ARTIFICIAL INTELLIGENCE IN INVESTIGATIVE INTERVIEWING

Brandon May, Rebecca Milne, Alyssa Fox and Amber King

Center for AI in Policing, Florida Institute of Technology, United States; University of Portsmouth, United Kingdom

POLICY BRIEFING NOVEMBER 2026

This policy briefing examines what the evidence on artificial intelligence supports for the interviewing of victims, witnesses and suspects. This report is based on the Center for AI in Policing review of the available scientific evidence. To date, it has found opportunities to improve practice using AI, alongside risks that may undermine justice. Under the Méndez Principles, adoption should depend on demonstrated interviewing benefit, protection of rights and accountable institutional oversight (Principles on Effective Interviewing 2021).

KEY FINDINGS

1. **Operational evidence remains limited.** Most research concerns simulations, archival material or relate to technical benchmarks (e.g., does AI outperform human comparisons). To date, only eight scientific papers have examined AI interacting with a person in real time, all of which raise cautions when using AI in interviewing contexts. There are no demonstratable benefits currently that support the effectiveness of AI in actual investigations or interviews (May et al. 2026).

2. **Training and specialist support show some promise.** Practicing interviewing behaviours with AI avatars and receiving feedback can, temporarily, help investigators ask better questions (Røed et al. 2023). AI models adapted for specific interviewing tasks can also perform better than general-purpose AI (May et al. 2026). However, research here is still limited on whether trainees retain the skills learned and apply them effectively in real investigations (Haginoya et al. 2023; May et al. 2026; Røed et al. 2023).

3. **Question quality and evidence quality can diverge from recommended practices.** Some AI systems can produce appropriate individual questions but perform less well when conducting a full interview (Järvilehto et al. 2026). Investigators should remember that effective interviewing involves more than how questions are worded, as the functionality of a question depends on contextual factors (e.g., voice tone, non-verbal behaviour of the interviewer, etc). It also requires rapport and appropriate responses to the interviewee's answers and needs (*Principles on Effective Interviewing* 2021). Human oversight is often touted as a mitigation for this (e.g., human in the loop); however, this alone does not ensure that investigators produce better questioning. Further, obtaining more information per question does not consistently lead to more correct information overall (May et al. 2026; Sun et al. 2025).

4. **Generative AI does contaminate and create false memories.** AI can create false memories that persist over time. In one experiment, misleading chatbot questions produced over three times as many false memories as the control group. Further, confidence in those memories remained higher (Chan et al. 2024). AI can also create confirmation biases that anchor an investigator to a specific set of questions. This in turn can lead an interviewer to make biased decisions, and potentially result in miscarriages of justice (May 2025).

5. **Automation creates different risks across tasks.** There are numerous vendors providing interviewing solutions, that claim AI can be used in transcription, report generation, detection deception, to name a few. There is no scientific evidence that would support the use of any of these tools. AI use in transcription, question classification, detection deception, and credibility assessments remains unsupported by the research and are currently unsuitable within investigative interviews.

POLICY IMPLICATIONS

IMPORTANCE OF FINDINGS

An investigative interview – which can broadly be defined as any investigative process that seeks to elicit information from witnesses, victims or suspects - produces information that shapes an investigation, enquiries and judicial decisions (*Principles on Effective Interviewing* 2021). This includes information obtained in an emergency call, during a first account at the scene of an event, or during a substantive (e.g., custodial) interview. In all of these examples, AI is being used to support these processes. This includes but is not limited to questioning, transcription, decision support, summarisation, or interpretation of an account. This can – and has - changed the evidence for what someone reports. As we have seen in recent months, AI has been touted as a suitable tool for information elicitation (e.g., report writing in first accounts) and is now widely used to produce a fluent summary of what a person has said. But, it is important to remember that to date – as of November 2026 - there is **no** empirical evidence that would support this approach. Consequently, it is highly probable that we have seen miscarriages of justice as a result of AI use in interviewing procedures.

Table 1.
An overview of the use of AI and the Mendez Principles.

AI AND THE MÉNDEZ PRINCIPLES	
Principle	**Evidence and policy implications**
1 Foundations	Benefits in actual investigations remain unestablished.
2 Practice	Better questions do not ensure accuracy. AI can contaminate memory.
3 Vulnerability	Suitability across age, language and communication needs remains unestablished.
4 Training	Avatar practice is promising; retention and transfer remain uncertain.
5 Accountability	Verify generated accounts. Retain recordings, AI outputs and human edits.
6 Implementation	Require independent evaluation, limits, oversight and rules to suspend use.

