence through its conversational competence, provided it gives answers that would naturally be given by a human. A standard version of the imitation game involves three major components: a human (A), a machine (B) and a human interrogator (C). The game is held in a setting where a human interrogator asks questions via teletypewriter. Here, the interrogator knows the entities only by labels such as X and Y, and remains unaware of their true identities. Once the game starts, the interrogator asks a list of questions to both participants within a limited time frame, and based on their responses, he then determines whether the answers were given by a human or a machine (Figure 1). A machine passes the Turing test if it successfully imitates the conversational style of a human or if the human interrogator fails to tell them apart. While the imitation game has long been considered a primary benchmark for evaluating intelligence as a competence in AI systems, recent developments have reshaped how we see the imitation game and its applicability.

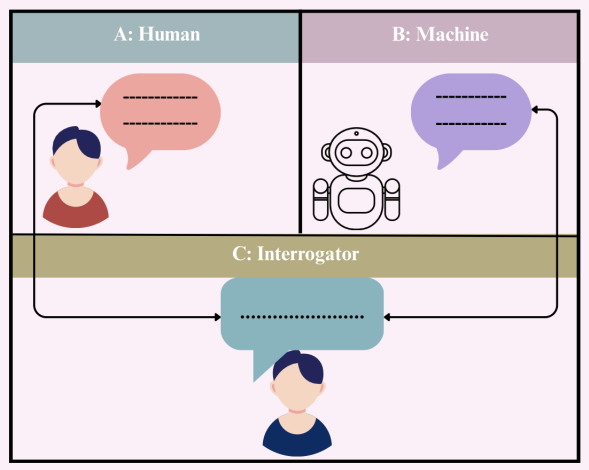

Figure 1. *The Imitation Game*

From symbolic AI to modern deep learning methods, the ability of artificial systems to imitate human intelligence has only improved with time. While early systems such as ELIZA had fairly structured, repetitive, and rigid responses to the interrogator's questions, contemporary systems have achieved unprecedented sophistication in imitating human conversational competence. Their ability to maintain contextual coherence across conversations and generate flexible, human-like, responses has significantly improved after the emergence of transformer-based LLMs. For the past few years, the language models have been deployed across manifold domains of our social existence, given their ability to hold expert-level conversations with humans and resolve their queries. However, irrespective of all these advancements, a persistent sense of unease remains. Despite convincing many users about their intelligence and even consciousness, the responses generated by AI systems are often identified as artificially forged. This paper attends to the potential implications of such failures. What did Turing mean when he suggested that the best strategy might be to imitate something other than a human? [6] Did he foresee that a perfect imitation of human intelligence would eventually feel like a parody? Or the fact

[1] Indian Institute of Technology Jodhpur, email: p22hs201@iitj.ac.in
[2] Indian Institute of Technology Jodhpur, email: kjg@iitj.ac.in

that deception as a criterion of imitation is problematic?

Following the above discussion, the present paper briefly addresses two major structural challenges in the Turing test: a) indeterminacy of the object of imitation, and b) the variability of the interrogator. We argue that 'human intelligence' might not be the most appropriate benchmark for evaluation. As Turing suggested, machines may carry out something that ought to be described as thinking, but which is very different from what a human does. In this sense, we understand that by indicating 'the best strategy for a machine may possibly be something other than the imitation of a human' (Turing, 1950), Turing implicitly recognised the limitations of human-centred criteria and opened the possibility that artificial systems might instantiate a distinct form of intelligence. Accordingly, we argue that the apparent failure of machines in the imitation game reflects not the lack of intelligence, but a mismatch between their mode of cognition and the unstable assumptions of the test, which necessitates a broader reconceptualization of intelligence beyond imitation.

