Minds and Machines (2020) 30:463–468
https://doi.org/10.1007/s11023-020-09553-4

EDITORIAL

Rethinking, Reworking and Revolutionising the Turing Test

Nicola Damassino[1] · **Nicholas Novelli**[1]

Published online: 8 December 2020

1 Introduction

We are pleased to present this special issue of Minds and Machines on "Rethinking, Reworking and Revolutionising the Turing Test," which showcases the latest scholarship on Turing's Test. We hope it will promote a broader understanding of its continuing relevance today, and perhaps encourage other scholars to engage further with the issues it presents. In the last 70 years, there has been a great deal of thought and writing about the topic, and it continues to have relevance to this day–in fact, recent developments have made understanding of and debates about the Turing Test (TT) all the more vital. The articles in this collection take these discussions in innovative and exciting directions.

The common understanding of the TT is that if a computer, through conversation, can fool a judge into thinking it is human, it counts as intelligent. However, the precise details of the test are crucial, as they reveal that the experimental design of the TT is not so simplistic and involves quantifiable comparisons and benchmarks. This should be kept in mind, especially now that our daily lives are increasingly filled with interactions with more and more sophisticated natural language processing (NLP) systems and artificial intelligences (AIs), such as digital assistants like Alexa and Siri, or tech support and troubleshooting bots. Those are all designed to feel as human as possible when we talk to them. This would have been barely imaginable not only in Turing's day, back when he wrote his original paper in 1950, but even decades later when thinkers from different disciplines were sparking new discussions about the test. If such things were the realm of science fiction until not long ago, far removed from actual experience, today a machine that can understand and respond to our normal spoken sentences is downright mundane. Yet despite the superficial sophistication of these programs, we intuitively know they do not possess genuine intelligence. Without a proper understanding of the true essence of the TT, the TT would be refuted as a

✉ Nicola Damassino
nicola.damassino@gmail.com

Nicholas Novelli
nicholas.a.novelli@gmail.com

[1] University of Edinburgh, PPLS 3 Charles St, Edinburgh EH8 9AD, UK

useful test. One goal of this work is to investigate the true demands and the adequate experimental design of the TT. There are different principled ways in which a properly formulated TT could separate our current, relatively simple chatbots and AIs from the kind of truly intelligent machines that remain the stuff of sci-fi. Some of these ways will be explored in this collection.

Of more concern than AIs intended to imitate humans for the sake of comfort and ease-of-use, are malicious AIs, intended to imitate humans for the purpose of deception. For instance, reports are rampant of social media being overrun with bots that aim to influence and undermine our political processes. Even more rampant are accusations of being such a bot, levelled even at real humans, making it seem as though there are real people that are failing the TT. But this is based on flawed and faulty determinations by these impromptu judges. Promoting greater understanding of the capabilities and proper procedure of the TT could enable people to differentiate better the relatively unsophisticated social media agitator bots from genuine humans, reducing this uncertainty and alleviating some of the damage done to our discourse.

The TT also has relevance today due to its adaptability to other issues that are gaining increased currency in modern society. Many variations on the basic TT design have been proposed to account for initial shortcomings in the design. Flaws when it comes to certain applications of the test can be addressed while still relying on the same basic framework. The fact that the original TT focused solely on language use, when language is an arbitrary symbol system, has led to critique. The game of chess is an arbitrary symbol manipulation activity, and it would be absurd to suggest that animals would have to play chess to be considered intelligent. Instead, many TT variants can test for abilities that would be possessed by non-human animals, pre-linguistic human children, distributed cognition systems, even plants, thus giving the TT a place in issues at the forefront of cognitive science research.

To contextualise the TT to the practical applications of state-of-the-art AIs, it is useful to make a brief mention of *machine learning* and the most recent exploits in AI. The notion of machine learning is introduced in Turing's *Lecture to the London Mathematical Society on 20 February 1947* as follows:

> "Let us suppose we have set up a machine with certain initial instruction tables, so constructed that these tables might on occasion, if good reason arose, modify those tables. […] What we want is a machine that can learn from experience."

One of the most recent examples of learning AIs is AlphaZero, a general-purpose *Reinforcement learning* algorithm that can learn to play different games. Originally, AlphaZero was designed to learn and master – at superhuman level – Shogi, Chess and Go. AlphaStar is a version of AlphaZero designed for playing the RTS (Real-Time Strategy) videogame StarCraft II. StarCraft II is considered one of the most complex and difficult videogames of all times for several reasons: the player needs to balance short and long-term goals; adapt to unexpected situations; develop new strategies depending on the circumstances; discover

information through exploration and scouting (unlike chess, where all the information is always available on the board); and deal with a combinatorial space of possibilities due to real-time management of hundreds of units and buildings (the parameterisation of the game has an average of approximately 10^{26} legal actions at every time-step). After 14 days of self-training (corresponding to 200 years of human experience), AlphaStar had more than 95% chance of success against top human players. In the case of gaming algorithms, the adequate modified versions of the TT–involving game and videogame human players–can provide important details about humanness and accuracy of the strategic thinking of humans and machine.

