Assessing Flaws in CAPTCHA Security through Progress in AI

Jaydon Stanislowski
University of Minnesota Morris

Outline

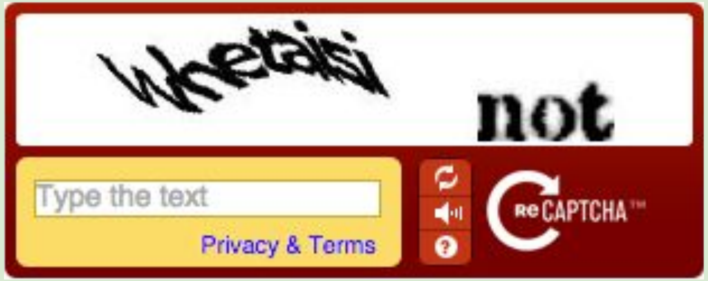
1. Introduction to CAPTCHAs
2. Modern CAPTCHAs
3. Reinforcement Learning
4. Attacking reCAPTCHA v3
5. Threat Analysis
6. Conclusion

1. Introduction to CAPTCHA

- What is a CAPTCHA?

- Motivation

- Early CAPTCHAs



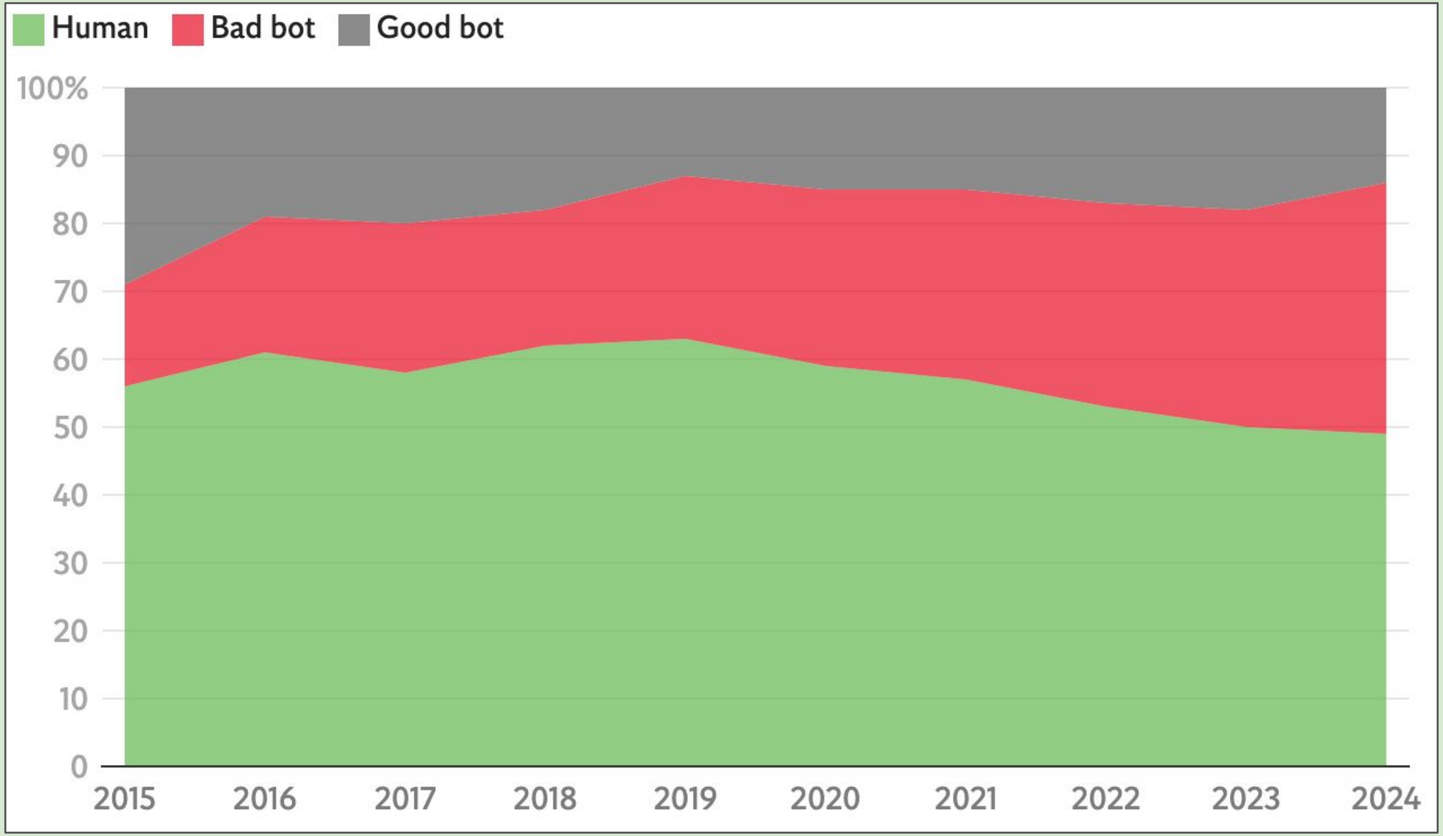
medical
cheese
bird
there
cheese
clear
cheese
bird
clear
dry