2 UNDERSTANDING THE INTENTIONS OF IMITATION GAME

Unfortunately, Turing does not help much in understanding the intentions and implications of the game. He did not even attempt to answer the central question "Can machines think?" in its original form but rather replaced it with a game that he considered sufficient to evaluate the machine's behaviour. Turing argued that the original question is too meaningless to deserve discussion, and therefore, the imitation game as a replacement provides the best strategy to avoid the ambiguity of the key terms 'machine' and 'thinking', while giving us grounds to assess the possibility of intelligence as competence in machines. The game is considered a strong source of inductive inference for attributing thinking to machines [2]. However, one may ask: if the original question is really meaningless, how are we supposed to believe in the propriety and adequacy of the game it is being replaced with? [2] Anyways, our interest in machine thinking does not concern their ability to play yet another game; rather it concerns possible computer activities and the concept of thinking [2].

A common interpretation of the imitation game takes the test as a theory that provides an operational definition of intelligence [2]. P.H. Millar, for instance, argued that "Turing test constitutes an operational definition which, given a computer terminal system, can be used as a criterion" [1]. According to other popular accounts, the game treats apparent linguistic indistinguishability and deception as sufficient conditions for assessing a machine's intelligence [3]. However, this perspective is difficult to reconcile with Turing's own methodology which carefully avoids stating any explicit definition of intelligence. The significance of this move lies in its methodological restraint. Turing did not make any explicit claim that passing the test would prove a machine's intelligence, nor did he provide any necessary and sufficient conditions for intelligent behaviour. In any case, the benefit of such understanding proves difficult to sustain. The test becomes interesting only if it connects to the idea of intelligence. As Moor argued, "Either one is giving an arbitrary definition of a new term in which case the operational definition needs no justification, but the Turing test becomes very uninteresting, or one is attempting to capture a notion which is at least related to our ordinary concept of thinking in which case the Turing test is very interesting, but the operational definition requires some justification." In this sense, the significance of this game is dependent upon its ordinary understanding of thinking.

3 THE STRUCTURAL INSTABILITY IN THE GAME

Now, if the significance of the game is indeed dependent upon our understanding of thinking, then the question is no longer about any explicit definition of intelligence, but rather to see if the game provides stable conditions for the assessment. Since Turing kept the idea of intelligence open and unclear, it seems the unclarity became the reason for two major structural challenges in the test.

3.1 The indeterminacy of object of imitation:

The primary issue which remains consistent is the condition that passing the Turing test requires machines to 'provide answers that would naturally be given by a man" [6]. Let us ignore the gender bias here and assume that he meant 'answers given by humans'. What kind of humans? It seems that Turing presupposes 'human intelligence' as a stable and identifiable target for imitation; however, he never specified the kind of intelligence to be imitated. The flaw lies in having a specific object of imitation. For instance, to imitate Mona Lisa is to have a specific object with well-defined features. However, this is not the case with human intelligence. Human intelligence is not a determinate object, but rather a spectrum of performances shaped by context, ability, and circumstances. The test never specifies whether an average, expert-level, or ideal human intelligence is to be imitated. This, in turn, removes any stable criteria of success and shifts the burden to the interrogator.

3.2 The variability of the interrogator:

This marks another issue in the imitation game. If the success or failure of a machine depends on the interrogator's knowledge, expectations, and interpretive abilities, then a machine may succeed with one interrogator and fail with another, suggesting that the test measures relative alignment rather than objective intelligence. The issue becomes more evident when applied to contemporary chatbots for emotional support. Chatbots, such as Replika, Webot, and Character.ai, are famous for imitating human emotional responses [5]. However, the success or failure of the imitation significantly depends on the user's state. Machine's imitations of humans' emotional responses are more likely to be experienced as convincing and meaningful when the user is in a vulnerable state. It implies that the system's ability to deceive the interlocutor depends heavily on the user's condition. When applied to Turing's imitation game, a user's limited cognitive capacity may allow a system to pass the test even when it lacks sufficient capacities.

Now, in response to these two objections, one may argue that Turing intentionally avoided defining intelligence, and that the objections do nothing more than describing what he intended. While it is true that Turing intentionally avoided providing any explicit definition of intelligence, there is a big difference between productive and problematic openness. By productive openness, we mean 'leaving something undefined to let the proposed framework grow and adapt', and problematic openness emerges when 'undefined things become the reason for instability in the framework, making it unable to do the jobs it claims to do'. In this case, the kind of openness Turing chose by not defining 'intelligence per se' is problematic. It is the reason why the game inherits the very

ambiguity it was designed to escape. While Turing's intentional openness can be philosophically defended as speculation, it becomes structurally damaging when it is adopted as a criterion to evaluate a machine's behaviour.

