

Enhancing Human Verification Systems to Address AI Agent Circumvention and Attributability Concerns

By: Yogev Angelovici, Anish Ganga, Saathvik Kannan, Zihao Zhou

Table of Contents

Table of Contents	2
Introduction	3
Implementation	3
Technical Implementation	3
Policy Implementation	3
Feasibility	3
Technical Feasibility	3
Policy Feasibility	3
References	3

Introduction

In a digital landscape increasingly shaped by automation, AI has transformed from a fringe technology into a critical infrastructure for many online services. This integration is largely due to revolutions in artificial intelligence capability in fields such as machine learning, which enable machines to mimic human actions increasingly, simultaneously limiting the ability to distinguish between AI and human actions [Rutschman, 2019]. This is particularly concerning when it comes to the concept of attributability. Without a clear mechanism to understand whether an action is carried out by a human or an AI, the attribution of an action being taken isn't possible. For example, suppose an AI agent were to make a decision violating a law under the guise of a human making that decision. In that case, there is no proper way to provide recourse in the legal system because there is no verification system that delineates between an AI agent and a human.

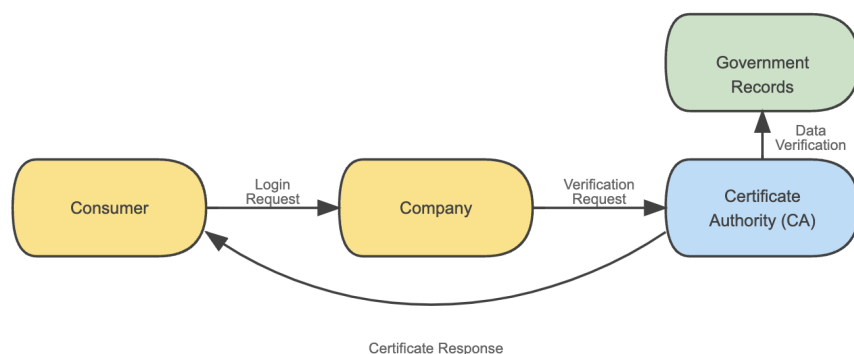
This problem's fundamental challenge is rooted in distinguishing between human users and automated entities. Since its inception in 2000, CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) has served as a critical digital gatekeeper, intending to distinguish humans from robots and thwart any automated attempts to exploit web services [Burling, 2012]. As AI is increasingly able to replicate human behavior, the question isn't whether a CAPTCHA will stop an agent but when will the agent be able to surpass the CAPTCHA.

The initial solution of the CAPTCHA industry to the increasing capabilities of AI to imitate human behavior and capabilities was to release harder user questions, which have taken more time to fulfill and caused more user friction [Searles & Tsodik, 2023]. However, multiple studies have continued to show that there is little to no difference between the ability of users to pass these benchmarks, or even surpass humans at these tests. A recent 2024 study showed that 100% of modern reCAPTCHA v2 tests can be solved with AI capabilities [Plesner et al, 2024]

To address this gap, a fundamental shift in perspective is required—an inversion in the way we perceive and handle AI. Rather than simply trying to outsmart ever-improving AI bots, we must adopt an approach that redefines the standards for human-AI differentiation and places a proactive identification and validation of human users as a priority over the disqualification of AI agents.

Implementation

Technical Implementation



Our proposed structure of this validation of human users is modeled off of the most comparable low trust situation on the modern internet, namely the certification of websites as trustworthy. The basic idea of website certification is that a website proves its trustworthiness to a Certificate Authority (CA), after which it is issued a

certificate containing a public and private key set. This certificate proves its trustworthiness to other web users, and the public and private key set can be used to prove that the certificate the website claims is theirs [Cloudflare, ND].

We can solve this by emulating the idea of a website certificate, instead of replacing the current CA standards of ‘trustworthiness’ with proof of human identity that can be claimed by human users. When a user needs to confirm their human identity, they upload a government-recognized identification document, which is then cross-referenced against a secure government database through an AI-enabled process. This process occurs through a third-party CA specifically designed for human identity verification. Once the ID is uploaded, the CA performs a real-time check against the relevant government databases to confirm the ID’s authenticity and the human identity it represents.

The verification system uses a recursive search across multiple databases, potentially extending beyond national boundaries. For example, if the user’s home country is not recognized or lacks adequate records, the system automatically extends the search to other national databases, which may be maintained by other countries’ respective certificate authorities. Each nation could maintain relations with its own set of CAs that it trusts to verify users against its records, enabling the verification system to probe databases in a cascading manner until a match is found or all options have been exhausted.

