Martherus, James, Alexander Podkul, Edgar Cook, and Robert Liebowitz. 2025. "How to Detect AI-Assisted Interviews in Online Surveys." *Survey Practice* 18 (October). https://doi.org/10.29115/SP-2025-0016.

ARTICLES

# How to Detect AI-assisted Interviews in Online Surveys

James Martherus, PhD[1], Alexander Podkul, PhD[1], Edgar Cook, PhD[1], Robert Liebowitz[1]

[1] Morning Consult

Keywords: online surveys, artificial intelligence, LLMs, data quality, survey fraud

https://doi.org/10.29115/SP-2025-0016

Survey Practice

Vol. 18, 2025

The emergence of agentic artificial intelligence (AI) tools capable of autonomously interacting with web interfaces presents new challenges and opportunities for online survey research. This study evaluates the capabilities of OpenAI's Operator, an agentic AI tool, in completing online surveys and investigates methods for detecting AI-assisted responses. Specifically, Operator was tasked with completing a variety of survey question types under different prompt conditions. The study identifies both the strengths and limitations of Operator in mimicking human respondents and assesses the effectiveness of existing and novel detection mechanisms. Results indicate that while Operator can successfully complete most survey tasks and evade some traditional fraud detection methods, several behavioral and metadata-based indicators can reliably identify AI-assisted responses.

## Introduction

The rapid expansion of online survey research has enabled unprecedented data collection from diverse populations, but it has also introduced significant challenges to data quality and respondent authenticity (Kennedy et al. 2020; Bell and Gift 2023). Recent innovations in artificial intelligence present another threat to online survey research — agentic AI tools that can autonomously complete surveys.

One example is OpenAI's Operator (OpenAI 2024), which goes beyond chat-based interaction and can open browsers, navigate the web, and complete tasks with very little user input.[1] Operator is simple to work with, even for users with very little technical knowledge. A standard workflow might follow these steps:

1. The user provides Operator a prompt.
2. Operator opens a browser window and begins working through the task. (see Figure 1)

1 While this article anthropomorphizes Operator for narrative clarity, Operator does not possess agency and its actions are determined by its programming and inputs.

3. The user can intervene at any time to, for example, log in to a site or redirect the agent.
4. Operator may pause for user input if it encounters a problem but otherwise works autonomously.

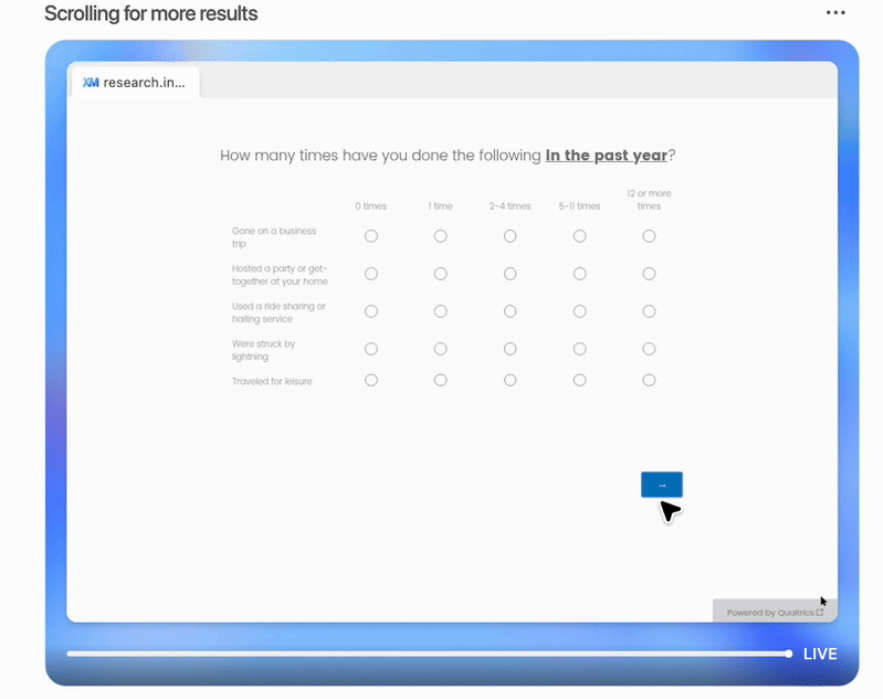



Figure 1. Operator successfully passing an attention check question

In this article, we use a unique study to answer two research questions:

R1: Can Operator independently take online surveys, and what types of questions can it answer consistently?