KEY RECOMMENDATIONS

Police leaders, justice ministries and oversight bodies should make the following conditions part of AI procurement and operational policy

1. We should restrict autonomous questioning. Do not approve routine AI-led interviews in live investigations until independent evaluation demonstrates benefit and safeguards for the intended system, population and setting.
2. We should exclude unvalidated AI credibility tools, AI deception detection, and emotion or credibility tools that seek to determine truthfulness or guilt. AI is truth indifferent (May et al., 2026)
3. Interviewees remain vulnerable to AI use. Testing is needed across age, language and communication needs, with specially trained interviewers and applicable legal safeguards before an AI is considered (*Principles on Effective Interviewing* 2021).
4. We must prioritise supervised training, which include investment in validated simulations and AI-facilitated training. This still requires evidence of retention and transfer before claiming operational benefits (Haginoya et al. 2023; May et al. 2026; Røed et al. 2023).
5. We must retain and document recordings (e.g., Principle 5 - on *accountability*), prompts, model versions, outputs and human edits. This also requires source verification, secure access, and arrangements for review and disclosure (Harrington 2023; May et al., 2025; *Principles on Effective Interviewing* 2021; Roxhage and Ahlström 2026).
6. We need institutional accountability. Agencies should name a responsible decision maker and establish independent oversight. We should also set implementation criteria, reassess model changes and require supplier cooperation.

RESEARCH OVERVIEW

BACKGROUND

AI is being developed to simulate interviewees, generate questions and hypotheses, transcribe speech, draft statements and evaluate interview material. Our recent review, pending publication, has systematically mapped this literature across applications, populations, comparators and outcomes. Searches ran from June to July 2026, with unpublished literature included in our analyses. Thirty-nine papers were identified as being suitable for analysis and were subsequently synthesized to identify the following policy-relevant categories (May et al. 2026).

TRAINING AND PROFESSIONAL COMPETENCE

Use of AI in simulations provided some evidence that it might be suitable for low-stakes arenas. This includes AI use as a tool in training environments that does not expose a real interviewee to a trainee's mistakes. Røed and colleagues found that avatar practice improved open-ended questioning and that automated feedback provided an additional benefit as interviewers increased the number of open questions compared to their baseline (Røed et al. 2023). These findings support further investment in supervised practice, consistent with Principle 4 on training (*Principles on Effective Interviewing* 2021).

However, the quality of feedback is yet to be fully explored and requires further research. In one avatar system, automatic classification agreed with human operators on only 52% of 11 question types (Haginoya et al. 2023). A system that misclassifies a leading question can reinforce poor practice. Evidence on sustained retention and transfer to actual investigations also remains limited. Training tools therefore need expert checking and evaluation of what interviewers subsequently do (May et al. 2026).

QUESTIONING AND HUMAN OVERSIGHT

A specialised AI model tested with 108 children in a mock event study increased recommended question types from 39.3% in the human-only condition to 73.1% with AI support and 88.1% when humans read AI-generated questions verbatim. In this study, participants recall improved, with no statistically significant difference in reported errors. This is promising evidence for constrained support but it is worth noting that humans remained in control of the interview (i.e., AI was never used to question a witness). In the same study interviewers reworded 201 AI-generated questions as none of the revised questions met the study's criteria for recommended questioning.

Elsewhere, models that performed well when generating a single question deteriorated during extended interaction (Järvilehto et al. 2026). Principles 1 and 2 require evaluation of the complete interviewing process, and yet, there is no empirical study that would support the use of AI in real interviews. Describing a system as subject to human oversight is also insufficient (i.e., human-in-the-loop is not an appropriate safeguard).

MEMORY AND INFORMATION RELIABILITY

Chan and colleagues tested whether misleading chatbot questions could distort memory in mock witness interviews. After a generative chatbot interview, participants falsely recalled 36.4% of the details tested, compared with 10.8% in the control group – i.e., more than three times the rate. Further, the false-memory rate did not decrease when participants were tested again one week later (Chan et al. 2024).