Once the verification is successful, the CA issues a digital certificate to the user, cryptographically proving their human identity. This certificate can be saved locally on the user’s device and remains valid for a specific period, after which re-verification is required. When visiting websites requiring verification, users simply present this certificate, which is then validated by the site’s backend with the CA for real-time human verification. This local certificate storage allows users to interact seamlessly across verified sites without repeatedly undergoing identity checks, streamlining the user experience while ensuring robust, AI-resistant human verification that adapts to global standards.

Policy Implementation

From a policy perspective, we have to understand how policymakers can encourage the creation and normalization of this verification system for the Internet. The current model and policy standards for website certificates were set back in the 1990s, when the internet desperately needed a solution for low-trust interactions on the internet and when every actor in the industry had the same incentive to create a shared system of proving trustworthiness. Policymakers need to recreate both the incentives to make a shared system as well as lower barriers to entry for companies to check for human verification.

This involves a two-pronged approach to implementation. The first prong of that approach is raising the incentives to adopt and implement robust verification systems. This can be achieved through financial incentives such as grants, tax breaks, or subsidies for companies that invest in advanced verification technologies. Additionally, establishing regulatory frameworks that recognize and reward adherence to standardized verification protocols will encourage businesses to prioritize trustworthiness and security. By making the adoption of these systems both economically beneficial and legally advantageous, policymakers can motivate a wide range of organizations to participate in the shared verification infrastructure.

The second prong focuses on fostering strong partnerships between the government and trustworthy CAs. Policymakers should collaborate with established CAs to ensure seamless integration of verification processes with government records. This partnership can involve creating secure channels for CAs to access and cross-reference government databases, enabling efficient and accurate identity verification. Additionally, the government can provide support through streamlined access to necessary data, clear guidelines for compliance, and resources to enhance the technical capabilities of CAs. By facilitating easy and reliable checks against government records, the government can lower the barriers for companies to implement human verification

systems. This collaborative approach not only enhances the credibility and effectiveness of verification processes but also ensures that the verification infrastructure is robust and aligned with national security and privacy standards. Through these partnerships, policymakers can drive the widespread normalization of verification systems, thereby strengthening trust and security across the internet when it comes to proving a user to be human.

With this implemented, the problem of attributability is shifted to focus on trying to validate the humanity of the user versus rooting the artificial intelligence out. If artificial intelligence can replicate human intelligence and continuously surpass checks previously thought as only possible to solve by humans, then the identification will always have to evolve because the tests will continue to be passed. However, by shifting the focus on utilizing human documentation to ensure that the user is human, then they can use this verification as access to actions requiring human verification. The impact can be seen in two different ways: attributability and role limitations. For attributability, by identifying when an illegal or harmful action is human-driven as opposed to AI agent-driven, the proper systems, and recourse can be taken based on the result of the test. If the illegal or harmful action was human-driven, this allows identification of the culprit and subsequent punishment under the legal system. While we can't necessarily punish AIs for illegal or harmful actions, this system allows governments to isolate whoever provided the AI access to a human-only system by checking the human verification certificate it used to access the website. This allows the handler of the model to be identified and held as culpable under the law. This also allows governments the power to subsequently shut down the model's access to human-only sites, amongst other options. For role limitations, the concept of AI agents is to carry out actions on behalf of humans but certain roles/actions may be designated as human only. For example, social media posting is a topic of big discussion due to the generation of false content that spreads misinformation. By requiring a guarantee of human status before acting such as posting, we can limit the ability of AI agents to carry out actions deemed unsafe at a mass scale.

Feasibility

Technical Feasibility

From a technical standpoint, the technologies needed to implement this verification system are already available and ready for use. The modern internet operates on low-trust scenarios using certificate frameworks, and this proposed system builds on that foundation by adding user identity verification. Instead of only servers proving their trustworthiness, the system allows servers to verify real and trustworthy users. The main technical challenges are integrating the verification process into existing login systems and enabling users to store their trust certificates securely.

