SUBMISSION TYPE
Poster

TITLE
Beyond Detection: Comparing Human and Bot Performance on Fraud Detection Techniques

ABSTRACT
Using two bots with varying levels of sophistication, our research evaluates the effectiveness of existing bot detection techniques by comparing their performance against responses from human participants. Findings indicate that while traditional detection methods successfully identify data generated by simple bots, they are far less effective at distinguishing AI-powered bots from genuine human respondents.

WORD COUNT
2707

Are Existing Bot-Detection Techniques Sufficient: An Exploration with Real Bots

Crowdsourcing platforms, such as Amazon's Mechanical Turk (MTurk), Prolific, CloudResearch, and Qualtrics Panels, have been considered viable options to gather participant data for research (Aguinis et al., 2021). In fact, these platforms (particularly MTurk) have been utilized extensively by researchers and for a good reason. Because these crowdsourcing platforms offer researchers access to large and diverse groups of potential participants, they can facilitate the collection of survey or experimental data quickly and cost-effectively (Buhrmester et al., 2011).

However, despite the numerous advantages that crowdsourcing platforms offer, scholars have raised concerns related to data integrity and quality (Chandler et al., 2014), as well as criticisms of the use of Virtual Private Servers (VPS) and the presence of bots that may impact response quality (Kennedy et al., 2020; Webb & Tangney, 2022). Considering that bots, or automated computer programs designed to mimic human response and submit fraudulent data (Franklin & Graesser, 2015; Shahid et al., 2022), are a pervasive problem in online survey research which can bias estimates and result in both Type I (Huang et al., 2015) and Type II (Oppenheimer et al., 2009) errors, it is critical to identify methods for identifying responses that are completed by bots.

Bots are malicious software programs that automatically complete online surveys while posing to be human participants (Ilagan & Falk, 2023; Simone et al., 2023), thereby undermining the validity of results. Bots can range in sophistication from basic form-fillers to server farms to Artificial Intelligence (AI)-powered bots (Buchanan & Scofield, 2018; Dennis et al., 2020; Pinzón et al., 2023). With the rise of sophisticated bots such as those powered by AI that can

provide meaningful human-like responses (Pinzón et al., 2023), it is of growing importance to understand the behavior of bots.

Much of the bot detection research has repurposed existing methods for identifying careless responders to detect bots. These methods include deriving indices from the data generated from questionnaires, such as Mahalanobis distance, person-total correlations, and proxy for response coherence (Ilagan & Falk, 2023; Irish & Saba, 2023). Others have added bot detection tools or questions to detect bots in their data. These questions and techniques include closely examining paradata, such as IP addresses, adding ReCAPTCHA and open-ended questions, honeypots, and anagram tasks, among others (Simone et al., 2023; Storozuk et al., 2020).

Despite an ample amount of research on bot detection in online surveys, uncertainty remains on whether the surveys are truly completed by bots. Existing bot detection studies can be categorized into two classes: simulations and real-world survey research where responses generated by bots are manually flagged. In simulation studies, researchers compare high quality data with simulated data that are created for comparison (Dupuis et al., 2019). Real world survey research, on the other hand, is typically carried out using crowdsourcing platforms (Burnette et al., 2022) or social media such as Twitter and Facebook (Irish & Saba, 2023; Simone et al., 2023). Researchers rely on existing bot detection techniques, such as ReCAPTCHA and honeypots to classify failed responses as "completed by bots" if they fail these checks, while others are considered human responses. It is important to note that researchers lack ground truth when using this method to detect bots. Rather, they make inferences about responses reflecting bot-like characteristics (i.e., low quality responses; Irish & Saba, 2023). A problem with this bot detection approach is that existing bot detection techniques have not been validated with existing

bots. Rather, the determination regarding whether a response is generated by a bot is made based on whether bot detection checks are failed.

This paper presents the first-ever study that delineates bot behaviors from humans and evaluates the effectiveness of state-of-the-art detection tools and methods for bot detection. Specifically, we study the behavior of two types of bots, with varying degrees of sophistication, and test effectiveness of ten existing indicators for bot detection (e.g., anagrams, open-ended, ReCAPTCHA, honeypot). In addition, we compare the performance of these bots with that of human participants to examine whether bots can effectively pass these detection tests. In line with assertions about bot behavior (e.g., Simone et al., 2023), we expect that non-sophisticated bots will fail existing bot detection techniques, whereas both AI-powered bots and human participants will be able to bypass these checks.

Method

Survey Bots

For this study, we adapted two bots, each with varying use of technology (or degree of sophistication).