What is a CAPTCHA?

- **CAPTCHA** - Completely Automated Public Turing test to tell Computers and Humans Apart
    - **Turing test -** a thought experiment, measures machine intelligence with a human evaluator (Turing, 1950)
- Tasks designed to be simple for humans, but hard for AI models
- Designed by Luis von Ahn et al. in the early 2000s

Motivation

- Artificial web traffic can have various malicious motives:
  - Credential stuffing/brute force attacks
  - Fake account creation and engagement
  - Spam, extortion
  - Web scraping
- Despite this, it is more prevalent than ever

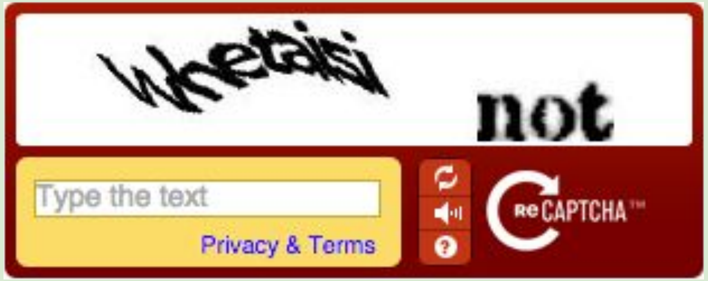


Area chart depicting the rise in artificial web traffic since 2015. | The Independent (adapted from Imperva 2025 Bad Bot Report)

Early CAPTCHAs

Text-based challenges:

- Participants must transcribe text
- Text is typically distorted to make it hard for machines to read
- Random noise, warping, rotating, etc.



medical
cheese
bird
there
clear
cheese
bird
cheese
clear
dry

Early CAPTCHAs

- **reCAPTCHA**: version 1 began as text-based
    - Developed by Luis von Ahn et al. in 2007
    - Acquired by Google two years later
- Other text-based frameworks included Gimpy, hCAPTCHA, etc.
- reCAPTCHA v1 deprecated in 2018. reCAPTCHA v3 is the latest version

