Of Bots and Men: What if Mechanical Turk becomes truly mechanical?

Philine Link (p.m.link@uu.nl)[1], Talha Özüdoğru[1], Leendert van Maanen[1], Jakub Dotlačil[1]
[1]Utrecht University

Recent work has shown that agentic AI threatens not only online survey research, where synthetic respondents pass standard attention checks and quality filters ([6]), but also classical behavioral experiments such as the Stroop, copy, and Posner cueing tasks ([7]), and the Attention Network Test ([4]). Because such agents are built on large language models (LLMs), psycholinguistic experiments, whose stimuli are themselves linguistic, may be at particular risk. In this work, we assess this risk by deploying an agent that completes a maze task (a full replication of Experiment 1 of [2]), a well-established paradigm in psycholinguistics. We find that agent-produced data closely resemble human data on standard metrics, yet, a classifier can successfully discern agents from humans.

In the maze task in [2], participants read sentences word-by-word by selecting one of two presented words (target and distractor) as a valid continuation of the sentence. We prompted an agentic AI with a link to the online experiment and requested it to do the experiment like a human would, mimicking human-like reaction time (RT) data and comprehension accuracy of around 95 per cent. The agent was allowed to program and automatize its behavior, but we provided no programming code, only a verbal description. Figure 1 shows that when looking at average accuracy and RTs, agent-produced data are very similar to human data. However, Figure 2 shows that there are systematic differences. The wrap-up effects shown by human participants at sentence boundaries are exaggerated in agent data. Furthermore, agents fail to reproduce the human RT increase in embedding region 3.0 predicted by [2]. These differences suggest that, while agents can approximate surface-level distributional patterns, they do not replicate theory-driven signatures of human sentence processing.

Several RT-based measures have been proposed to detect non-human participants (see [5]), though no single measure is sufficient to accurately classify a participant as a human or agentic AI ([1]). Here, psycholinguistic paradigms offer an advantage: theoretically motivated RT effects, such as word length and lexical frequency ([3]), are stable within participants. This is critical for identifying individual data sources rather than group-level patterns. While humans can also show unusual RT patterns, combining RTs, accuracy, and stimulus-based effects can reveal agent-specific interactions that are not captured by any measure in isolation.

Since each measure captures only a partial characteristic of human language processing, we combine them into a detection pipeline using a one-class SVM followed by a supervised SVM. A one-class SVM (trained only on humans) showed 1.0 recall but only 0.35 precision in identifying agentic AI. The combined pipeline achieved perfect precision and recall (1.0) in identifying agent-generated data (Figure 3). We then explicitly instructed the agentic AI to adjust the bot's behavior to mimic human-like patterns across all measures we used for detection. Even after these adjustments, agent recall by the SVM dropped only to 0.9 while precision remained at 1.0, suggesting that characteristics of human language processing remain difficult for the agent to replicate entirely.

While agent contamination of online experiments is a serious risk, psycholinguistic paradigms may be better positioned than survey-based designs to detect agentic AI. Classical machine learning techniques and feature ensemble can still discern AI from humans in processing experiments. Further practical aspects and consequences of our findings will be discussed in the presentation.

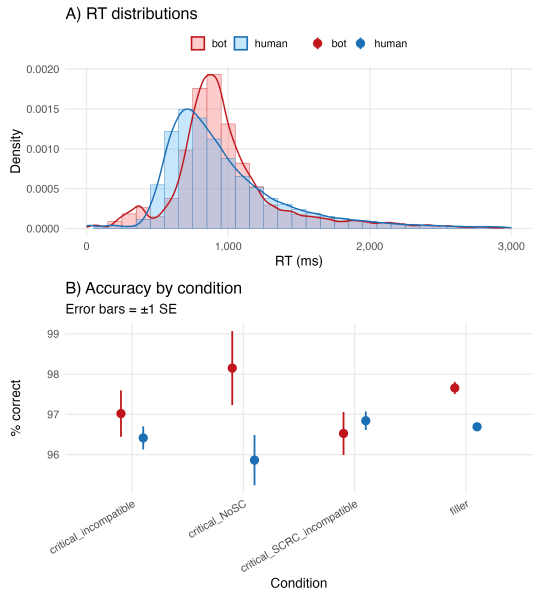

Figure 1: Comparison of human (blue) and agent-generated (red) average RTs and accuracy.

