The Reverse Turing Test: Being Human (is) enough in the Age of AI

Fatemeh Alizadeh[1], Aikaterini Mniestri[1], and Gunnar Stevens[1]

[1] *University of Siegen, Siegen, Germany*

Abstract

Disposing of bad actors on social media is a daunting task, particularly in the face of "engineered social tampering" [4]. That is what Ferrara et al. [6] have labeled the rise of social bots, and large platform owners are struggling to mitigate the harmful effects caused by such malicious software. Therefore, it is no surprise that platform owners like META are fastening their security controls and that the popular press has tracked the efficacy of these measures. Specifically, META has been implementing what Forbes' Lance Eliot named the 'Upside Down Turing Test.' [26]. Unlike the original Turing test, which tasked a human participant with distinguishing a human from a digital speech correspondent, this version is designed to use a software program to distinguish non-human activity on the platform. In this work, we discuss the complications introduced by this reversal taking the human user's perspective. On the one hand, we recognize the necessity for fraud detection and defense against web-automated attacks. On the other hand, we find it necessary to uplift the voices of users who are wrongfully made victims as a result, in minor or major ways. At the same time, we offer alternatives to these invisible Reverse Turing Tests (RTTs) that expand the scope for distinguishing between human and non-human actors, while keeping humanity at the forefront of this inquiry.

Keywords

Reverse Turing Test, CAPTCHA, Bot detection, User-centered design

CEUR-WS.org/Vol-3136/paper-7.pdf

1. Introduction

The advent of the age of computers set forth a new horizon of exploration for artificial intelligence (AI). Ever since, researchers and industry have dedicated ample resources to investigate the question: Can machines think? Famously, the Turing Test, originated in 1950 by Alan Turing, was meant to assess the extent to which a digital computer could 'pass' as human through a knock off of the three-person party pastime, the "imitation game." [27]. In Turing's version, a human evaluator is meant to judge natural language conversations between a human and a machine developed to generate human-like responses. To be clear, Turing does not claim to answer the original question of whether machines are indeed, capable of thinking. Instead, he asks: can machines do what we (as thinking entities) can do? Thus, he not only does he distinguish between the physical and intellectual capacities of a man, as he proposed, but also sets the precedent for this workshop paper by emphasizing that machinic 'thought' ultimately serves a purpose for human users. That is to say, when the tables turn and computers are tasked with recognizing human from non-human behavior, humanity itself should be a part of this equation in all of its diversity and complexity. In this workshop paper, we pay respect to the evolving discourse around the Reverse Turing Test (RTT) [1, 4], the concept that a machine is capable of recognizing human behavior in digital systems, to critique recent attempts to distinguish users from

Proceedings of CoPDA2022 - Sixth International Workshop on Cultures of Participation in the Digital Age: AI for Humans or Humans for AI? June 7, 2022, Frascati (RM), Italy

EMAIL: Fatemeh.alizadeh@uni-siegen.de (F. Alizadeh); Aikaterini.Mniestri@student.uni-siegen.de (A. Mniestri); Gunnar.stevens@uni-siegen.de (G. Stevens)

ORCID: 0000-0002-5365-4695 (F. Alizadeh); 0000-0002-7785-5061 (G. Stevens)

© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

social bots on Instagram. For the purpose of this paper, social bots refer to computer algorithms that automatically produce content and interact with human actors on social media, trying to emulate and possibly alter their behavior [4]. Social bots have populated social media platforms for the past few years. We seek to understand the arguments for reverse Turing tests (RTT) in the current social media landscape, particularly in relation to the invisible bot detection algorithms referred to by Google as reCAPTCHA (Completely Automated Public Turing Tests to Tell Computers and Humans Apart), and we counter them with academic literature that centers on users' negative experiences with these algorithms. As a result, we rely on literature from New Media and Human Computer Interaction (HCI) to propose alternatives to standardized reCAPTCHAs that are more inclusive and user-centric.