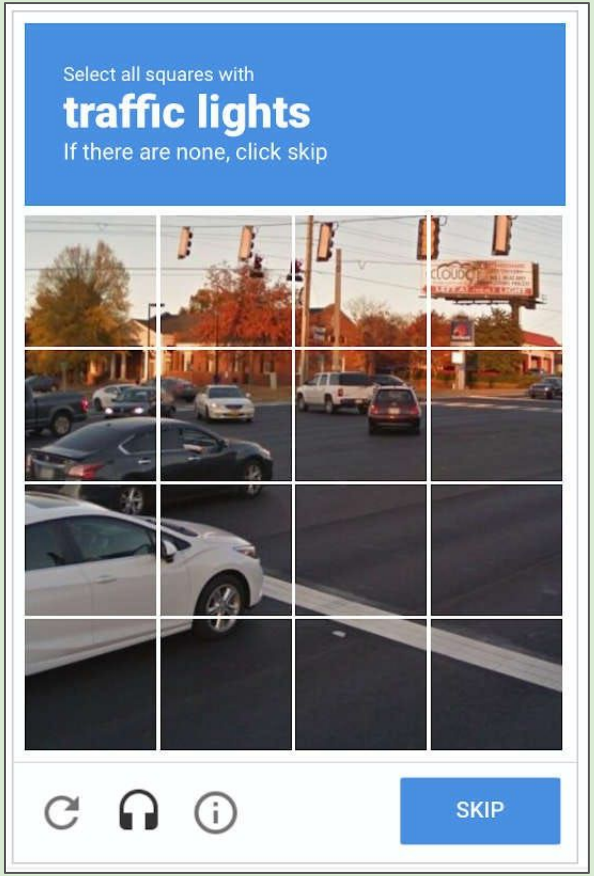
2. Modern CAPTCHAs

- Common Types
- reCAPTCHA v3
- Why reCAPTCHA?

Common Types

Many challenge types exist now, but are outside of the scope of this presentation

- Image-based challenges
- Audio-based challenges
- Spatial reasoning challenges



Select all squares with

traffic lights

If there are none, click skip

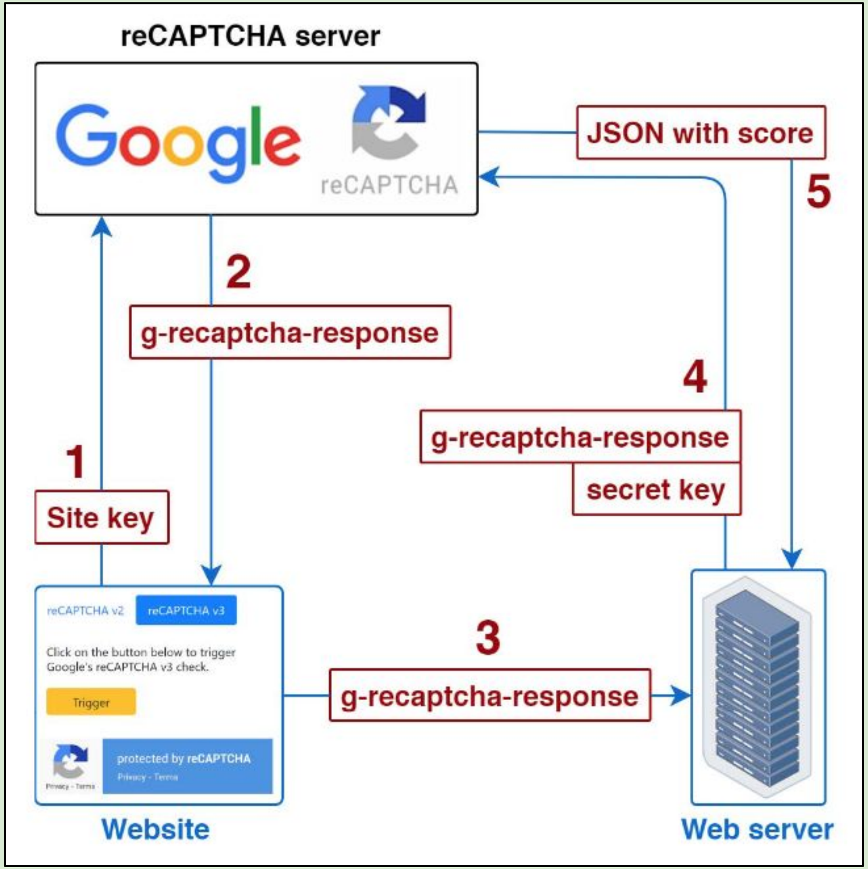
reCAPTCHA v3

- Intended to reduce friction for real users, removing the actual "challenge," entirely invisible
- Uses **behavior metrics**, calculates likelihood of a bot based on browser activity (cookies, inputs, etc.), directly embedded into site interactions
- Earlier version, reCAPTCHA v2, did this using a checkbox challenge (i.e. not invisible)

