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

Table of contents

01

Introduction

02

Implementation

03

Feasibility

04

Concluding Remarks

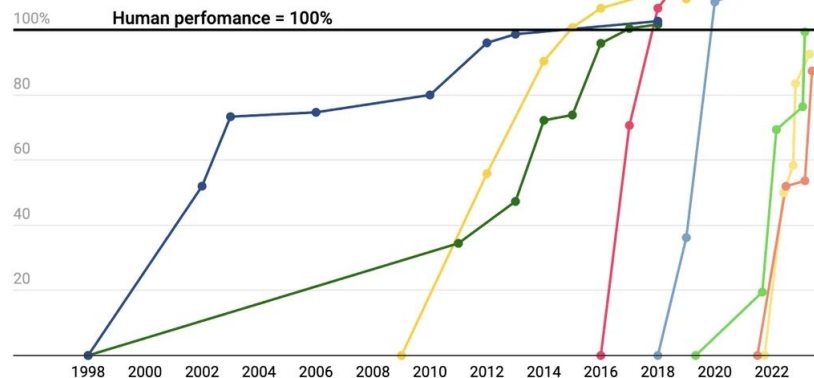
Introduction

- The increasing sophistication of AI has made CAPTCHA tests, designed to differentiate humans from bots, largely ineffective.
- A 2024 study found that AI can solve 100% of modern reCAPTCHA v2 tests, highlighting the need for new methods of human verification.
- We need a fundamental shift in perspective on how to evaluate between humans and AI, as AI continues to expand

AI has surpassed humans at a number of tasks and the rate at which humans are being surpassed at new tasks is increasing

State-of-the-art AI performance on benchmarks, relative to human performance

● Handwriting recognition ● Speech recognition ● Image recognition ● Reading comprehension
● Language understanding ● Common sense completion ● Grade school math ● Code generation



For each benchmark, the maximally performing baseline reported in the benchmark paper is taken as the "starting point", which is set at 0%. Human performance number is set at 100%. Handwriting recognition = MNIST, Language understanding = GLUE, Image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehension = SQuAD 2.0, Speech recognition = Switchboard, Grade school math = GSK8k, Common sense completion = HellaSwag, Code generation = HumanEval.

Chart: Will Henshall for TIME • Source: ContextualAI

TIME

02

Implementation

Technical Implementation

Modeled on Website Certification

- Users obtain digital certificates from a specialized Certificate Authority (CA) to prove human identity.

Secure Verification Process

- Upload government-issued ID.
- AI-enabled cross-references with global government databases.

Global Recursive Verification

- Searches multiple national databases through trusted international CAs.

Digital Certificate Issuance

- Certificate includes public/private keys.
- Stored locally, valid for a set period with periodic re-verification.

Seamless User Experience

- Present certificate for real-time verification on websites.
 - Ensures secure, streamlined interactions across verified platforms.
-

REAL BANK.

Verify your humanity with Real* first.

Verify

Experience seamless banking with Real Bank. Join thousands of satisfied customers who trust Real Bank for their banking needs.

Create your Account

Identity Verification

Enhanced Security Verification

1. Security Key Upload



Upload your security key file
Drag and drop or click to browse

2. Complete Verification

Username

Enter your username

Verification Message

Enter verification message

Complete Verification

Policy Implementation

- Policymakers need to encourage the creation and normalization of the human verification system by recreating incentives & lowering barriers for companies compliance
 - This involves a two-pronged approach:
 - First, raising incentives to adopt robust verification systems through financial means like grants, tax breaks, or subsidies, and reward adherence to verification protocols.
 - Second, policymakers should foster strong partnerships between the government and trustworthy Certificate Authorities (CAs) to ensure seamless integration of verification processes with government records, enabling efficient and accurate identity verification.
 - By facilitating reliable checks against government databases and providing streamlined data access, we can enhance the credibility and effectiveness of verification processes
-

Intended Consequences

- The proposed system could help to identify the human handlers responsible for AI agents that engage in harmful or illegal actions online.
 - If an AI agent is found to be responsible for harmful online behavior, the system can trace it back to the human who granted it access to the "human-only" online space.
 - The human handler could then be held accountable for the AI agent's actions.
 - A human verification system could also be used to restrict the actions of AI agents in certain online spaces.
 - For instance, AI agents could be prohibited from posting on social media, which would help to reduce the spread of misinformation.
 - Restricting AI agent actions in this way would require a guarantee of human status before allowing certain actions
-

03

Feasibility

Technical Feasibility

Built on Proven Frameworks

- Certificate-Based Trust: Utilizes the modern internet's existing certificate frameworks.
- Uses same decentralized network model as the current framework

Enhanced Verification

- Seamless Integration
- Authentication Protocols: Compatible with OpenID Connect and OAuth2 for easy incorporation.
- CA Checks: Embeds Certificate Authority checks into existing login workflows.
- Secure Certificate Storage

Example Device Integrations: for key verification:

- macOS/iOS: Keychain
 - Windows: Trusted Platform Module (TPM)
 - Android: Hardware-backed Keystore
-

Policy Feasibility

Clear and Robust framework:

- The proposed human verification system, modeled after website certification, 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.

Possible Drawbacks:

- privacy issues, the need for international cooperation, and risking excluding people who lack government-issued ID or live in areas with limited digital access.

Key areas moving forward:

- Creating procedures for certificate management, conducting pilot programs, establishing international collaboration for cross-border verification, and drafting guidelines for data privacy
-

04

Concluding Remarks

Conclusion: Securing a Trustworthy Digital Future

Implement Robust Human Verification

- Utilize government-issued IDs and digital certificates to distinguish humans from AI.

Enhance Accountability and Trust

- Ensure accurate attribution of actions and foster genuine online interactions.

Foster Policy Support and Collaboration

- Incentivize adoption, build partnerships with Certificate Authorities, and prioritize privacy protections.
-