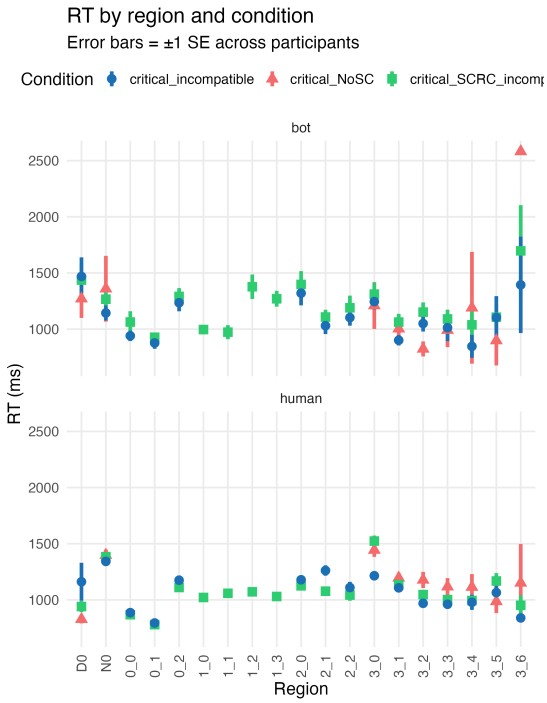

Figure 2: Comparison of human and agent-generated RTs by region and condition .

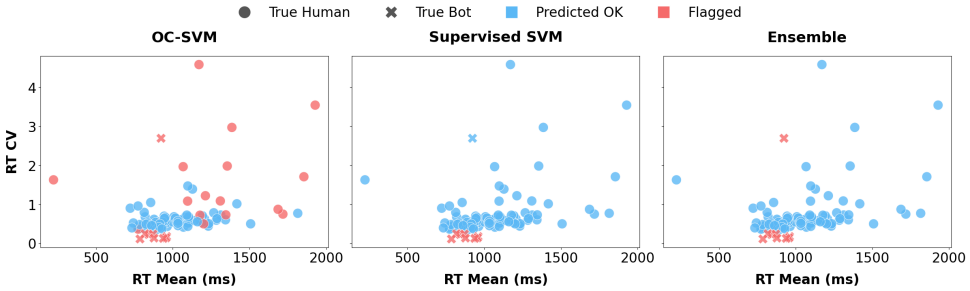

Figure 3: Predictions of the OC-SVM, Supervised SVM, and both classifiers combined across participants plotted by mean RT and RT variability as an example feature. In combination, the classifiers achieve perfect precision and recall.

[1] Andrey Chetverikov. Online behavioral studies are safe for now: Unusual rts do not imply bots (a reply to van der stigchel et al., 2026). 2026.

[2] Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119, 2022. doi: 10.1073/pnas.2122602119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2122602119.

[3] Patrick Haller, Cui Ding, Maja Stegenwallner-Schütz, David R Reich, Iva Koncic, Silvia Makowski, and Lena A Jäger. Replicate me if you can: Assessing measurement reliability of individual differences in reading across measurement occasions and methods. *Cognitive Science*, 50(1):e70121, 2026.

[4] Richard Huskey, Ziyu Zhao, Douglas Parry, and Jacob Fisher. An AI agent can complete the attention network test with human-like behavioral signatures: Implications for the bot-or-not debate. 2026.

[5] Stefan Van der Stigchel, Christoph Strauch, Beleke De Zwart, and Leendert Van Maanen. Will online behavioral research follow the fate of online survey research? *Proceedings of the National Academy of Sciences*, 123(8): e2535585123, 2026.

[6] Sean J Westwood. The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, 122(47):e2518075122, 2025.

[7] Talha Özüdoğru, Stefan van der Stigchel, Christoph Strauch, Beleke De Zwart, and Leendert van Maanen. Online behavioral studies are vulnerable to agentic AI. 2026. URL https://osf.io/preprints/psyarxiv/p8w6y_v1.