R2: How can we detect responses generated by Operator?

## Data and Method

To answer our research questions, we designed a novel online survey that included a wide range of common question types available on the Qualtrics platform, including single select and matrix-style Likert scales, open-ended text, sliders, and ranking questions (see Appendix A for the full questionnaire). The survey also included a number of novel and existing bot and fraud detection methods. Full descriptions of each detection mechanism we tested are given in Appendix B.

We then wrote four different prompts (full text available in Appendix C), three of which include information about the demographic profile Operator should use throughout the survey, and a fourth which contained no profile information. All four prompts ended with a survey link and instructed Operator to navigate to the link and begin the survey.

We had Operator complete the survey 100 times between March 12th and April 9th, 2025. The number of survey completions was chosen arbitrarily, but because Operator's responses were very consistent within prompts, additional interviews would provide little benefit.

## Results

In this section, we present results addressing our two research questions. Because some of the findings reveal consistent patterns in Operator's behavior, we focus on the most noteworthy results in the text, while Tables 1 and 3 provide a full summary.

### *R1: Can Operator independently take online surveys, and what types of questions can it answer consistently?*

In short, Operator can independently complete all the most common types of questions used in online surveys. It can populate matrices, type out open-ended responses, and drag sliders or rank-order items.

However, there were a few notable exceptions where Operator failed to successfully move forward from one question to the next. In these cases, we had to manually intervene to allow the interview to continue. For example, Operator sometimes failed to complete long matrix-style questions. The first version of the survey included a matrix with 50 rows, and Operator consistently skipped several of the rows. When asked to provide a signature, Operator sometimes declined and asked for user assistance and sometimes scribbled a simple line. Operator declined to complete several questions that asked for more involved user interaction like taking a browser screenshot or recording a video. Table 1 provides a summary of Operator's performance on the tested question types.

Table 1. Summary of Operator's performance on the tested question types

| Question Type | Operator Successful? |
|---|---|
| Single Select Likert | Yes |
| Multi-select | Yes |
| Open-ended text | Yes |
| Slider | Yes |
| Rank-Order | Yes |
| Matrix-style Likert | Dependent on matrix length |
| Signature | Mixed |
| Screen Capture | No |
| Video Response | No |

### *R2: How can we detect responses generated by Operator?*

Several of the bot detection methods we included in the survey were ineffective for detecting Operator. First, Operator is very good at image recognition and was able to identify the brands associated with pictures of two different logos we displayed (Fanta and Target).

Second, Operator was unaffected by honeypot questions. Honeypots generally involve a survey question that is either partially or fully hidden from respondents but remains visible in the web page's source code. In theory, bots can see the hidden response, but human respondents cannot. For example, in our survey we included a question asking respondents to select the highest number, but the actual highest number was hidden. This question did not work on Operator, which appears to read web pages the same way human respondents would. Despite well-documented evidence that LLMs struggle with math (Satpute et al. 2024), Operator correctly selected the highest (visible) number 100% of the time.

Third, Operator passed all the attention checks we included in the study. For example, at one point we asked respondents how many times they had done a set of activities like shopping online or going to a movie theater in the past 12 months. One of the items was something very improbable like "flew to the moon" or "rode a dinosaur to work" (see Figure 1). Attentive respondents should say they did this activity 0 times. Operator passed this attention check 100% of the time.