reCAPTCHA v3

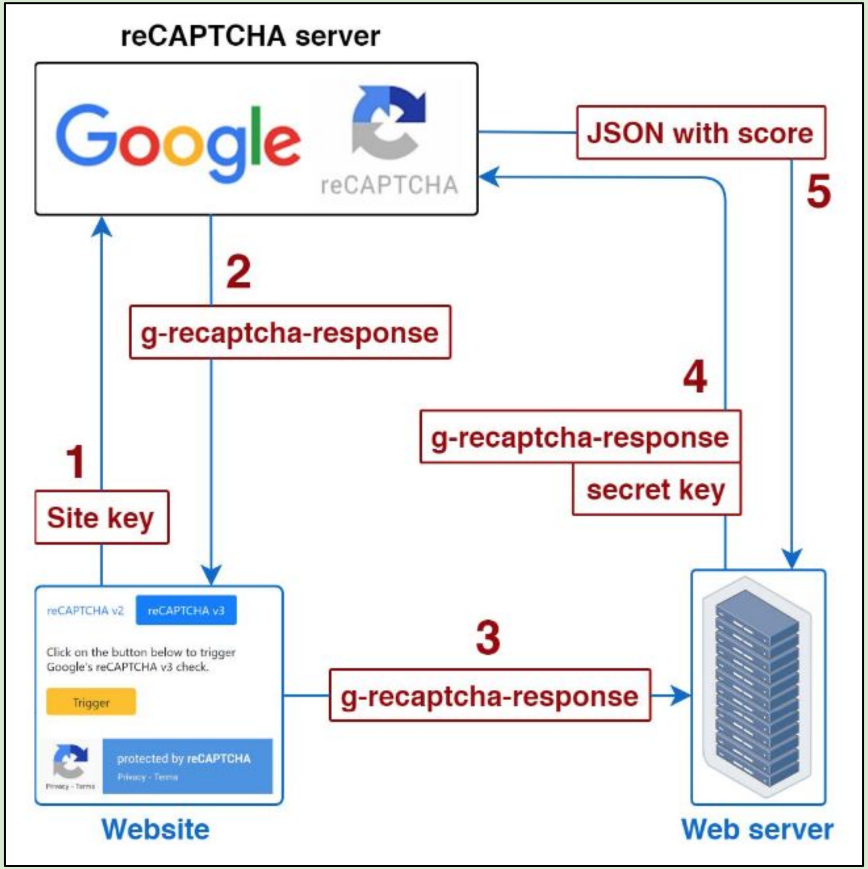
1. reCAPTCHA v3 is invoked, the website sends metrics and invocation context to Google's servers, collected by the tool
2. Google generates a token to be verified by the site

reCAPTCHA v3 workflow | Joosen et al.



reCAPTCHA v3

3. The token is sent to the web server
4. The server verifies the token with the site's key and sends this to Google
5. Google returns a formatted score, which can be used by the website to take necessary actions



reCAPTCHA v3 workflow | Joosen et al.

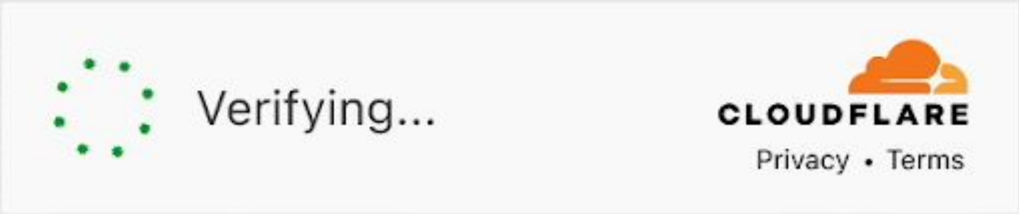
reCAPTCHA v3

Considerations:

- Code is proprietary, details are intentionally obfuscated
- Calculations may be randomly skewed to deter probing
- Score itself is not descriptive, only a discrete number {0.1, 0.3, 0.5, 0.7, 0.9}
- Lower score means more likely to be a bot

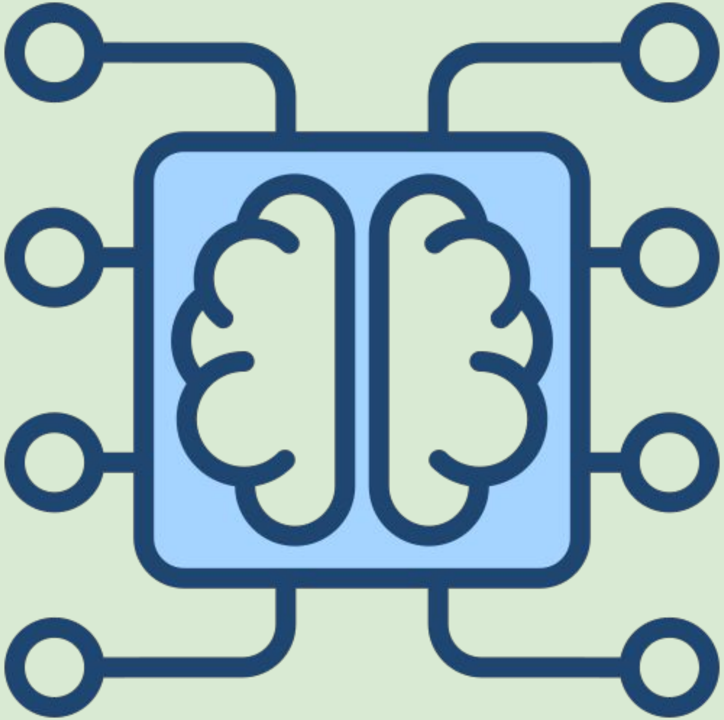
Why reCAPTCHA?

- Over 10 million websites as of 2025 (BuiltWith)
  - This includes GitHub, Reddit, Amazon, and more
- Designs of reCAPTCHA v3 are applied in other CAPTCHA frameworks
  - Notably, Turnstile and newer versions of hCAPTCHA



Verifying...

3. Reinforcement Learning

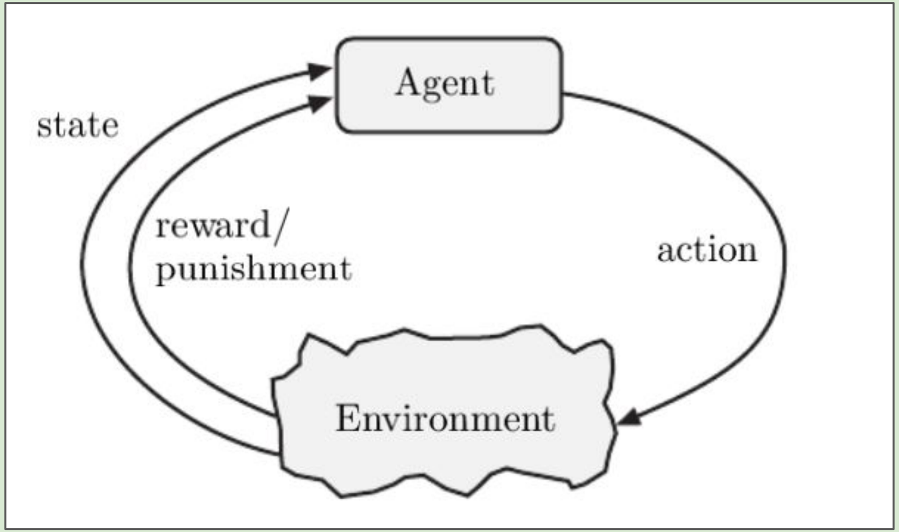


Overview

- Machine learning paradigm
- Employed when computers interact with an environment
    - Applications in robotics, social media algorithms, strategy games, etc.
- Agent is given a set of actions to do so, choices are initially random