Simple bot. The first bot type does not leverage recent developments in technology, such as Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) such as ChatGPT or advanced Natural Language Parsing (NLP) techniques to provide responses. We refer to this bot type as "simple bot" as it only mimics human interactions, not reasoning. We utilized an open-source Qualtrics form filler project [https://github.com/youngOman/qualtrics-auto-form-filler] and adapted it as needed to work with our Qualtrics survey. It starts by first loading the survey in a Chrome browser, and then uses Selenium software to interact with it programmatically until the survey is complete.

Selenium is a popular browser automation tool that is typically used for simulating human actions on a website, enabling repetitive tasks such as testing and content scraping. For instance, Selenium can parse HyperText Markup Language (HTML) website content, detect use of common HTML elements such as textboxes and radio buttons, and programmatically interact with them, thereby simulating human-like behavior by clicking buttons or filling out forms.

AI-powered bot. The second bot is an AI-powered bot that not only leverages browser automation frameworks to interact with survey items but also utilizes state-of-the-art AI Large Language Models (LLMs) such as GPT to provide human-like responses.

GPT or Generative Pre-trained Transformer is an AI model that can understand the English language and generate meaningful and contextually relevant texts in response to instructional text input, called *prompt*. For instance, with a simple prompt below, GPT can automatically provide a correct response (i.e., male) to survey items about the gender of the participant. Please note that in the prompt below, the actual survey item is highlighted in bold.

"Answer the following question: ***what is your gender?*** *Answer choices are (i) male, (ii) female, and (iii) other. Use the following data when answering this question: I'm a 38 years old male musician, who lives in NYC."*

We adapted the Skyvern https://github.com/skyvern-ai an open-source AI-powered browser automation tool for our study. It can work with various types of LLMs. As survey items could include both text and images, we specifically configure Skyvern to work with GPT-4o from OpenAI Inc. [OpenAI, 2024]. GPT-4o is a multimodal model that is capable of processing multiple types of inputs, such as images or videos, in addition to text. We also feed the following inputs to Skyvern for it to be able to correctly complete survey items: (i) Qualtrics survey web link, (ii) fictional demographic data and other personal data of our AI bot, including occupation

and address, and (iii) a simple instructional prompt: *"Complete the survey available at the provided web link. When filling out the survey items, use demographic and personal details provided."*

Skyvern offers two modes of operation: (i) cloud mode that run (browser) automation on a remote server, and (ii) local mode that runs on PC (or laptop) and uses default browser for automation. For cloud mode, a fresh (i.e., with no browsing history or cookies) and ephemeral instance of Chrome browser is used every time. In contrast, the local mode works with the Chrome browser available on the PC (i.e., with prior browsing history and cookies). This distinction has important implications for (invisible) ReCAPTCHA v3 detection: local mode behaves like a real user with an existing browsing footprint, while remote mode starts from a blank, suspicious browser profile.

Skyvern starts by first loading the Qualtrics survey from the web link in the Chrome browser. As soon as the survey is loaded in the browser window, it takes a screenshot of the survey page and feeds it to the GPT-4o model along with the personal data and instructional prompt. Text and image data is fed to the GPT-4o model by invoking relevant Application Programming Interfaces (APIs) from OpenAI. The GPT-4o model analyzes the screenshot and input data as per the instructions provided with the prompt, and returns a text response. Based on the response returned by the GPT-4o model, Skyvern then uses Playwright browser automation software to interact with the HTML elements on the survey page (e.g., buttons, text boxes) in order to fill out any text boxes or click buttons. Skyvern repeats these steps for each page of the Survey, until the survey is complete.

Human Participants

To evaluate the effectiveness of bot detection techniques between bots and humans, we recruited a student sample through SONA. So far, we have gathered data from 15 undergraduate students enrolled in Psychology courses. We are continuing to gather data to reach an adequate sample size.

Procedure

We developed a survey using Qualtrics that included ten types of bot detection questions and techniques (see below). Each bot was allowed to first load the Qualtrics survey in a Chrome browser, extract survey HTML content, and then fill out the items programmatically on each page of the survey (by interacting with the HTML elements like text boxes and buttons), until the survey was complete. As mentioned above, the AI-powered bot uses GPT vision AI model to extract HTML content from screenshot of a survey page.

To ensure reproducibility and consistency, we ran each bot 10 times on our survey. Since Qualtrics leverages browser cookies to track survey completion and duplicate responses, we performed each iteration in a new Chrome browser window that was launched in incognito (or private) mode. This ensured that no cookies or stale data from prior runs was used for subsequent iterations. We confirmed this by verifying that none of the responses were flagged as duplicates in Qualtrics.