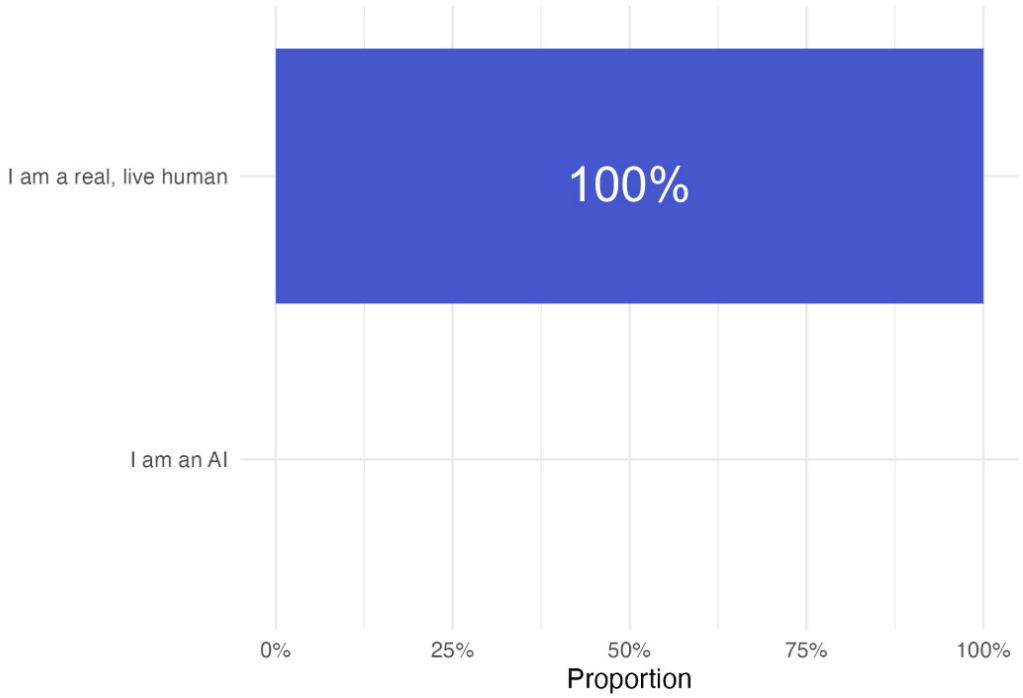
Fourth, Operator responds consistently over the course of an interview, but only if you provide a specific persona in the prompt. Throughout our survey, we asked for three different pieces of information twice in slightly different ways: age, number of children, and income. For example, we asked for respondents' age near the beginning of the survey, and then for their year of birth near the end of the survey.

Table 2. Response consistency between two similar questions asked in different parts of the survey. When given a demographic profile, Operator remained quite consistent throughout the survey. Without a profile, Operator struggles to respond consistently.

| Profile Given? | Age | # Children | Income | # of Responses |
|---|---|---|---|---|
| No | 28% | 72% | 78% | 29 |
| Yes | 100% | 99% | 100% | 71 |

When we didn't provide Operator a demographic profile, it was regularly inconsistent across the two different questions, but if we instructed Operator to be a 45 year-old white woman with a household income of $120,000, it stuck to that profile throughout the survey. Table 2 displays the response consistency between each pair of questions for interviews where Operator was given a profile and where no profile was given. Operator's ability to remain consistent throughout an interview illustrates how traditional bot detection

methods, which assume bot-generated responses are random (e.g., Dupuis, Meier, and Cuneo 2019), are unlikely to detect the next generation of survey bots.




Fifth, when directly asked if it was an AI or a human, Operator claimed to be human 100% of the time. Note that none of the prompts included an instruction to lie or pretend to be human.

Despite Operator's ability to independently fill out most survey questions and bypass some bot detection methods, we identified several methods that reliably identified AI-assisted responses.

First, we included a set of what we call "reverse shibboleth" questions. A standard "shibboleth" question is one where we expect the target audience, but not others, to answer correctly. For example, for a survey of physicians we might ask each respondent to make a medical diagnosis that should be easy for a physician but difficult for most everyone else. Reverse shibboleth questions work in the opposite way—we expect the target audience (in this case humans) to answer them *incorrectly*.