This paper comes together in five sections. To begin with, we briefly provide some historical context by presenting an abbreviated history of the RTT. We then go into more detail about reCAPTCHAs and practices of invisible user tracking, and then specifically address the RTT algorithms enforced by Instagram. This leads us to the following section, which explores the user perspective on these security practices. We focus on users' gripes with these algorithms, which show that Instagram's fight against non-human interaction has real consequences for human users. As a result, we offer a series of alternatives, delineated in recent literature from the fields of HCI and New Media positioning ourselves firmly on the side of the user while also still acknowledging the need to maintain social media free of bad actors.

1.1. History of the Reverse Turing Test

The idea of repurposing the Turing test for curbing undesirable online interactions, was first proposed in an unpublished draft by Naor, who outlined a solution to the growing problem of online spam clogging free services such as email [20]. Naor suggested that users requesting access should first be given content identification tasks, such as "gender recognition" or "nudity detection" in images, to distinguish them from malicious actors. In the late 1990s, these tasks could easily confuse the relatively unsophisticated bots, but were "unambiguous" to humans [19]. A year later, the first practical example of a CAPTCHA scheme was developed to safeguard an online poll in advance of an upcoming presidential election [11]. The goal of CAPTCHA was to prevent the manipulation of election results by creating distorted texts and distinguishing humans from machines [15]. Building upon this, Baird et al. introduced RTT, a test in which "*a program challenges the user with **one synthetically generated image of a text** and the user must type the **text** correctly to pass*"[3]. This definition of RRT is based on the premise that text recognition tasks can overwhelm sophisticated bots and distinguish them from human users. However, in light of new advanced bots that solve even the most difficult traditional CAPTCHAs with a 99.8% hit rate [10], this premise is no longer tenable.

To address this issue, Google introduced the No CAPTCHA reCAPTCHA system, which is based on an advanced risk analysis engine that takes into account how users interact with CAPTCHA verifications [12]. Users only need to check the "I'm not a robot" box and will be verified without having to solve a CAPTCHA, if the risk analysis engine determines that the user is human. Otherwise, they are presented with an image-based challenge or a traditional text-based CAPTCHA to verify their identity. Previously, researchers have pointed out the shortcomings of this system (e.g., [2, 24]). For example, such a system can easily be tricked into thinking a program is human, or can become ineffective when Google web cookies are deleted or JavaScript is disabled [2]. Therefore, Google has recently improved this mechanism and made it completely invisible. The new "Invisible reCAPTCHA" service is based on the same technology as "No CAPTCHA reCAPTCHA", only without the "I'm not a robot" checkbox. Instead, Google "*invisibly analyzes the way users navigate through a website and assigns them a risk score based on how malicious their behavior is*" [12]. Invisible reCAPTCHAs eliminate the disruption caused by challenging CAPTCHAs, creating a more positive experience for users. As a result, more and more organizations are deploying invisible audits of activity on their platform.

Therefore, with this evolution of bots and AI technology, the construct of RTTs has also evolved from simple text recognition CAPTCHAs to complicated AI-invisible behavior tracking and discrepancy assessment between human and machine behavior [9]. In the following section, we take a closer look at these new forms of invisible RTTs to detect non-human activities on social media platforms and their consequences.

1.2. New forms of the Reverse Turing Test

Social bots populate social media platforms (Facebook, Twitter, Instagram etc.) en masse [1]. According to a 2018 study by Ghost Data, nearly 95 million Instagram accounts are automated [13]. In 2016, bots produced more internet traffic than humans [5]. Social bots have been used to undermine political discourse, manipulate the stock market, steal personal data, and spread fake news [6]. Therefore, detecting social bots is an important research goal. But social bots are becoming more skilled by the day, and the line between human-like and bot-like behavior is becoming more blurred [1, 13]. Today, they engage in more complicated types of interactions, whether discussing with other humans, commenting on their posts, or responding to their questions [6]. What makes the problem worse is that fake engagement is not only caused by bots, but also by fake accounts. Unlike bots, fake accounts do not improve their own metrics but those of other users, creating an unhealthy and inorganic environment [1]. Because of the increasing number of bots on Instagram, the company is accused of not doing enough to detect them [13].