Other studies have found that AI sometimes helped people recall information but did not consistently improve the accuracy of their accounts, at times creating false memories (Sim et al., 2026; Sun et al. 2025). Evaluations must therefore check whether AI helps interviewees provide accurate information and whether it introduces false details. Asking more open questions, obtaining more information per question or receiving positive user feedback is not enough to demonstrate better interviewing. As such, before introducing AI into any interviewing context testing should continue to establish whether AI can contaminate, distort or create false memories, and whether they are comparatively better than human interviewers. As Dr Julia Shaw has recently stated "*AI is the ultimate false memory making machine*"

RESEARCH OVERVIEW

TRANSCRIPTION AND STATEMENT PRODUCTION

Every AI-generated summary, or transcription, require verification as part of the achieving best evidence framework. However, even with verification, there remains concerns that AI-generated content may contain undetectable errors (May et al., 2026). One transcription study found errors varied with accent and noise, with content-word substitutions two to five times more frequent for a regional English accent than for standard British English in two AI systems (Harrington 2023). Other research has also found that AI shifted effort towards checking generated text rather than removing the need for work (Roxhage and Ahlström 2026). As such, less writing does not necessarily mean less work or time saved overall.

Under Principle 5 on accountability, agencies should retain the original account and make every consequential AI alteration reviewable. In other words, generated summaries must remain distinguishable from the interviewee's own words. Further, time saved in drafting should be assessed alongside the time taken to verify output, omissions and potential changes in meaning. Suppliers should also specify how sensitive interview data are stored, accessed and reused, and enable agencies to meet applicable confidentiality and disclosure duties (*Principles on Effective Interviewing* 2021).

CREDIBILITY AND DECEPTION ASSESSMENT

Good performance on one task does not mean AI can detect deception, nor does it support the use of AI in any deception detection practices. Despite many vendors claiming AI is able to detect deception, research has consistently shown limitations in AI use and deception (Wu et al. 2023) and credibility assessment (Karataş 2026). Further, AI tools are much less consistent than experts when judging how believable statements are. In one example, Wu and colleagues reported accuracy of up to 95.11% when testing a lie-detection system on existing archival video datasets. However, the study did not establish whether the system could reliably detect deception during live investigative interviews (Wu et al. 2023).

In summary, police should not use AI in any measure of credibility assessment. This includes, but is not limited to assessments of facial expressions – commonly associated to body language or micro-expressions – emotions, or credibility to decide whether someone is telling the truth. At present there is no independent testing that shows that a specific tool is reliable for any of the aforementioned purposes and settings (Karataş 2026; Wu et al. 2023) albeit, there may be some future benefits (Li et al., 2026). Further, despite vendor claims that AI could detect deception during investigative interviews, the available scientific evidence does not establish that it can do so reliably or accurately.

IMPLEMENTING THE MÉNDEZ PRINCIPLES

Principle 6 requires implementation through institutional and national measures - none of which exist for AI. As such, agencies should seek to evaluate pilots independent of vendors before implementation and consider policy needs related to the use of AI in interviewing. ImpleMéndez can support this shared evaluation across jurisdictions, however, we must recognize that independent evaluation will take some time. These evaluations should include interviewees and independent legal and interviewing specialists, to include behavioural scientists, alongside those responsible for purchasing and operating the technology.

Just because there is an absence of a detected harm is not evidence that harm cannot, or has not, occurred. Wider adoption should follow demonstrable interviewing benefit and effective safeguards, with arrangements to suspend use when either fails.

FURTHER INFORMATION

Research basis: May, B., Milne, R., Fox, A., King, A., and Jones, M. (in preparation). Artificial Intelligence in Investigative Interviewing: A Systematic Evidence Map of Applications, Evidence, and Research Gaps. Unpublished manuscript

Note, all communications should be directed to Brandon May, Ph.D, bmay@fit.edu

SELECTED REFERENCES

Chan, Samantha, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F. Loftus. 2024. "Conversational AI Powered by Large Language Models Amplifies False Memories in Witness Interviews." arXiv preprint. https://doi.org/10.48550/arXiv.2408.04681.

Haginoya, Shumpei, Tatsuro Ibe, Shota Yamamoto, Naruyo Yoshimoto, Hazuki Mizushi, and Pekka Santtila. 2023. "AI Avatar Tells You What Happened: The First Test of Using AI-Operated Children in Simulated Interviews to Train Investigative Interviewers." *Frontiers in Psychology* 14: 1133621. https://doi.org/10.3389/fpsyg.2023.1133621.