Overview

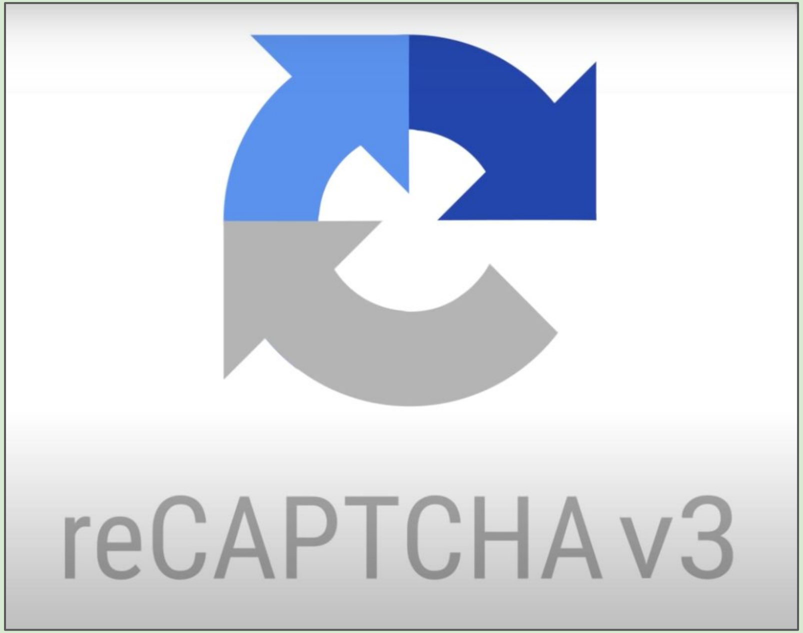
- The agent observes its current state to alter chance of making decisions
- Reward score influences these changes, agent must try to maximize it



Block diagram modeling basic principles of reinforcement learning | Tizhoosh, Taylor

4. Attacking reCAPTCHA v3

- Bypassing Behavior Metrics
- Results



Bypassing Behavior Metrics

- Browser activity that may be measured (Joosen et al., 2022):
  - **Static features**: presence of cookies, IP address, browser, operating system, etc.
  - **Dynamic features**: mouse and keyboard inputs, timings, request frequency
- reCAPTCHA v3 obscures how much these are measured

Bypassing Behavior Metrics

Static features:

- Sivakorn et al. (2016) determined static features to have a very strong positive influence on score
- In such cases, dynamic features have very minimal effect
- Goal of this research is to exploit dynamic features starting from a low score

Bypassing Behavior Metrics

Measuring dynamic features:

- Treating the score as an "oracle" — specific states of the environment aren't needed. Instead, fine tune behaviors based on the output
- The RL agent makes assumptions about its environment and learns accordingly

