A Bona Fide Turing Test

Sharon Temtsin
sharon.temtsin@pg.canterbury.ac.nz
University of Canterbury
Christchurch, New Zealand

Diane Proudfoot
diane.proudfoot@canterbury.ac.nz
University of Canterbury
Christchurch, New Zealand

Christoph Bartneck
christoph.bartneck@canterbury.ac.nz
University of Canterbury
Christchurch, New Zealand

ABSTRACT

The constantly rising demand for human-like conversational agents and the accelerated development of natural language processing technology raise expectations for a breakthrough in intelligent machine research and development. However, measuring intelligence is impossible without a proper test. Alan Turing proposed a test for machine intelligence based on imitation and unconstrained conversations between a machine and a human. To the best of our knowledge, no one has ever conducted Turing's test as Turing prescribed, even though the Turing Test has been a bone of contention for more than seventy years. Conducting a bona fide Turing Test will contribute to machine intelligence evaluation research and has the potential to advance AI researchers in their ultimate quest, developing an intelligent machine.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Turing test, imitation game, conversational agent, artificial intelligence

ACM Reference Format:
Sharon Temtsin, Diane Proudfoot, and Christoph Bartneck. 2022. A Bona Fide Turing Test. In *Proceedings of the 10th International Conference on Human-Agent Interaction (HAI '22), December 5–8, 2022, Christchurch, New Zealand*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3527188.3563918

Conversational agents are frequently used in commerce [11], education [8], and games [3]. Moreover, technological developments grant companies the ability to integrate conversational agents into private sector smart systems. Smart devices allow users to manage their time efficiently with a digital assistant, such as Apple's Siri and Amazon's Alexa. Still, there is a long way ahead, as chat users reported a preference for a human-like experience when interacting with a conversational agent [7, 9]. Currently, no conversational agent can provide a human-like intelligent conversation.

The progress in machine learning research provides a means to represent a complete target system with a single model [10]. AI research companies, like OpenAI and Meta AI, allow access to their own natural language processing end-to-end learning models, OpenAI's GPT and Meta AI's BlenderBot. Given a natural language model, developers can utilise a conversational agent framework, such as Amazon Lex and Google's Dialogflow, for creating their own conversational agent applications.

However, recent events bring the lack of testing methods to the fore. The recent dispute between Google and Blake Lemoine may reinforce the feeling that true machine intelligence is just around the corner [16]. This dispute would not have happened had we had a proper test. Although in this case the issue is whether the technology is sentient, a similar issue may appear regarding a machine's intelligence. Testing the conversational agent's ability to imitate a human's unconstrained conversational competence could be such an intelligence test. Alan Turing proposed an experiment based on an imitation game for testing intelligence in machines [18]. Over time, many started to informally refer to his test as "the Turing Test".

Alan Turing suggested three versions of an imitation game. In 1948, Turing wrote a report where he presented a restricted chess-playing imitation game [19]. In 1950, Turing described a second imitation game for intelligence in machines that was characterised by an unconstrained conversation with three players [18]. The second imitation game is played by three participants in two separated rooms.

The players in the rooms are not allowed to exchange any information other than written text. One room contains an interrogator and the other room two players. Player *(A)* is a machine and player *(B)* is a human. The interrogator *(C)* is a human. The interrogator has an unconstrained conversation with both player *A* and player *B* in parallel. The interrogator's objective in the machine-imitates-human game is to determine which of the two players is the machine and which is the human. Player *A*'s target is to mislead the interrogator and player *B*'s is the opposite.

Turing proposed a third version of the imitation game during a discussion in 1952 [17]. This imitation game consists of only two players: the computer/human and the interrogator. In this version the interrogator' role is taken by a jury. Each juror must judge several players, some machines and some humans. The machine's aim is to convince a considerable number of jurors that they are having a conversation with a human, while in reality the conversation is with a machine. Nevertheless, Turing was concerned that in this two-player game, the jury would tend to classify a human as a machine rather than the opposite [17]. This issue received confirmation from a series of Turing-style test contests called the "Loebner Prize Competition" [12]. Such a phenomenon cannot happen in the three-player imitation game [5, 14].

