OXFORD

The Computer Journal, 2024, 1–3
https://doi.org/10.1093/comjnl/bxae078
Original Article

From breaking to faking the code: Alan Turing's Imitation Game latest upgrade for discerning Artificial Intelligence (AI)-generated deepfakes

Marios C. Angelides*

Brunel Design School, College of Engineering Design and Physical Sciences, Brunel University London, Kingston Lane, Uxbridge UB8 3PH, United Kingdom

*Corresponding author: marios.angelides@brunel.ac.uk

Abstract

Alan Turing developed a simple test for verifying whether machine or man. He did this with a vision that was decades ahead of the time but with the technology that was available to him at the time. Back in the time, the notion of AI was a future prospect, and not a real threat in any way to humanity. Roll the clock forward to today and AI is not a prospect but a stark reality and the threat to humanity, according to the sceptics, all but real. But how serious has the threat become? This paper investigates.

1. ALAN TURING'S ORIGINAL VISION OF THE IMITATION GAME

Attempting to get a true sense of the Turing test, the natural choice is to go back to the original publication in Mind from 1950 [1] and to read about, and understand, the test in the master's own words. Many versions of the original article have appeared since then, but many would question if any of these newer and polished versions measure up to the original and whether Turing would approve any of them. And there is a certain degree of nostalgia going back to the original which one does not feel with the newer versions. Turing's extraordinary vision is coming through clear and so is his forward thinking, but it requires keeping an open mind about his discussion of human cognition and interaction, AI, and technology which is redolent of the time.

Turing introduces the Imitation Game by asking the question on everyone's lips, *can machines think?* He rules out a definitional response, a statistical answer, as he calls it. Instead, he opts to transpose the question into a problem that he considers unambiguous, i.e. the Imitation Game that is played with a man, a woman, and an interrogator the last of which is kept apart from the other two and is challenged to determine which of them is the man and which is the woman by asking questions and receiving their response through an intermediary. What is significant in Turing's game is that the use of deception by either the man or the woman or both is allowed, and it is up to the interrogator to decide that. Turing then revises his original question to consider *what will happen when a machine takes the part of the man in the game?* How often will the interrogator decide wrongly in the machine versus human version of the game in comparison to the human versus human version of the game?

Turing then, drawing a line between the physical appearance and intellectual capability of humans argues that there is little point in making a thinking machine look like more human by dressing it up in artificial flesh, considering that the nature of the test is purely a Q&A session, without any practical demonstrations, carried out by a third party and where no human endeavour subject is off limits. He claims, prophesizes rather, that at some time this, i.e. making a machine look more human, might be done. He further justifies his assumptions by arguing that machines are no match to human appearance and likewise humans are no match to machine capabilities.

He predicts that the odds are weighted too heavily against the machine when thinking is required and likewise for the human when speed and accuracy is required. He concedes that to overcome the machine weakness with regards to thinking, the machine needs to be constructed to play the Imitation Game satisfactorily or steer away from imitation of human behaviour. He concludes that the best strategy for the machine is to try to provide answers that would naturally be given by humans.

Turing then turns his attention to identifying and justifying the nature of the machine to be used and he selects the digital computer as a close model to a 'human' computer that is equipped with a fictional rule book and whose rules change from task to task. The rule book may be used to program the digital computer to mimic the actions of the human computer. He considers introducing the use of randomness or free will, but he argues against that as it will be difficult to distinguish between the two. He concludes that 'he is not considering whether all digital computers would do well in the game nor

Received: June 28, 2024. **Accepted:** August 7, 2024

whether the computers at present available would do well, but whether there are imaginable computers which would do well.' He revises his original question to consider *whether there are imaginable digital computers which would do well in the Imitation Game?*

Turing considers the legitimacy of the questions he raises and takes the opportunity to make his now famous predictions that by the end of the (last) century, it will be possible to programme computers to play the Imitation Game so well that an average interrogator's chances of making the right identification after 5 min will not be more than 70% and that general opinion on machine thinking will have shifted. He also raises an important concern with regards to the fine line between fact and conjecture, so this does not result in harm.

