INTELLIGENCE · ARTIFICIAL AND COLLECTIVE INTELLIGENCE

"Risks of Automated Intelligence: Hallucination, Bias and Overconfidence"

Automated tools can produce intelligence quickly, but they can also invent facts, reproduce unfair patterns and sound more certain than they should. This article explains these three risks, the human tendency to over-rely on automation, and the practical controls that keep them in check.

ModQu research article · 1,889 words · 21 references (APA 7th edition)

Automated systems now carry out a growing share of intelligence work. They classify documents, summarise reports, score risks, flag anomalies and draft text. Their speed and scale are real advantages. But automated intelligence has characteristic ways of going wrong, and they differ from the ways human analysts go wrong. Three stand out: **hallucination**, the production of false information that looks genuine; **bias**, systematic error that falls unevenly on different groups; and **overconfidence**, certainty that is not matched by accuracy. A fourth risk sits with the people who use the systems: the tendency to rely on them more than they deserve.

This article explains each risk, gives documented examples, and sets out controls that any organisation using automated intelligence can apply.

Hallucination

Hallucination is most associated with language models, which generate text by predicting plausible continuations rather than by retrieving verified facts. Ji et al. (2023), in a survey of the research, define hallucination as generated content that is nonsensical or unfaithful to the source content. They distinguish content that contradicts the source from content that cannot be verified from it at all. Both are common. Maynez et al. (2020) examined machine-written summaries of news articles and found that a large proportion contained information not present in the original, much of it incorrect.

Hallucinations are dangerous because they are fluent. A fabricated statistic is presented in the same polished style as a true one. Invented references are a familiar case: asked for sources, a model may produce citations with real-sounding authors, titles and journals that do not exist. In a widely reported American case, lawyers filed a brief citing several judicial decisions that a chat assistant had invented, and the court imposed sanctions (Mata v. Avianca, Inc., 2023). The lawyers had asked the tool whether the cases were real, and it assured them they were.

In intelligence work, a hallucinated detail can be harder to catch than in a court filing, because readers seldom check each source. A false figure in a market summary, a misattributed quotation in a briefing, or an invented event in a chronology can travel through several products before anyone notices.

Bias

Bias in automated systems usually starts with data. A model learns from examples, and if the examples reflect past discrimination, unequal coverage or poorly chosen labels, the model learns those patterns too. Barocas and Selbst (2016) explain how this can happen even when no one intends it and when sensitive attributes such as race or sex are excluded, because other variables act as proxies. Mehrabi et al. (2022) catalogue many sources of bias, from how data are collected to how outcomes are measured and how models are evaluated.

Several studies have documented the consequences. Buolamwini and Gebru (2018) tested commercial facial analysis systems and found error rates for darker-skinned women far higher than for lighter-skinned men. Obermeyer et al. (2019) examined an algorithm widely used in American hospitals to identify patients for extra care. Because it used past health-care spending as a measure of need, and less had historically been spent on Black patients with the same conditions, it systematically underestimated their needs.

Risk scoring in criminal justice has been especially contested. Journalists at ProPublica reported that a widely used recidivism tool wrongly labelled Black defendants as high risk at nearly twice the rate of white defendants (Angwin et al., 2016). The tool's developer responded that its scores were equally accurate for both groups in another sense. Both claims could be true. Chouldechova (2017) and Kleinberg, Mullainathan and Raghavan (2017) showed mathematically that, when groups differ in their underlying rates, no score can satisfy all common definitions of fairness at once. Choosing which kind of fairness matters is a value judgement, not a technical one. Dressel and Farid (2018) added that the tool was no more accurate than predictions made by untrained people recruited online.

For business intelligence, the same mechanisms apply wherever models score customers, applicants, suppliers or regions. A model trained on past sales may underrate markets the company has never served well. A model trained on past hiring may learn the preferences of past managers. These are not only ethical problems but analytical ones: the intelligence is wrong for part of the population.

Overconfidence

The third risk is that automated systems express more certainty than their accuracy justifies. Guo et al. (2017) showed that modern neural networks, although more accurate than older models, are often worse **calibrated**: when they report 90% confidence, they are right much less than 90% of the time. Figure 1 illustrates the pattern.