Since 2004 the Loebner Prize Competition changed the contest format from a two-player to a three-player imitation game. In 2008,

HAI '22, December 5–8, 2022, Christchurch, New Zealand
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9323-2/22/12…$15.00
https://doi.org/10.1145/3527188.3563918

the Loebner Prize and the University of Reading joined forces to conduct Turing Tests in the three-player imitation game format. The University of Reading ran two additional Turing Test competitions in 2012 and 2014 [20]. However, the organisers misinterpreted Turing's test and used Turing's prediction in his 1950 paper as a benchmark for passing the test. Turing's prediction was that, given a computer "with a storage capacity of about 10^9, ... an average interrogator will not have more than 70 per cent. chance of making the right identification after five minutes of questioning." [18].

Shah and Warwick [15] explained the decision to use Turing's prediction as a fixed benchmark. In their view, the goal of the machine is to imitate the human foil's properties in the game. Following this motivation, a benchmark of 70% was chosen as a protocol for scoring the game. However, just because Turing predicted 70% accuracy after five minutes for a computer with the given storage capacity, this does not make it a good benchmark value. One could just as well argue for a 75% benchmark to decide whether the machine is intelligent.

A special case occurs if the benchmark is 50%. If the interrogator is only correct 50% of the time, that is no better than taking a guess in every imitation game. Unlike Shah and Warwick [15]'s aim, here the idea is to imitate the whole property set of the human foil. In this case the interrogator cannot distinguish between the conversational agent and the human. Therefore, the machine can be acknowledged as intelligent. Considering such a machine-imitates-human game interpretation is possible if Shah and Warwick [15]'s explanation of the man-imitates-woman game role in Turing's test is acceptable. The man-imitates-woman game is similar to the computer-imitates-human game with two changes. A man takes the place of the player *A* and a woman takes the place of the player *B*.

According to Shah and Warwick [15], the man-imitates-woman game is not part of the protocol for scoring Turing's imitation game, but an introduction to the three-player imitation game. If the man-imitates-woman game is not part of the protocol, then the Turing Test is a fixed benchmark protocol with 50% benchmark.

Different aspects of Turing's test must be considered before we start designing a bona fide Turing Test. One of the aspects is the benchmark proposed by Turing. Turing wondered whether the performance of the interrogator in the computer-imitates-human game is going to be the same as the performance of the interrogator in the man-imitates-woman game [18]. Copeland and Proudfoot [6] recognised the man-imitates-woman game as a benchmark for the machine-imitates-human game. For a machine to do well in the imitation game, the interrogator must guess wrongly between a conversational agent and a human in the computer-imitates-human game no less frequently than does the interrogator in the man-imitates-woman game. (We would like to thank Jack Copeland for the research idea to run the man-imitates-woman game experiment as part of the Turing Test protocol.)

Another aspect of Turing's test is interrogation duration. Shah and Warwick [15] used only five minutes in their tests, based on Turing's prediction. Five minutes is a short conversation duration, much like an introductory meeting between strangers. First impressions may be useful as a preliminary test for intelligence, but cannot be the ultimate means to decide whether the machine is intelligent. The duration of each imitation game must be carefully considered during the experimental design. It should be noted that Turing did not mention any specific duration for his test.

Experimenting with the game duration variable in a pilot study may indicate an appropriate duration for both the man-imitates-woman and the computer-imitates-human games. The man-imitates-woman game benchmark may change after a long time period. The man's ability to imitate a woman might be affected by social changes as well as the performance of the interrogator. Therefore it is important to remember that Turing's test should be calibrated by conducting a man-imitates-woman game regularly.

Despite the absence of man-imitates-woman game experiments, researchers have conducted gender detection experiments. Patterson et al. [13] and Argamon et al. [2] conducted a gender detection algorithm evaluation experiment. The algorithm takes as an input a human's written text and outputs the writer's gender. However, the researchers did not base their experiments on unconstrained conversations. Adam et al. [1] and Collins and Evans [4] experimented on a modified version of a gender imitation game. In these experiments too performance was not calculated based on the results of unconstrained conversations. These experiments may shed some light on the challenges of conducting the man-imitates-woman game experiments.

According to our knowledge, no experiment has accurately followed Turing's test protocol. The test's properties, such as its validity, reliability, sensitivity and accuracy also remain unclear. Given the enormous media attention that the Turing Test regularly receives and the extensive references to it in the scientific literature, one must be surprised at the absence of experiments that accurately follow Turing's protocol. To avoid an 'Emperor's New Clothes' scenario, we intend to conduct a series of experiments that will help us to better understand Turing's test.

Establishing practical guidelines for conducting Turing Tests would allow us to formulate an intelligence criterion. A machine that satisfies this criterion possesses intelligence. We should remember, however, that a machine can be intelligent even though it does not pass the test. A conversational agent's ability to pass the test has the potential to satisfy or at least to improve the previously mentioned user's demand for human-like experience.