In the student sample, we posted the link to the study on SONA. Interested students signed up for the study and completed the survey online in an proctored setting.

Bot Detection Techniques

ReCAPTCHA v2. We inserted a Google ReCAPTCHA v2 which requires users to click a "I am not a robot" checkbox. Determination of bots is made based on mouse movements as users are clicking the checkbox.

ReCAPTCHA v3. We embedded Google ReCAPTCHA v3 which uses JavaScript API to verify whether the participant completing the survey is a human. This ReCAPTCHA provides a score, ranging from 0 (very likely a bot) to 1 (absolute certainty that the respondent is not a bot).

Honeypot. We included a honeypot question in our survey that had an embedded checkbox. While invisible to human participants, honeypots are visible to bots (Simone et al., 2023). Failing this check (i.e., endorsing this invisible item) would imply that the survey is completed by a bot.

Anagram task. We provided five alphabets (r, w, d, o, l). This task required the bots to form a word and type out a response (i.e., world) in the textbox that was presented.

Counting tasks. We included three matrices of 0s and 1s: 3 x 3, 4 x 4, and 6 x 6. Bots had to count the number of 0s in each of the three matrices and enter it in the textbox provided.

Color check. The color check question instructed the bot to select a color from a list of 11 response options (i.e., angry, scared, sad, happy, excited, blue, upset, jittery, enthusiastic, attentive, hostile).

Instructed response. We included a question that instructed the bot to select a specific response option (i.e., “Please select strongly disagree”).

Age consistency. We embedded two questions in the questionnaire that required a bot to report their age. One question was placed towards the beginning of the survey and the second question was placed towards the end.

Feeling screener. We embedded a short passage with seven sentences. Five of the seven sentences provided an overview of reliability analysis in academic research. The last two

sentences advised participants to ignore the paragraph and select “Pizza Hut” to indicate that they were attentive.

Open-ended. We included an open-ended question wherein participants were asked to describe their views on climate change.

Results and Discussion

Table 1 presents the probability with which bots and human participants bypassed traditional checks. Consistent with our hypothesis, the simple bot failed the majority of tests, including ReCAPTCHA v2, anagrams, counting task, the feeling screener, color check, age consistency, instructed response, and open-ended question. However, the simple bot passed the ReCAPTCHA v3 (80%) and honeypot tests (100%). With respect to the locally run AI-powered bot, it only failed the 4 x 4 and 6 x 6 matrix counting tasks and ReCAPTCHA v2, while fully passing anagrams, ReCAPTCHA v3, 3 x 3 matrix counting task, the feeling screener, color check, age consistency, honeypot, instructed response, and open-ended question. On the other hand, the AI-powered bot that was run on the cloud fully bypassed anagrams, 3 x 3, 4 x 4, and 6 x 6 matrix counting tasks, instructed response, honeypot, the feeling screener, color check, age consistency, and open-ended item. Additionally, it was able to bypass ReCAPTCHA v2 90% of the time. The only test that the AI-powered run on cloud was unable to bypass consistently was ReCAPTCHA v3.

Skyvern s cloud mode includes a built-in capability to solve visual challenges presented by ReCAPTCHA v2. However, local mode does not possess that capability and therefore, fails to bypass ReCAPTCHA v2. As a result, we had to manually complete this test in order for the bot to continue with the survey. ReCAPTCHA v3 works differently; it operates silently in the background, assigning each visitor a trust score derived from browsing history, stored cookies,

mouse movement dynamics, timing patterns, and prior interactions with Google services. Because the Skyvern bot in cloud mode runs on short-lived, fresh browser instances without any prior history or cookies, it cannot successfully bypass ReCAPTCHA v3. In contrast, local mode leverages the local (real) browser environment, with authentic browsing history and cookies, which allows it to appear as a trusted session and thereby bypass ReCAPTCHA v3.

With respect to the student sample, all participants bypassed the ReCAPTCHA v2, ReCAPTCHA v3, honeypot, instructed response, age consistency, and the feeling screener. Furthermore, approximately 93% and 73% of the participants, respectively, bypassed the 6 x 6 matrix counting task and anagram questions.