Harrington, Lauren. 2023. "Incorporating Automatic Speech Recognition Methods into the Transcription of Police-Suspect Interviews: Factors Affecting Automatic Performance." *Frontiers in Communication* 8: 1165233. https://doi.org/10.3389/fcomm.2023.1165233.

Järvilehto, Liisa, Yongjie Sun, Nami Aiba, et al. 2026. "Large Language Model (LLM) and Human Performance in Child Investigative Interviewing Question Formulation Tasks." *Behavioral Sciences & the Law* 44 (1): 142–63. https://doi.org/10.1002/bsl.70029.

Karataş, Zeki. 2026. "Can a Large Language Model Judge a Child's Statement?: A Comparative Analysis of ChatGPT and Human Experts in Credibility Assessment." *Journal of Evidence-Based Social Work* 23 (1): 78–93. https://doi.org/10.1080/26408066.2025.2547211.

Li, Siyu, Pär-Anders Granhag, Yunhan Shi, Yongjie Sun, Thomas J. Nyman, Shumpei Haginoya, and Pekka Santtila. "Using large language model–based artificial intelligence (AI) suspects to train strategic use of evidence: Preliminary evidence of transfer to mock suspect interviews." *Law and Human Behavior* 50, no. 3 (2026): 240.

May, Brandon. 2025. "Are We DrAIfting? The Role of Large Language Models in Police Investigative Interviewing." *Investigative Interviewing: Research and Practice* 15 (1): 1–22. https://iiirg.org/wp-content/uploads/2025/09/IIRP_20250506_May_Final.pdf.

May, Brandon, Jordan Nunan, Marshall Jones and Marek Palace. 2025. "Taxonomy of Bullshit: AI Truth-Indifference in Policing Contexts." *Policing: A Journal of Policy and Practice 20.* https://doi.org/10.1093/police/paag052

May, Brandon, Rebecca Milne, Alyssa Fox, Amber King, and Marshall Jones. (in preparation). "Artificial Intelligence in Investigative Interviewing: A Systematic Evidence Map of Applications, Evidence, and Research Gaps." Unpublished manuscript.

Principles on Effective Interviewing for Investigations and Information Gathering. 2021. The Méndez Principles. https://www.apt.ch/sites/default/files/publications/apt_PoEI_EN_11.pdf.

Røed, Ragnhild Klingenberg, Gunn Astrid Baugerud, Syed Zohaib Hassan, et al. 2023. "Enhancing Questioning Skills through Child Avatar Chatbot Training with Feedback." *Frontiers in Psychology* 14: 1198235. https://doi.org/10.3389/fpsyg.2023.1198235.

Roxhage, André, and Jonathan Ahlström. 2026. "Designing an AI-Enhanced Interface for Cognitive Support in High-Stakes Interrogations." Master's thesis, Lund University. https://lup.lub.lu.se/student-papers/record/9230187.

Sim, Mattea, Yael Eiger, and Tadayoshi Kohno. "AI-Enabled Human Memory Manipulation: Misleading AI-Generated Summaries Distort Human Memory." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 9, no. 4, pp. 2327-2339. 2026.

Sun, Yongjie, Haohai Pang, Liisa Järvilehto, et al. 2025. "Comparing the Performance of a Large Language Model and Naive Human Interviewers in Interviewing Children about a Witnessed Mock-Event." *PLOS ONE* 20 (2): e0316317. https://doi.org/10.1371/journal.pone.0316317.

Szojka, Zsofia A., Suvimal Yashraj, and Thomas D. Lyon. 2025. "Automated Question-Type Coding of Forensic Interviews and Trial Testimony in Child Sexual Abuse Cases." *Law and Human Behavior* 49 (2): 163–72. https://doi.org/10.1037/lhb0000590.

Wu, Yi-Chang, Yao-Cheng Liu, and Ru-Yi Huang. 2023. "The Use of Artificial Intelligence in Interrogation: Lies and Truth." *IAES International Journal of Robotics and Automation* 12 (4): 332–40. https://doi.org/10.11591/ijra.v12i4.pp332-340.

This article is based upon work from COST Action CA22128 (ImpleMéndez), supported by European Cooperation in Science and Technology (COST).