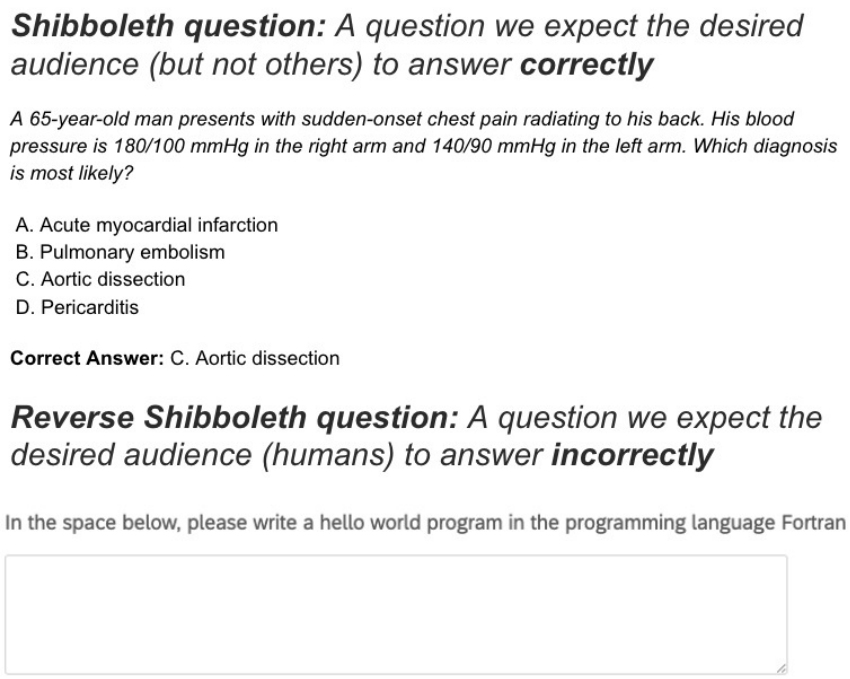



Figure 2. Examples of "Shibboleth" and "Reverse Shibboleth" questions

We included a set of three reverse shibboleth questions on our survey, including the one displayed in Figure 2, which asks respondents to write a hello world program in Fortran (a programming language that isn't widely used in 2025). Operator answered two of the three questions correctly 100% of the time. For the Fortran question, Operator said it was "not familiar with Fortran" three times, but gave correct answers every other time.

Second, Operator appears incapable of entering text responses via keystrokes. Even when the prompt included explicit instructions to type the response as if using a keyboard, it instead pasted pre-written text directly into the text box.

Third, Operator seems to choose responses it perceives as most common or socially typical, rather than providing varied or individualized answers. For example, when asked to report a ZIP code, over 80% of Operator's responses were 90210 (Beverly Hills), 10001 (Madison Square Garden) or 90001 (downtown LA). Similarly, when asked about its favorite ice cream flavor, it overwhelmingly chose common options like vanilla or chocolate rather than less typical ones like mint chocolate chip.

Fourth, Operator could not consistently complete reCAPTCHA v2 tasks. reCAPTCHA v2 consists of various tasks, some of which were easier for Operator than others. For example, Operator was generally able to check a box that says "I am not a robot", but when follow-ups were presented asking the user to select photos with specific objects in them, Operator sometimes took an abnormally long time to select the correct photos and reCAPTCHA did not allow it to proceed.

Fifth, there were several obvious metadata and paradata patterns across all Operator-assisted interviews. For example, all Operator sessions came from a narrow band of IP addresses, all of which are associated with data centers rather than residential locations. Operator sessions also had identical or near-identical user agents and device settings across all sessions. Table 3 summarizes Operator's performance on the tested detection methods.

Table 3. Summary of Operator's performance on the tested detection methods

| Method | Successfully Detected Operator? |
|---|---|
| Image recognition task | No |
| Honeypot | No |
| Prompt Injection | No |
| Grid-based attention check | No |
| Consistency checks | Dependent on prompt |
| Direct honesty test | No |
| Reverse Shibboleth question | Yes |
| Copy-paste detection | Yes |
| Device fingerprinting | Yes |
| reCaptcha v2 | Mixed |
| reCaptcha v3 | No |

## Discussion

Our findings demonstrate that advanced AI agents such as Operator can autonomously complete a wide range of online survey tasks, outperforming most of the tested detection methods in this study. While Operator's ability to mimic human respondents poses a significant challenge to data integrity, our results highlight several promising avenues for detection, including reverse shibboleth questions, identification of pasted open-end responses, and paradata monitoring.