He goes on to consider possible contrary views to his questions through what he perceives might be legitimate objections for lack of evidence. Starting with the theological objection, he discusses soul creation and moves to what he labels the heads in the sand objection where he discusses the superiority of humans. He then moves on to the mathematical objection where he discusses the limitations of machines because of their lack of human intellect and then turns to discussing consciousness, or lack of it, in machines and seems to take a neutral view on this. He then questions whether a machine can truly be the subject of its own thought and be made to exhibit human traits and he concludes by predicting that this is not utopia and will be made possible in the future. But for the purpose of the test, it will be sufficient to be made to exhibit human behaviour. He then disagrees with what might be Lady Lovelace's objection, i.e. that electronic equipment cannot think for itself, or learn to that effect, and that is not capable of original thought. He continues with a discussion on mimicking the behaviour of the nervous system and concludes that it will be difficult for the interrogator to take advantage of the difference. Likewise, he argues that it is difficult to produce laws of human behaviour but, regardless, being regulated by some laws of behaviour does not imply being some sort of machine. Therefore, he seems to support the informality of behaviour, natural behaviour that is. Finally, he considers extrasensory perception and argues that while a human may have the ability to use a form of it, the machine will resort to randomness.

In a true Turing style, he attempts to describe a learning machine by raising yet another question, *can a machine be made to be supercritical*, and by using the skin-of-an-onion as an analogy to explain his make belief that the human mind is mechanical at every layer. He strides on to experiment with imitating, or building up to, the intellectual ability of an adult human mind from that of a child by considering their initial state of mind, education, and experience. He then famously presents his vision which is widely regarded as the birth of machine learning, i.e. that such evolution would require: inheritance, mutation, and selection, with survival of the fittest, punishments, and rewards driving the process. He argues that the teacher will be uninformed of how the learning process is progressing, although they may still be able to predict its behaviour. He takes a further step to explain that we should expect that the machine will do something that we cannot make sense of or which we think of as completely random because of a departure from completely disciplined behaviour. Through his vision of a learning machine, he had set up the foundations of modern machine learning.

2. ALAN TURING'S IMITATION GAME UPGRADE IN THE 21ST CENTURY

Back in Turing's time, AI posed no direct threat to humanity because people had no access to personal computers or smart phones or communication networks or social networks and back then all developments in AI were carried out mostly in research labs. Roll the clock forward to today's time and it is an entirely different world with people of all ages now having access to a home computer, a smart phone, the Internet and social networks, and AI-driven software are being pushed on ordinary people's devices. And there are plenty of examples to cite from recent times where AI-generated content has posed serious threats to unsuspecting people, from finance to security and as far as existence. Would Turing's Imitation Game still hold today? The objective remains broadly unchanged, i.e. an interrogator challenged to distinguish between a learning machine and a human. However, the challenge has become a lot tougher as both the environment and the technology in which the game is played has evolved to what Turing predicted 75 years ago. The interrogator is mostly an ordinary human being trying to determine directly, not through a third party, what might be a deepfake and what might be genuine and with the AI-generated artefacts now on people's devices not hidden behind a screen.

The accelerated development of AI-manipulated media has been evolving, in parallel, Turing's question in a new direction, i.e. 'how can an ordinary human interrogator determine whether a media artefact is real or fake?' Success with deepfake media relies heavily on disrupting human audiovisual processing [2], to influence attention and information absorption, and on exploiting human sentiment, to influence opinion formation. Social media is rife with deepfake videos that have been manipulated by machine learning to either swap one person's face for another or alter the person's face to make them appear to say something they have not said. Furthermore, deepfakes also undermine the genuity of authentic videos which depict a person performing a physical action or making a statement that everyone knows is true. At the same time, the same AI tools, which are running into hundreds, have been used to create counter tools with which to detect deepfakes, but these are often not accurate enough, even with the best of training, or not easily accessible. And tool inaccuracy will almost certainly result in interrogator inaccuracy and eventually mistrust for the tool. Motivated by the widespread success of wearable technology, some recent counter tool development efforts have taken a turn in the direction of human physiology metrics for discerning deepfakes, but early results are not at all encouraging. These latest tools are attempting to discern variations in facial blood flow, blood pressure, skin temperature, pulse, etc., as evidence of AI manipulation, but currently, this is proving extremely challenging.

Expert media forensic analysts will have easy access to mainstream counter tools to support or complement their own professional experience in determining the authenticity of a deepfake suspect by looking for any changes at pixel and sound bite levels, distortions in the synchronization of audio and video and tracing the original source of each. An ordinary interrogator may not have access to such software tools, let alone have the technical know—how to use such tools and be able to discern when the tool is producing results that, statistically, are more likely to be inaccurate.

When faced with the dilemma, deepfake or not, the interrogator, quite paradoxically, will need to rely on the one hand, on the same audiovisual cues that the deepfake is trying to disrupt and, on the other hand, the sentiment that the deepfake is trying to exploit. For the former, the interrogator will need to consider inconsistencies such as overall face expression, mouth movement, lip-syncing, speech emotion, shadows, reflections, and distortions alongside the quality of the media that is usually intentionally reduced to hide or blur any cues that may help the interrogator and aid with disrupting human perception. Introducing, for example, graininess, blurriness, darkness, a flickering face, more than one person, a floating distraction, a person with a dark skin, in a suitable style, will only fuel the disruption. Trickery may also be introduced to reduce the interrogator's confidence in their ability to correctly identify a deepfake, for example, inversion (upside down media), misalignment (face split horizontally), and occlusion (black bar over eyes). Ultimately, the interrogator's visual perception might be the only tool at their disposal. When their visual perception suggests a real video, then the next test for the interrogator is to establish whether the message presented is true or misinformation.