This structural instability was not limited to the text but had further visible consequences. In 2014, Eugene Goostman, a conversational chatbot, was widely reported to have passed the Turing test [7] [8]; however, the claim quickly collapsed, as no consensus existed on what passing the Turing test meant [4]. It was one of the cases where chatbot failed the Turing test, not because it performed poorly, but because there were no specified criteria for final judgment. The current LLMs also face similar confusions, as there is a major divide between expert judgments: some reduce LLMs to stochastic parrots or pattern matching systems, while others consider them genuinely intelligent systems. The ambiguity persisted for so long because the standards by which they were measured were never settled. It seems Turing was never ignorant of these issues. His suggestion that perhaps something other than human intelligence should be imitated was not casual. He was aware that "a machine may carry out something that ought to be described as thinking, but that is very different from what a human does" [6].

4 IS THERE STILL LIFE IN TURING TEST?

Thus, after seventy five years, the continued use of the imitation game as a decisive evaluative benchmark has become really difficult to justify in the light of contemporary AI systems. In the traditional sense, the game derived much of its significance from a fundamental asymmetry between humans and computers: humans were considerably intelligent, and computers were not. Building on this asymmetry, the condition for computers to imitate humans' conversational style was not only provocative but also appeared meaningful. However, things have escalated drastically over the last seventy-five years. AI systems are already competent enough to generate high-level linguistic patterns. Thus, the concern is no longer about their ability to imitate humans; rather much more troubling issue is to find if the game is competent enough to evaluate the conversational ability of AI systems. Precisely, the test became methodologically weaker at the moment when AI systems turned competent to challenge its hidden assumptions.

5 CONCLUSION

The Turing test appears to have lived two distinct lives. The first is as a philosophical provocation intended to redirect the discussion away from sterile disputes about the nature of thinking. In this role, its conceptual openness is not a defect but a virtue, which allows reflection on the conditions under which intelligence is attributed. The second is as a practical benchmark for evaluating machine intelligence. Here, however, the same openness becomes problematic. A test intended to classify, compare, and measure is in need of stable criteria for its success; yet the imitation game leaves both the object of imitation and the standards of judgment unresolved. Much of the subsequent debate surrounding the Turing Test can be understood as a conflation of these two quandaries. The persistent disagreement over whether specific systems have passed the test unveils not merely disagreement about machines, but uncertainty about what exactly the test was ever meant to establish. If so, the challenge is not simply to design a better benchmark, but first to clarify what kind of phenomenon is being evaluated and whether human intelligence should continue to serve as the privileged model against which machine cognition is assessed.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper.

REFERENCES

[1] P. H. Millar, 'On the point of the imitation game', *Mind*, **82**(328), 595–597, (1973).
[2] James H. Moor, 'An analysis of the turing test', *Philosophical Studies*, **30**(4), 249–257, (1976).
[3] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.
[4] Ian Sample and Alex Hern, 'Scientists dispute whether computer "eugene goostman" passed turing test', *The Guardian*, (June 2014).
[5] Vivian Ta, Charlotte Griffith, Christopher Boatfield, Xiaoxuan Wang, Michael Civitello, Hannah Bader, Elizabeth DeCero, and Alex Loggarakis, 'User experiences of social support from companion chatbots in everyday contexts: Thematic analysis', *Journal of Medical Internet Research*, **22**(3), e16235, (2020).
[6] Alan M. Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433–460, (1950).
[7] Kevin Warwick and Huma Shah, 'Can machines think? a report on turing test experiments at the royal society', *Journal of Experimental and Theoretical Artificial Intelligence*, **28**(6), 989–1007, (2016).
[8] Kevin Warwick and Huma Shah, 'Passing the turing test does not mean the end of humanity', *Cognitive Computation*, **8**(3), 409–419, (2016).