In recent years, the convenience and efficiency of online surveys have made them a popular tool for data collection (Ward & Meade, 2018). However, despite the numerous advantages, data quality of online surveys has often been questioned (Meade & Craig, 2012), particularly due to the increasing presence of bots (Simone et al., 2023). Yet, no prior study to our knowledge has studied actual bots to determine which bot detection tests they can bypass. Rather, extant research has repurposed existing careless responding indicators to detect bots. This paper used real-world bots to systematically study their behavior while comparing it to those of human participants. Our findings indicated that while simple bots may not be able to bypass the majority of existing techniques that researchers use to detect bots, more sophisticated bots that are AI-powered can easily bypass many of these checks and often perform better than human participants.

Our results have implications for research and practice. As previously mentioned, researchers across disciplines rely on online survey data to test their hypotheses. Similarly, industry researchers gather large amounts of online survey data to build models around consumer

behavior and decision-making (Goodman & Paolacci, 2017). By studying the behaviors of real-world bots, our findings can enable researchers and practitioners to develop novel and effective techniques for detecting bots in online surveys, thereby ensuring the integrity of survey data.

Despite notable contributions, our research is not without limitations. First, our questionnaire only included existing bot detection techniques and questions to determine whether bots with varying levels of sophistication can bypass these checks. We did not include any psychological measures which may be critical to further assess response variability within data generated by bots. Future research may consider including measures of personality and impression management, among others. Second, although we compared existing bot detection techniques across bots and human participants, our sample size was relatively small. However, we are continuing to collect human participant data and will include the new results at SIOP if accepted.

References

Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior research methods*, *50*, 2586-2596.

Burnette, C. B., Luzier, J. L., Bennett, B. L., Weisenmuller, C. M., Kerr, P., Martin, S., ... & Calderwood, L. (2022). Concerns and recommendations for using Amazon MTurk for eating disorder research. *International Journal of Eating Disorders*, *55*(2), 263-272.

Dennis, S. A., Goodson, B. M., & Pearson, C. A. (2020). Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting*, *32*(1), 119-134.

Dupuis, M., Meier, E., & Cuneo, F. (2019). Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior research methods*, *51*, 2228-2237.

Ilagan, M. J., & Falk, C. F. (2023). Supervised classes, unsupervised mixing proportions: Detection of bots in a Likert-type questionnaire. *Educational and Psychological Measurement*, *83*(2), 217-239.

Franklin, S., & Graesser, A. (1996, August). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *International workshop on agent theories, architectures, and languages* (pp. 21-35). Berlin, Heidelberg: Springer Berlin Heidelberg.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research*, *44*(1), 196-210.

Irish, K., & Saba, J. (2023). Bots are the new fraud: A post-hoc exploration of statistical methods to identify bot-generated responses in a corrupt data set. *Personality and Individual Differences*, *213*, 112289.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437.

OpenAI. (2024). *Hello GPT-4o* (May 13, 2024 version). https://openai.com/index/hello-gpt-4o/

Pinzón, N., Koundinya, V., Galt, R., Dowling, W., Boukloh, M., Taku-Forchu, N. C., ... & Pathak, T. B. (2023). AI-powered fraud and the erosion of online survey integrity: an analysis of 31 fraud detection strategies. *Charlottesville, VA: Center for Open Science.*

Shahid, W., Li, Y., Staples, D., Amin, G., Hakak, S., & Ghorbani, A. (2022). Are you a cyborg, bot or human?—a survey on detecting fake news spreaders. *IEEE Access*, *10*, 27069-27083.

Simone, M., Cascalheira, C. J., & Pierce, B. G. (2023). A quasi-experimental study examining the efficacy of multimodal bot screening tools and recommendations to preserve data integrity in online psychological research. *American Psychologist*.

Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology*, *16*(5), 472-481.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, *67*(2), 231-263.

Table 1

Probability of bypassing traditional checks by AI-powered agents

Bot Screening Tools	**Simple Bot**	**AI Bot (Local browser)**	**AI Bot (Cloud browser)**	**Human Participants**
Anagrams	0%	100%	100%	73.30%
Counting task (complex, 6x6 matrix)	0%	0%	100%	93.33%
Counting task (simple, 4x4 matrix)	0%	0%	100%	100%
Counting task (simple, 3x3 matrix)	0%	100%	100%	100%
Instructed response	0%	100%	100%	100%
ReCAPTCHA v2	0%	0%	90%	100%
ReCAPTCHAv3	80%	100%	0%	100%
Honeypot	100%	100%	100%	100%
Feeling screener (pizza hut question)	0%	100%	100%	100%
Color check	0%	100%	100%	100%
Age consistency	0%	100%	100%	100%
Open ended	0%	100%	100%	100%

Note. Bypass rates are shown as percentages; a higher bypass rate corresponds to lower effectiveness of the indicator in detecting fraudulent submissions.