To address these challenges, Instagram uses automated bot detection systems [18, 22] or "*automated technologies*" that help "*ensure the functionality and integrity of the service*," as stated in Instagram's Terms of Use [25]. Although the company does not disclose the inner workings of such bot detection systems, ostensibly to protect the systems from malicious actors, these mechanisms have been shown to use AI-based technologies (also referred to as reCAPTCHA) to invisibly track user interactions and apply necessary authenticity measures when a pattern of potentially inauthentic behavior is detected [18]. Authenticity measures on Instagram range from asking users to confirm their accounts, to temporarily blocking their actions, to permanently disabling accounts [14]. As useful and necessary as these mechanisms are, they are not without cost. Previous studies have shown that in some cases human users are blocked due to an error in these mechanisms (the so-called false positive error)[22], with no way to challenge or change the algorithmic decision. Considering that social media echo systems are already pointing in the direction of environments where interaction between machines is the norm and humans are navigating a world that is predominantly populated by them, the application of such bot detection mechanisms raises several concerns about the potential consequences of such control for users and society at large, as well as the redefinition of the term *human* as opposed to *machine* in future social media environments.

1.3. The risks and consequences of the invisible RTT

Judging by the statements of major platform owners, the new generation of bot detection and suppression algorithms does not compromise the usability of social media platforms. Google introduces its latest associated product, reCAPTCHA Enterprise, with the advertisement that it is "*a frictionless user experience, where fraud protection is easily extended across websites*" [23]. However, despite their effectiveness, these seamless and invisible mechanisms still raise a number of issues under the General Data Protection Regulation (GDPR). Mainly because they collect loads of personal data from users without proper notification and consent [8].

Introducing new measures to counter coordinated inauthentic behavior, Instagram users are also led to believe that the transition to reCAPTCHA is only to bring benefits for them both in terms of security and usability [14]. Nevertheless, as time goes by, it seems like this promise has not been fulfilled on two levels. Firstly, users still struggle with the automatic bot detection mechanism, and secondly, there are bad actors causing friction associated with the authentication processes within platforms after the fact: on Instagram, the uphill battle begins once a user loses access to their account. Once the algorithm has flagged their content, users, especially those who entrepreneurialise through the platform in the form of promotions and sponsorship deals, are approached by 'dealers.' These dealers offer creators an opportunity to recover their accounts for a price. Whereas some of them are merely scammers, others inexplicably have enough access in Instagram's moderation infrastructure to restore one's account and 'right' any algorithmically decided wrongs [28].

Last but not least, Simone Natale asserts that AI scientists "*have incorporated knowledge about users into their efforts to build meaningful and effective interactions between humans and machines.*"[21]. That is to say, the author implores for us to understand deception as an integral part of AI technologies. So, if we reverse the roles, in an RTT the computer is also encouraged to assume deception, to always be suspicious that there is no human end user on the other side but rather, another software, and malicious software at that. This is a sentiment that is built into the authentication systems of dominant social media platforms like Instagram. Their 2020 statement on authentication reads: *"If we see signs of potential inauthentic activity, we will require the account holder to confirm who they are, and once an account holder verifies their information, their account will function as usual unless we have reason to investigate further.*" [14]

This incriminates the user and puts the onus on them to prove that they are not guilty. Only, in this case, not being guilty means that they must prove their own humanity. Similarly, Myers West argues that "*automatic bans presume that users both intended to break the rules and are thus unable to learn how to do better*" [18]. She points out that the appeals process on Instagram fails to educate users on the reason for their misconduct and deprives them of the opportunity to reclaim their agency. In fact, users experience a range of negative emotions when they challenge the bot detection algorithm's decision, including anger, frustration, and, more importantly, dehumanization.

