INVESTIGATION · METHOD AND THE FUTURE

Artificial Intelligence in Investigations: Capabilities and Risks

Artificial intelligence can search, transcribe, translate, match and summarise evidence at a scale no investigator can match, but it also makes errors that look convincing. This article explains what AI can do in investigations, where it fails, and the safeguards that keep it useful.

ModQu research article · 971 words · 11 references (APA 7th edition)

Artificial intelligence (AI) is now built into many of the tools investigators use. It ranks documents for review, transcribes and translates interviews, recognises faces and objects in images, links records that refer to the same person, flags unusual transactions and, with large language models, drafts summaries and answers questions about case material. These capabilities can save large amounts of time. They also introduce new kinds of error, some of which are hard to see. This article describes the main capabilities, the main risks, and how investigators can use AI without letting it decide what the evidence shows.

What AI can do

Most AI used in investigation is machine learning: software that learns patterns from examples rather than following rules written by a programmer. Its main uses fall into five groups.

- **Search and review.** In technology-assisted review, a model learns from reviewers' decisions and ranks the remaining documents by likely relevance. Grossman and Cormack (2011) found that this approach could find relevant documents at least as effectively as exhaustive manual review while requiring far less human reading.
- **Speech and language.** Transcription, translation and summarisation make large volumes of audio and foreign-language material accessible.
- **Images and video.** Models detect faces, vehicles, weapons and text in images, and can compare a face against a database.
- **Pattern detection.** Models flag anomalies in financial, network or communications data for human review.
- **Generative assistance.** Large language models draft documents, extract names and dates, and answer questions about material they are given.