These challenges can be addressed with straightforward solutions. Companies can incorporate the verification system into their login processes by using established authentication protocols like OpenID Connect and OAuth2, which allow seamless integration without disrupting the user experience. By embedding Certificate Authority (CA) checks into the login workflow, websites can validate user certificates efficiently. For users, secure certificate storage is manageable with built-in device features such as Keychain on macOS and iOS, TPM on Windows, and hardware-backed Keystore on Android. These secure storage options protect private keys and ensure certificates are accessible only when needed, providing both security and ease of use. Overall, the system can integrate smoothly with current sign-in methods while maintaining strong security and user accessibility.

Policy Feasibility

The verification system addresses a critical and growing challenge in our digital landscape. As AI systems become increasingly sophisticated in mimicking human behavior, traditional CAPTCHA-style verification methods are proving inadequate. The proposal builds upon established and proven infrastructure, particularly the Certificate Authority (CA) model that has successfully secured website verification for decades. The technical foundation is solid, utilizing existing secure storage mechanisms (TPM, Keychain, Keystore) and standard authentication protocols (OpenID Connect, OAuth2). The system's integration with government databases provides a robust verification mechanism that would be significantly harder to circumvent than traditional behavioral tests. It also offers a clear framework for accountability, enabling the tracking of digital actions back to verified human actors.

The system requires more international cooperation and database integration across jurisdictions, which may present technical and diplomatic hurdles. Privacy concerns may arise, as the system would create a repository of personal identification information, making it a potential target for cyberattacks. User adoption might face resistance due to privacy concerns and the friction introduced by the verification process. Additionally, the system might inadvertently exclude individuals without government ID or those from countries with limited digital infrastructure.

To further address these concerns, several key actions are needed:

- Develop a certificate lifecycle, including issuance, renewal, and revocation procedures.
- Create a pilot program within a single jurisdiction to test the system's effectiveness
- Establish international working groups to develop protocols for cross-border verification and data sharing.
- Draft privacy protection guidelines that address data storage, transmission, and access controls.

Conclusion:

As artificial intelligence continues to evolve and permeate the digital landscape, the challenge of distinguishing between human and AI-driven actions becomes increasingly critical. The proposed human identity verification system, inspired by the Certificate Authority framework, offers a comprehensive solution to ensure clear attribution and enhance online security. By implementing a standardized verification process that utilizes government-issued identification and secure digital certificates, policymakers can effectively address the limitations of traditional CAPTCHA methods and mitigate the risks associated with AI impersonation.

Adopting this verification system not only strengthens accountability by enabling precise attribution of actions but also fosters trust in online interactions by ensuring that users are genuinely human. The benefits are multifaceted, including enhanced legal recourse, reduced misinformation, and controlled access to sensitive digital platforms, thereby safeguarding both individuals and society at large.

However, the success of this initiative relies on collaborative efforts between governments, Certificate Authorities, and private sector stakeholders to establish and normalize these verification protocols. Policymakers must prioritize the creation of supportive frameworks that incentivize adoption and ensure robust privacy protections, facilitating widespread implementation and user acceptance.

It is imperative that immediate action is taken to implement this verification system, as the technological capabilities of AI continue to advance. By proactively addressing the challenges of attributability and reinforcing human verification mechanisms, we can secure a more trustworthy and resilient digital future. Let us commit to these strategic changes to uphold accountability, protect legal integrity, and maintain the authenticity of human interactions in the digital realm.

References

Burling, Stacey. *CAPTCHA: The Story behind Those Squiggly Computer Letters*.

<https://phys.org/news/2012-06-captcha-story-squiggly-letters.html>. Accessed 27 Oct. 2024.

“How Does SSL Work? | SSL Certificates and TLS.” *Cloudflare*,

<https://www.cloudflare.com/learning/ssl/how-does-ssl-work/>. Accessed 27 Oct. 2024.

Plesner, Andreas, et al. “Breaking reCAPTCHA v2.” *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2024, pp. 1047–56. *arXiv.org*,

<https://doi.org/10.1109/COMPSAC61105.2024.00142>.

Rutschman, Ana Santos. “Artificial Intelligence Can Now Emulate Human Behaviors – Soon It Will Be Dangerously Good.” *The Conversation*, 5 Apr. 2019,

<http://theconversation.com/artificial-intelligence-can-now-emulate-human-behaviors-soon-it-will-be-dangerously-good-114136>.

Searles, Andrew, and Gene Tsudik. “Observing CAPTCHAS ‘in the Wild.’” *USENIX*, 21 Nov. 2023,

<https://www.usenix.org/publications/loginonline/observing-captchas-%E2%80%9Cwild%E2%80%9D>.