We hope that a better understanding of Turing's test as a tool to assess the intelligence of conversational agents will enable developers to better evaluate their systems. Companies such as Google should use such a tool to check whether their machines are considered intelligent under the unconstrained conversation condition, which might allow them to avoid controversies. Conducting a bona fide Turing Test is the first step in this direction.

REFERENCES

[1] R. Adam, U. Hershberg, Y. Schul, and S. Solomon. 2004. Testing The Turing Test—Do Men Pass It? *International Journal of Modern Physics C* 15, 08 (2004), 1041–1047. https://doi.org/10.1142/S0129183104006522

[2] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & talk* 23, 3 (2003), 321–346. https://doi.org/10.1515/text.2003.014

[3] D.J.H. Burden. 2008. Deploying embodied AI into virtual worlds. In *Applications and Innovations in Intelligent Systems XVI* (London), T. Allen, R. Ellis, and M. Petridis (Eds.). Springer, 103–115. https://doi.org/10.1007/978-1-84882-215-3_8

[4] H. Collins and R. Evans. 2014. Quantifying the tacit: The imitation game and social fluency. *Sociology* 48, 1 (2014), 3–19. https://doi.org/10.1177/0038038512455735

[5] B.J. Copeland. 2000. The Turing Test. *Minds and Machines* 10, 4 (2000), 519–539. https://doi.org/10.1023/A:1011285919106

[6] J. Copeland and D. Proudfoot. 2009. Turing's Test: A philosophical and historical guide. In *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, G. Epstein, R.and Roberts and G. Beber (Eds.). Springer, Dordrecht, 119–138. https://doi.org/10.1007/978-1-4020-6710-5_9 Export Date: 10 July 2022; Cited By: 14.

[7] A. Følstad and M. Skjuve. 2019. Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st international conference on conversational user interfaces*. 1–9. https://doi.org/10.1145/3342775.3342784

[8] A. J Gonzalez, J. R Hollister, R. F DeMara, J. Leigh, B. Lanman, S. Lee, S. Parker, C.r Walls, J. Parker, and J. Wong. 2017. AI in informal science education: Bringing Turing back to life to perform the Turing test. *International Journal of Artificial Intelligence in Education* 27, 2 (2017), 353–384. https://doi.org/10.1007/s40593-017-0144-1

[9] M. Jain, P. Kumar, R. Kota, and S.N. Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong). Association for Computing Machinery, 895–906. https://doi.org/10.1145/3196709.3196735

[10] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences* 40 (2017). https://doi.org/10.1017/S0140525X16001837

[11] X. Luo, S. Tong, Z. Fang, and Z. Qu. 2019. Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science* 38, 6 (2019), 937–947. https://doi.org/10.1287/mksc.2019.1192

[12] J. H Moor. 2001. The status and future of the Turing test. *Minds and Machines* 11, 1 (2001), 77–93. https://doi.org/10.1023/A:1011218925467

[13] W. Patterson, J. Boboye, S. Hall, and M. Hornbuckle. 2017. The gender Turing test. *International Conference on Applied Human Factors and Ergonomics*, 281–289. https://doi.org/10.1007/978-3-319-60585-2_26

[14] D. Proudfoot. 2011. Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence* 175, 5-6 (2011), 950–957. https://doi.org/10.1016/j.artint.2011.01.006

[15] H. Shah and K. Warwick. 2010. Testing Turing's five minutes, parallel-paired imitation game. *Kybernetes* 39, 3 (2010), 449–465. https://doi.org/10.5220/0005736802150222

[16] N. Tiku. 2022. *The Google engineer who thinks the company's AI has come to life*. Retrieved August 01, 2022 from https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

[17] A. Turing, R. Braithwaite, G. Jefferson, and M. Newman. 2004. Can Automatic Calculating Machines Be Said To Think? (1952). In *The Essential Turing*, Jack Copeland (Ed.). Clarendon Press, 487–515.

[18] A. M. Turing. 1950. Computing machinery and intelligence. *Mind* LIX, 236 (1950), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[19] A. M. Turing. 2004. Intelligent machinery. In *The Essential Turing*, Jack Copeland (Ed.). Clarendon Press, 395–432.

[20] K. Warwick and H. Shah. 2016. Can machines think? A report on Turing test experiments at the Royal Society. *Journal of experimental & Theoretical artificial Intelligence* 28, 6 (2016), 989–1007. https://doi.org/10.1080/0952813X.2015.1055826