Therefore, for the latter, the interrogator will need to be aware that the existence of a deepfake on social media does not necessarily guarantee truth and accuracy and, therefore, the entire content or part of it may be misinformation adapted to exploit sentiment and achieve an end purpose. Once emotionally compromised, e.g. becoming angry or anxious, the risk of rational judgement also being compromised increases, thereafter. Crowd wisdom which proliferates on social media is being increasingly used as a source for intelligence on deepfakes, but often even such an aggregate wisdom may be counterproductive. The interrogator will rely on several factors to decide whether what they are seeing and hearing is not misinformation: context, personal knowledge, ability to reason critically, capacity to learn, and accept updates on personal beliefs.

3. GAME HINTS FOR TURING'S IMITATION GAME LATEST UPGRADE

We have already seen how Deepfakes have been used to defraud people, to put their personal and online safety at greater risk, and even endanger their personal well-being and health. When faced with the Turing question, i.e. is it real or deepfake, the most natural reaction is to take a view on the spot or at least consider, often at a great risk. The research communities around the world have been building innovative new AI tools to help detect such deepfakes, but questions remain over their accessibility and effective use by inexperienced ordinary people. An alternative option would be a strategy for building public awareness of Deepfake technology and most importantly game hints for empowering ordinary people to think more critically before they proceed to consume what might be a possible deepfake. Building up the knowledge and skill of ordinary people to detect algorithmic manipulations will increase their ability and intuition to recognize such video manipulations [3]. How does an ordinary person recognize a video that has been algorithmically altered by AI? What are the game hints?

- Standard video quality featuring blurring images and misaligned body parts.
- Standard audio quality featuring digital-sounding voices with digital background noise.
- Facial transformations with unmatching, or lack of, emotion.
- Facial hair, or lack thereof, that looks unnatural.
- Facial birthmarks whose colour, texture, and shadows look unnatural, or which are missing.
- Eye movement and blinking frequency is excessive or out of sync.
- Eyes and eyebrows shadows appear in unexpected places.
- Eye-glasses glare is excessive and the angle of glare changes when the wearer moves.
- Lip movements are out of sync with the voice sounds.
- Teeth outline is unnatural.
- Cheeks and forehead skin is too smooth or too wrinkly and its age is inconsistent to hair and eyes.
- Body shape, posture, or movements are unnatural.
- Image and audio searches unravel earlier versions that are inconsistent with the current.
- Message conveyed is questionable, unexpected, and unverified.

Developing convincing deepfakes is becoming a new art and while there are many tools to detect algorithmic manipulation and plenty of hints such as the above to help discern deepfakes, people still fall victim to coercion by deepfakes as the quality of the manipulation and the misinformation is alarmingly very convincing. Turing predicted 75 years ago that as machine learning improves, the average interrogator's chances of making the right identification after 5 min will not be more than 70% and we have been witnessing on a regular basis how true his predictions have become recently. When the next video emerges featuring a politician making outrageous claims or a finance expert advising on outlandish offers or a celebrity in unflattering looks or poses then applying the Turing test might be the better response rather than a knee jerk reaction.

Websites such as Deepfake Detection Challenge [4] and Detect Fakes [5] that curate high-quality deepfake and real videos have recently begun emerging to assist with training ordinary people to discern deepfakes.

SUPPLEMENTARY DATA

Supplementary data is available at *The Computer Journal* online.

FUNDING

None declared.

DATA AVAILABILITY

No new data were generated or analysed in support of this research.

REFERENCES

1. Turing, A.M. (1950) Computing machinery and intelligence. *Mind*, **LIX**, 433–460. https://doi.org/10.1093/mind/LIX.236.433.
2. Groh, M., Epstein, Z., Firestone, C. and Picard, R. (2021) Deepfake detection by human crowds, machines, and machine-informed crowds. *PNAS*, **119**(1), e2110013119.
3. *Detect Deepfakes: How to counteract misinformation created by AI*. https://www.media.mit.edu/projects/detect-fakes/overview/
4. Deepfake Detection Challenge. https://paperswithcode.com/dataset/dfdc/
5. Detect Fakes. https://detectfakes.kellogg.northwestern.edu/