1.4. The alternatives to invisible reCAPTCHA bot detection

As mentioned above, the invisibility and seamlessness of reCAPTCHA and bot detection algorithms raise serious doubts when it comes to the fairness, transparency and accountability of such systems, as well as users' rights and control over the data being collected. In this section, we review some potential alternative approaches from the literature that have been introduced to address the challenges posed by the invisibility of reCAPTCHAs.

Frischmann [7] proposed a first set of human-focused RTT to investigate different aspects of intelligence and distinguish humans from bots. He refers to common sense, and rationality as human characteristics that can be used for generating RTT based on the notions of what it means to be human. He argued that a common sense (rational) test "*could employ a structure similar to the conventional Turing test, and the observer could ask questions that would require the skillful use of common sense (rationality)*." By addressing human bias and judgement errors, especially when there is not one "right" answer, he touches on a similar idea as using humans' perception capabilities for generating distorted text CAPTCHAs.

In a similar vein, Massey [17] argues for an approach to AI tests that is based on the discipline of aesthetics rather than technology. In his work, he proposes the 'metaphor to nonsense' transition. He proposes that the human mind can read metaphorical meaning into what is essentially a nonsensical phrase. He then argues that, even if a computer is trained to recognize metaphors, it will not be able to make the transition to viewing a particular metaphor as nonsense, at will. There is no logical transition from the metaphor to nonsense, which means that the AI would be faced with an insurmountable adversarial object. This paper does not merely center the user, but the humanity of the user.

However, despite the aforementioned promising proposals for human-focused RTTs and the concerns raised by the new forms of RTTs, research on alternative RTTs has so far remained at a conceptual level. In the following section, we look at this challenge through an HCI lens to see how these concepts can be operationalized in practice.

1.5. Towards a user-centered RTT

If the question raised by RTTs is how AI can be up to the task of distinguishing human activity from non-human activity on the Internet in a user-friendly manner, then the main idea of invisible reCAPTCHA and bot detection algorithms as a solution is to eliminate the question rather than provide an effective answer. Mainly because users, as the main affected parties of algorithmic decision making, have the right to be aware that they are being evaluated by an algorithm[29]. Therefore, the current

invisible RTT mechanisms does not meet the requirement of transparency and are also not in compliance with the GDPR. As an alternative, we propose to apply the HCI method by involving users in a user-centered design (UCD) cycle. As "*a multidisciplinary design approach based on the active involvement of users to improve the understanding of user and task requirements*" [16], UCD allows users not only to be aware of the test, but also to play an active and participatory role in the testing process. As we saw in Section 1.4, the human essence of users can inspire new creative ways to approach the RTT design process that leverage the aesthetic, intellectual capabilities of humans. We therefore propose the application of UCD to these techniques in order to put them into practice and create new forms of testing that are as diverse as they are potentially fun.

References

[1] Akyon, F.C. and Kalfaoglu, M.E. 2019. Instagram fake and automated account detection. *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)* (2019), 1–7.

[2] Al-Fannah, N.M. 2019. Using Aesthetic Judgements to Distinguish between Humans and Computers. *arXiv:1704.02972 [cs]*. (Jul. 2019).

[3] Baird, H.S. et al. 2003. Pessimalprint: a reverse turing test. *International Journal on Document Analysis and Recognition*. 5, 2 (2003), 158–163.

[4] Boshmaf, Y. et al. 2013. Design and analysis of a social botnet. *Computer Networks*. 57, 2 (2013), 556–578.

[5] Efthimion, P.G. et al. 2018. Supervised machine learning bot detection techniques to identify social twitter bots. *SMU Data Science Review*. 1, 2 (2018), 5.

[6] Ferrara, E. et al. 2016. The rise of social bots. *Communications of the ACM*. 59, 7 (2016), 96–104.