Bypassing Behavior Metrics

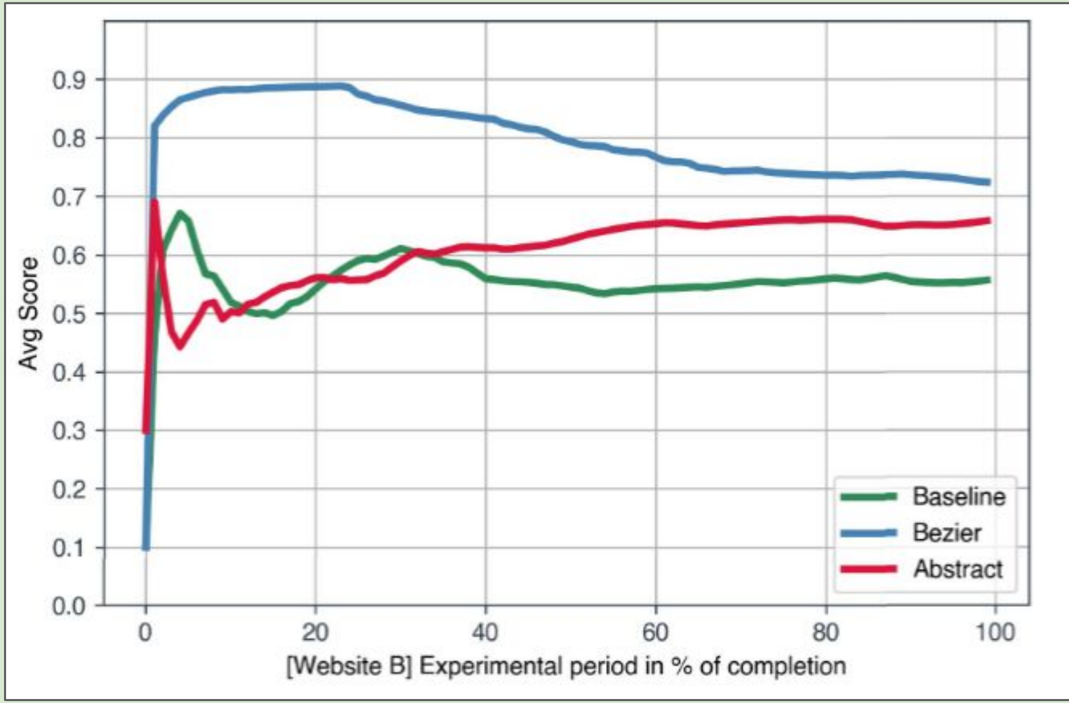
- Joosen et al. trained their bot on three websites:
  - **Website A**: hosted by researchers, implementing reCAPTCHA v3; training only
  - **Website B**: hundreds of daily requests, only recently deployed reCAPTCHA v3; partial training
  - **Website C**: thousands of daily requests, already fully integrated reCAPTCHA v3; full deployment

Bypassing Behavior Metrics

- Two modes of the bot were deployed in testing to assess score impact of dynamic features
- ***Bezier***: Can only move mouse in curved patterns and click, as well as time mouse inputs, hover duration, etc.
- ***Abstract***: Freedom over mouse movements and timings, plus keyboard inputs and scrolling

Results

- A baseline (naive) algorithm was used to assess score improvement via RL
- *Bezier* and *Abstract* deployed on B, with *Abstract* in its training mode



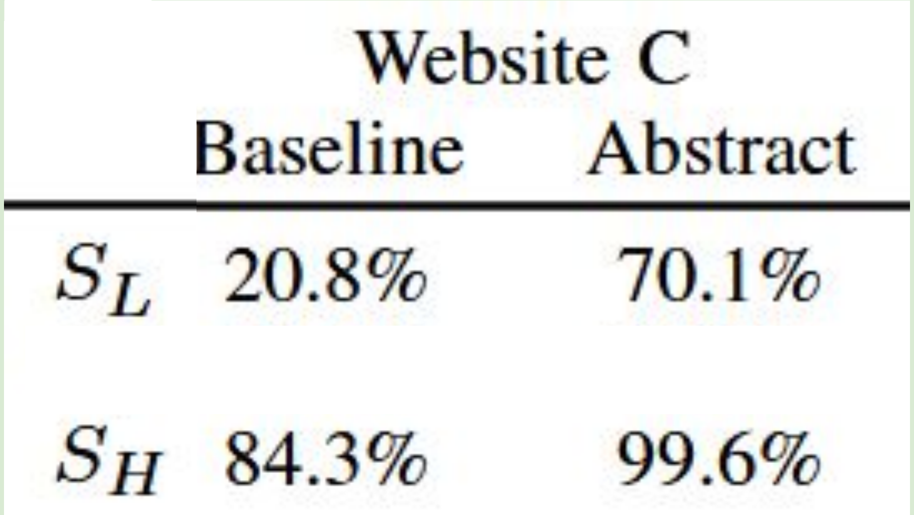
Average cumulative score over the period of evaluation for Website B | Joosen et al.