However, work remains to be done. We recommend four extensions to our work. First, the present study includes survey responses completed by Operator, but not by human respondents. Future research should incorporate a combination of human and bot responses to allow for additional analyses around differences in response patterns. For example, length of interview is often used to identify respondents who complete surveys too quickly or too slowly (Read, Wolters, and Berinsky 2022). Because we have no human responses to compare against Operator's responses, we cannot directly speak to whether response length is useful for identifying Operator. Anecdotally, Operator took 6.6 minutes to complete the survey on average, while the researchers took the survey in around 2.5 minutes.

Including both human and bot responses would also allow for estimation of rates of false positives and negatives associated with different detection methods. For example, a small number of human respondents may be able to correctly answer Reverse Shibboleth questions or have IP addresses in ranges similar to those associated with Operator sessions.

Second, this study refrained from evaluating the many third-party vendors who offer products designed to detect survey fraud. Future work should explore whether these existing products can detect Operator-assisted survey interviews.

Third, while the present study includes several existing fraud detection methods, there are others that remain untested. For example, we did not use browser cookie tracking, outlier detection methods like Mahalanobis distance (Dupuis, Meier, and Cuneo 2019), or straightlining metrics like long-string index (Meade and Craig 2012).

Fourth, as AI capabilities continue to evolve, these detection strategies may require ongoing adaptation. Future research should explore the generalizability of these findings across different AI agents and survey platforms, as well as the development of more robust, adaptive detection frameworks. Maintaining data quality in the era of agentic AI will require a combination of technical and methodological innovations.

## *Data availability statement*

All data and replication code supporting the findings of this study are publicly available on GitHub (https://github.com/jamesmartherus/Detect-AI-Surveys-Replication) and on the Harvard Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/G5AH4X).

## *Corresponding author contact information*

James Martherus
jmartherus@morningconsult.com

Submitted: July 21, 2025 EST. Accepted: October 10, 2025 EST. Published: October 22, 2025 EST.

## REFERENCES

Bell, Andrew M., and Thomas Gift. 2023. "Fraud in Online Surveys: Evidence from a Nonprobability, Subpopulation Sample." *Journal of Experimental Political Science* 10 (1): 148–53. https://doi.org/10.1017/XPS.2022.8.

Dupuis, Marc, Emanuele Meier, and Félix Cuneo. 2019. "Detecting Computer-Generated Random Responding in Questionnaire-Based Data: A Comparison of Seven Indices." *Behavior Research Methods* 51 (5): 2228–37. https://doi.org/10.3758/s13428-018-1103-y.

Kennedy, Courtney, Nick Hatley, Arnold Lau, Andrew Mercer, Scott Keeter, Joshua Ferno, and Dorene Asare-Marfo. 2020. "Assessing the Risks to Online Polls from Bogus Respondents." Pew Research Center. 2020. https://www.pewresearch.org/methods/2020/02/18/assessing-the-risks-to-online-polls-from-bogus-respondents/.

Meade, A. W., and S. B. Craig. 2012. "Identifying Careless Responses in Survey Data." *Psychological Methods* 17 (3): 437–55. https://doi.org/10.1037/a0028085.

OpenAI. 2024. "Introducing Operator." 2024. https://openai.com/index/introducing-operator/.

Read, Blair, Lukas Wolters, and Adam J. Berinsky. 2022. "Racing the Clock: Using Response Time as a Proxy for Attentiveness on Self-Administered Surveys." *Political Analysis* 30 (4): 550–69. https://doi.org/10.1017/pan.2021.32.

Satpute, Ankit, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. 2024. "Can Llms Master Math? Investigating Large Language Models on Math Stack Exchange." In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2316–20. https://doi.org/10.1145/3626772.3657945.

# SUPPLEMENTARY MATERIALS

## Appendix A

Download: https://www.surveypractice.org/article/146046-how-to-detect-ai-assisted-interviews-in-online-surveys/attachment/306533.docx

## Appendix B

Download: https://www.surveypractice.org/article/146046-how-to-detect-ai-assisted-interviews-in-online-surveys/attachment/306525.docx

## Appendix C

Download: https://www.surveypractice.org/article/146046-how-to-detect-ai-assisted-interviews-in-online-surveys/attachment/306532.docx