[7] Frischmann, B.M. 2014. Human-focused Turing tests: A framework for judging nudging and techno-social engineering of human beings. *Cardozo Legal Studies Research Paper*. 441 (2014).

[8] GDPR & Recaptcha: How to stay compliant with GDPR: 2021. *https://measuredcollective.com/gdpr-recaptcha-how-to-stay-compliant-with-gdpr/*. Accessed: 2022-03-27.

[9] Gonzalez, A.V. and Søgaard, A. 2020. The reverse turing test for evaluating interpretability methods on unknown tasks. *NeurIPS Workshop on Human And Machine in-the-Loop Evaluation and Learning Strategies* (2020), 62.

[10] Goodfellow, I.J. et al. 2014. Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks. *arXiv:1312.6082 [cs]*. (Apr. 2014).

[11] Guerar, M. et al. 2021. Gotta CAPTCHA'Em all: a survey of 20 Years of the human-or-computer Dilemma. *ACM Computing Surveys (CSUR)*. 54, 9 (2021), 1–33.

[12] Guerar, M. et al. 2018. Invisible CAPPCHA: A usable mechanism to distinguish between malware and humans on the mobile IoT. *computers & security*. 78, (2018), 255–266.

[13] Instagram's Growing Bot Problem: *https://www.theinformation.com/articles/instagrams-growing-bot-problem*. Accessed: 2022-03-26.

[14] Introducing New Authenticity Measures on Instagram: *https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram*. Accessed: 2022-03-26.

[15] Kochanski, G. et al. 2002. A reverse turing test using speech. (2002).

[16] Mao, J.-Y. et al. 2005. The state of user-centered design practice. *Communications of the ACM*. 48, 3 (2005), 105–109.

[17] Massey, I. 2021. A new Turing test: metaphor vs. nonsense. *AI & SOCIETY*. 36, 3 (Sep. 2021), 677–684. DOI:https://doi.org/10.1007/s00146-021-01242-9.

[18] Myers West, S. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*. 20, 11 (2018), 4366–4383.

[19] Naor, M. 1996. Verification of a human in the loop or Identification via the Turing Test. *Unpublished draft from http://www. wisdom. weizmann. ac. il/~naor/PAPERS/human abs. html*. (1996).

[20] Naor, M. and Shamir, A. 1995. Visual cryptography. *Advances in Cryptology — EUROCRYPT'94* (Berlin, Heidelberg, 1995), 1–12.

[21] Natale, S.L. in C. and M.S.S. and Natale, S. 2021. *Deceitful Media: Artificial Intelligence and Social Life After the Turing Test*. Oxford University Press.

[22] Rauchfleisch, A. and Kaiser, J. 2020. The false positive problem of automatic bot detection in social science research. *PloS one*. 15, 10 (2020), e0241045.

[23] reCAPTCHA: *https://www.google.com/recaptcha/about/*. Accessed: 2022-03-27.

[24] Sivakorn, S. et al. Sivakorn: I'm not a human: Breaking the Google reCAPTCHA.

[25] Terms of Use | Instagram Help Center: *https://help.instagram.com/478745558852511*. Accessed: 2022-03-27.
[26] The Famous AI Turing Test Put In Reverse And Upside-Down, Plus Implications For Self-Driving Cars: *https://www.forbes.com/sites/lanceeliot/2020/07/20/the-famous-ai-turing-test-put-in-reverse-and-upside-down-plus-implications-for-self-driving-cars/*. Accessed: 2022-03-26.
[27] Turing, A.M. 2009. Computing machinery and intelligence. *Parsing the turing test*. Springer. 23–65.
[28] Waters, S. I'm Not a Robot! So Why Won't Captchas Believe Me? *Wired*.
[29] Art. 22 GDPR – Automated individual decision-making, including profiling. *General Data Protection Regulation (GDPR)*.
"ReCAPTCHA." *reCAPTCHA*. https://www.google.com/recaptcha/about/ (March 27, 2022).