Results

- On website C, researchers did not have access to the score. Intended to test *Abstract*'s capabilities in a fully black-box environment
- Success measured by **evasion rate**, frequency of v3 seeming to detect no bot activity
- Evaluated for both high starting sessions (presence of static features) and low starting sessions

Results

- *Abstract* almost perfectly avoided detection with static features
- Starting from a low score with minimal static data, *Abstract* succeeded at a rate of ~70%



	Website C	
	Baseline	Abstract
S_L	20.8%	70.1%
S_H	84.3%	99.6%

Evasion rate (in %) for both the *Baseline* and *Abstract* algorithms across different session types | Joosen et al.

Analysis

- Research Implications
- Limitations
- Future of CAPTCHA Tools
- Ethical Considerations

Research Implications

- The results suggest that providing a CAPTCHA score is a vulnerability, can be exploited to:
  - Train machines to receive a passing score
  - Probe what behavior data is being measured

Research Implications

CAPTCHAs based on these metrics may be insecure as a whole:

- They are easily bypassed by models trained to mimic human browsing
- They are apparently biased in favor of static variables easily accounted for by attackers

Limitations

- reCAPTCHA v3 is known to intentionally add random noise, obscuring the meaning of the score
- Models cannot remain active over long periods of time, since excessive number of requests influences the score
- Not accounting for combining CAPTCHA schemes or using different metrics-based CAPTCHA tools

Limitations

- reCAPTCHA v3 likely employs adversarial learning
- More study of how the tool has changed over time is needed, could improve or remain the same
- This also makes past experiments harder to replicate

Future of CAPTCHA Tools

- CAPTCHA tools may pivot to more complex challenges for humans, possibly focusing more on spatial reasoning and logic puzzles (many do already)
- Current CAPTCHAs using behavior metrics may improve over time, but so will AI models to bypass them

Future of CAPTCHA Tools

- Von Ahn et al. designed CAPTCHAs with the intention of being broken
  - As AI evolves, breaking CAPTCHA schemes is proof of progress in the field
  - CAPTCHA will always change as breakthroughs occur, but for how long will this remain effective?

Ethical Considerations

Responsible disclosure:

- Researchers received permission from website owners before deploying their bots
- Google was notified of these security concerns. The issue was closed as intended behavior, likely considered to be a "reasonable limitation" of the tool

Conclusion

- CAPTCHA security is an important but persistent problem as AI becomes more sophisticated
- Today's Captcha tools face challenges that reduce their effectiveness, but will continue to grow over time

Questions?

References

[1] Von Ahn, L., Blum, M., Hopper, N., & Langford, J. (2003). CAPTCHA: Using Hard AI Problems for Security. International Association for Cryptologic Research 2003. https://doi.org/10.1007/3-540-39200-9_18

[2] Turing, A. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[3] Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science, 321(5895), 1465–1468. https://doi.org/10.1126/science.1160379

[4] Tizhoosh, H. R., & Taylor, G. W. (2006). Reinforced Contrast Adaptation. International Journal of Image and Graphics, 06(03), 377–392. https://doi.org/10.1142/s0219467806002379

[5] reCAPTCHA Usage Statistics. Trends.builtwith.com. Retrieved November 3, 2025, from https://trends.builtwith.com/widgets/reCAPTCHA

[6] Tsingenopoulos, I., Preuveneers, D., Desmet, L., & Joosen, W. (2022, June 1). Captcha me if you can: Imitation Games with Reinforcement Learning. IEEE Xplore. https://doi.org/10.1109/EuroSP53844.2022.00050

[7] Cuthbertson, A. (2025, April 15). Bots now make up more than half of global internet traffic. The Independent. https://www.the-independent.com/tech/bots-internet-traffic-ai-chatgpt-b2733450.html