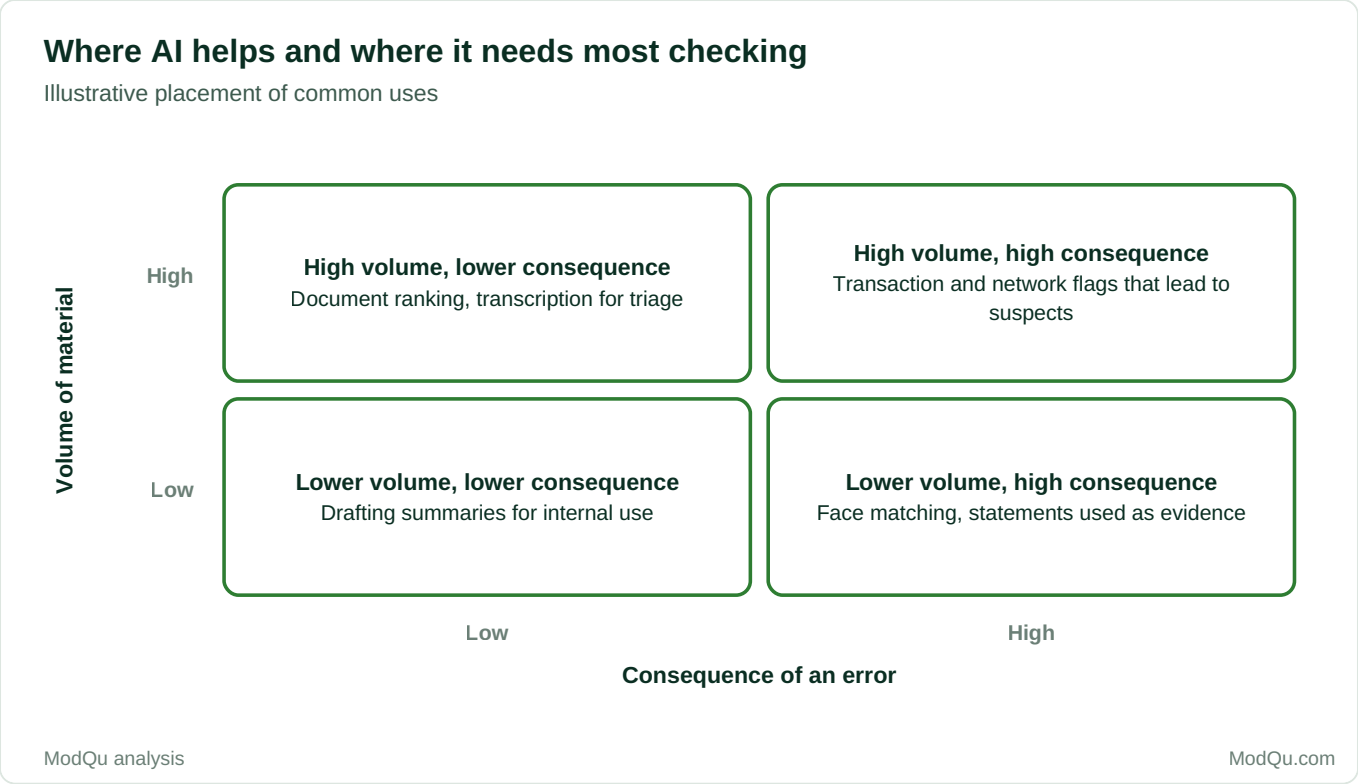


Figure 1. AI adds most value where volumes are large, but every use where an error could point at the wrong person needs independent human checking. Diagram by ModQu.

Where AI fails

Bias and uneven accuracy. A model is only as good as the data it learned from. The United States National Institute of Standards and Technology tested face recognition algorithms and found that many had higher false match rates for some demographic groups than others, with the size of the difference varying widely between algorithms (Grother et al., 2019). Buolamwini and Gebru (2018) found that commercial gender classification systems were much less accurate for darker-skinned women than for lighter-skinned men. Lum and Isaac (2016) showed that a predictive policing model trained on past drug arrests would send police back to the same neighbourhoods, because the arrest data reflected where police had looked, not where drug use occurred.

Errors that look convincing. Large language models can produce fluent text that is false, a problem known as hallucination (Ji et al., 2023). In *Mata v. Avianca, Inc.* (2023), a United States court sanctioned lawyers who had filed a brief citing court decisions that a chatbot had invented. In an investigation, an invented date, quotation or source in an AI summary can travel into a report unnoticed.

Opacity. Many models cannot explain why they produced a result. Rudin (2019) argues that for high-stakes decisions, people should prefer models that can be interpreted, rather than relying on after-the-fact explanations of black-box systems.

Automation bias. People tend to accept automated outputs and to stop looking for contradicting information (Parasuraman & Manzey, 2010). A face match or risk score can therefore narrow an investigation too early. Investigators may stop interviewing other witnesses or checking other suspects once a system has named someone, even though the match was only ever a probability. Comparisons have also shown that complex tools are not always more accurate than simple methods: Dressel and Farid (2018) found that a widely used recidivism risk tool predicted reoffending no more accurately than people with no criminal justice expertise.

Safeguards

Guidance such as the National Institute of Standards and Technology's AI Risk Management Framework emphasises mapping how a system will be used, measuring its performance and managing its risks over time (National Institute of Standards and Technology [NIST], 2023). For investigators, this means:

- **Treat outputs as leads.** A face match, flag or summary is a reason to look, not a finding.
- **Check against the source.** Every fact taken from an AI summary should be confirmed in the original document or recording.
- **Know the error rates.** Use tools that have been tested independently, on data similar to the case, including across demographic groups.
- **Record use.** Note which tool, version and settings were used and what was checked, so that the work can be reviewed and disclosed.
- **Protect data.** Do not paste confidential case material into public AI services that may store or reuse it.
- **Follow the law.** Some uses are regulated. The European Union's AI Act classes many law enforcement uses, including some biometric identification, as high-risk and subjects them to requirements for data quality, documentation and human oversight (Regulation (EU) 2024/1689, 2024).

A short example

The following hypothetical case illustrates these safeguards. An investigator reviewing 400 hours of recorded calls uses AI transcription and a language model to list every mention of a shipment. The model's list includes a call in which the suspect appears to name a date. Before relying on it, the investigator plays the recording and finds that the transcription misheard a word and the model then built a confident but false statement on it. The investigator corrects the record, notes the error in the case file, and adds a rule that every AI-extracted fact must be checked against the audio before it enters a report. The tool still saved weeks of listening, but only the checked facts were used.

Conclusion

AI lets investigators handle volumes of material that would otherwise be impossible, and in tasks such as document review it can match or exceed human performance. But it can be biased, opaque and confidently wrong, and people tend to trust it too much. Investigators who treat AI outputs as leads, verify them against the original evidence, know their tools' error rates and record how they used them can gain its benefits without letting it decide what happened.

References

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), Article eaao5580. https://doi.org/10.1126/sciadv.aao5580

Grossman, M. R., & Cormack, G. V. (2011). Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, *17*(3), Article 11.

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test (FRVT) part 3: Demographic effects* (NISTIR 8280). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8280

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article 248. https://doi.org/10.1145/3571730

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023).

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union, L*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x