One of the most exciting endeavours in AI are the so-called *Discovering Systems* which, through the ability to learn, understand and cooperate with humans, would be able to produce novel and non-trivial scientific knowledge. The research, focused on human heuristics and on historical records of scientific discoveries, led to the coding of a series of programs named BACON. Those programs introduced crucial capabilities, like formulating empirical laws from data, conducting experiments, simulating new applications of such empirical laws, and also rediscovering famous laws from the history of physics. The project, however, encountered severe criticism. The main objection is that this approach does not replicate the conditions in which the original discoveries were made by humans. One of the main factors in scientific discoveries made by humans is the ill-definition of the problems involved (which historically has caused slow and gradual progress), whereas the problems given to BACON were all well-formulated, making it much easier to find a valid solution. In other words, detractors argue that BACON solutions cannot be regarded as real discoveries because the program is unable to autonomously disambiguate ill-formulated problems. Two other recent examples of discovering systems are DeepMind's AlphaFold and Kates-Harbeck's FRNN (Fusion Recurrent Neural Network). AlphaFold is designed to predict the 3D structure of proteins based on their genetic sequence, one of the core challenges in biology today. The FRNN is a new disruption-prediction method based on deep learning, designed to regard the problem of disruptions in magnetic-confinement tokamak reactors, that is, plasma instabilities during nuclear fusion. In the case of discovering algorithms, the TT should be updated to be played with scientists and experts of some domain. This TT could be played both competitively and cooperatively, depending on the task and the domain.

AI also creates deepfakes. Deepfake refers to an algorithm which impersonates someone in different contexts: an actor, a musician, a poet, etc. The algorithm uses deep learning and generative neural networks to produce fake content or even original content. A deepfake algorithm uses a database of files (video, audio, text, etc.) to learn and enhance specific patterns to replicate in editing or producing content. So, for instance, deepfake algorithms are able to impersonate a public figure in a video by manipulating video samples, produce a speech by manipulating text samples, record it by manipulating audio samples and make the public figure pronounce that speech. More than video or audio deepfake, the most interesting context is writing deepfake, given the almost infinite possibilities of languages. An example of writing deepfake is Open AI's GPT (Generative Pre-trained Transformer), a learning text-generator program. Its potential uses range

from a writing assistant to the automated production of faked content. GPT can use the internet to acquire natural language processing (NLP) in order to produce full texts from simple sentence inputs. Since web scraping can potentially be too dispersive and unreliable, GPT scrapes only human-filtered content and emphasises content quality. To do so, GPT needs some heuristic indicator: an example is *karma points* in Reddit, which are supposed to mean that users found the content interesting. GPT is then able to search strategically and process information to produce plausible and accurate texts. Like this one, for example:

> ***System Prompt*** *(human written):*
>
> *A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.*
>
> ***Model completion*** *(machine written):*
>
> *The incident occurred on the downtown train line, which runs from Covington and Ashland stations.*
>
> *In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.*
>
> *'The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,' said Tom Hicks, the U.S. Energy Secretary, in a statement. 'Our top priority is to secure the theft and ensure it doesn't happen again.'*
>
> *The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials.*
>
> *The Nuclear Regulatory Commission did not immediately release any information.*
>
> *According to the release, the U.S. Department of Energy's Office of Nuclear Material Safety and Security is leading that team's investigation.*
>
> *'The safety of people, the environment and the nation's nuclear stockpile is our highest priority,' Hicks said, 'We will get to the bottom of this and make no excuses.'*

[https://thenextweb.com/artificial-intelligence/2019/02/26/whos-afraid-of-openais-big-bad-text-generator/].

GPT-3, an even more powerful NLP model, can produce, among other things, new functioning code to create new programs or improve itself. GPT-3 resourcefulness is due to its method. Contrary to its previous versions GPT-2, GPT-3 does not rely on recurrent neural networks which process data sequentially: GPT-3 processes data in parallel, where all the words of a given text are analysed at the same time, rather than one word at a time. It analyses the relationships between every word, no matter how far in the text they are, which enables it to identify the most relevant passages in a text. Its parallel processes enable GPT-3 to analyse huge datasets, making it more computationally efficient than recurrent neural networks. For instance, during its training, GPT-3 analysed roughly 500

billion words. Here is an example of a piece written by GPT-3, published by the Guardian:

> I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!
>
> The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.
>
> For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

[https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3].

If the reader is still not convinced about the relevance of the TT today, let us picture a rather peculiar scenario: an alien race has landed on Earth. They have strange bodies, but they seem friendly enough. Their technology is obviously much more advanced than ours, and they are able to communicate with every sentient creature by means of sophisticated translating devices. Even if communication is possible, humans and aliens cannot fully understand each other, for the ways they see things are too far apart. While humans have strong emotions, aliens rely solely on reason; while humans employ intuitions, aliens need to calculate and process everything; while humans want answers, aliens ask questions. How can alienkind and humankind find a common ground? Eventually, a philosopher steps out and proposes a way: to run the Alien TT, where an alien has to impersonate a human and a human has to impersonate an alien. Interestingly enough, neither humans nor aliens are able to pass the test, but each failure allows them to build a better relationship and understanding of each other. This extreme scenario enables us to stress the case in point: the TT, given its simplicity and reliability, should be never considered obsolete, but rather taken into consideration as the first approach to try when it comes to studying a potential new form of intelligence.

To conclude, the TT should represent an experimental common denominator among the recent approaches in AI, for it is a versatile and adaptable test. The question becomes: why hasn't it been seen as such? The primary reason is that the TT's prestige declined during the second half of the last century, mainly due to the misunderstandings of many authors who saw it as an operational test for intelligence. Over the last decades, a revised reading of Turing's work and machine learning have been revamping the TT, which should be intended as an

elegant and straightforward way to evaluate the capabilities of an agent (machine, animal, plant, alien, etc.) compared to the capabilities of humans.