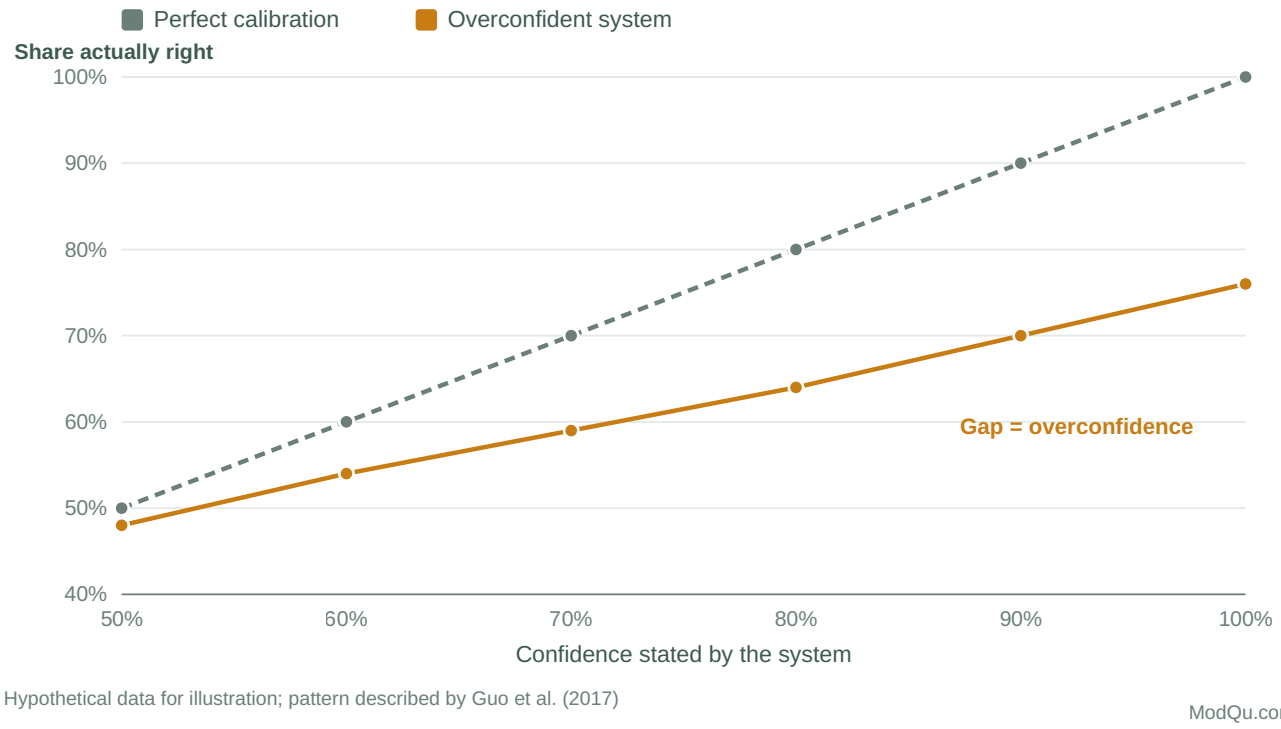


Figure 1. What overconfidence looks like. A well-calibrated system is right as often as it claims; an overconfident one claims more than it delivers. Hypothetical data for illustration; pattern described by Guo et al. (2017).

Overconfidence is aggravated by a related weakness. Models can learn shortcuts: features that happen to predict the right answer in training data but have nothing to do with the real task. Geirhos et al. (2020) describe how image classifiers, for example, may rely on background or texture rather than the object itself. Such models perform well in testing and then fail, often confidently, when conditions change. Amodei et al. (2016) list robustness to such changes in conditions among the central practical problems of AI safety.

In intelligence terms, a model that is confident and wrong is worse than one that is uncertain and wrong, because confidence discourages checking. Language models compound the problem by expressing every answer in the same assured tone.

Over-reliance by people

Automated systems are used by people, and people's behaviour creates the fourth risk. Research on automation in aviation, medicine and process control shows a consistent pattern. Skitka, Mosier and Burdick (1999) identified **automation bias**: people follow automated advice even when other information shows it is wrong, and fail to notice problems the system does not flag. Parasuraman and Manzey (2010) reviewed the evidence on complacency and found it affects experts as well as novices, and is not easily removed by training alone.

Bainbridge (1983) described the "ironies of automation". The more reliable an automated system, the less practice people get in doing the task themselves, and the less able they are to step in when the system fails. Lee and See (2004) argue that the goal is not maximum trust but **appropriate reliance**: trust that matches what the system can actually do, in the conditions where it is being used.

Controls

The four risks call for different controls, summarised in Figure 2.

Three risks of automated intelligence

What causes each, how to spot it, and how to control it

Risk	Main cause	Warning signs	Controls
Hallucination	Model generates plausible text, not checked facts	Claims with no traceable source; invented citations	Ground in sources; verify every claim
Bias	Training data or labels reflect past unfairness	Error rates differ between groups	Test by group; document data; review outcomes
Over-confidence	Stated confidence not matched to accuracy	Same assured tone for right and wrong answers	Measure calibration; show uncertainty
Human over-reliance	People trust a tool that is usually right	Fewer checks; errors pass unnoticed	Train, rotate, audit samples by hand

Compiled from Ji et al. (2023), Mehrabi et al. (2022) and Guo et al. (2017); table by ModQu

ModQu.com

Figure 2. Risks of automated intelligence and how to control them. Compiled from Ji et al. (2023), Mehrabi et al. (2022) and Guo et al. (2017); table by ModQu.

Against hallucination, the main control is grounding and verification. Give the system the source documents it should use, require every claim to be linked to a passage, and check those links. Treat any statement without a traceable source as unverified. Never use model-generated citations without confirming that each source exists and says what is claimed.

Against bias, test performance separately for the groups the system affects, not only on average. Document where training data came from and what they leave out. Gebru et al. (2021) propose **datasheets** for datasets, recording how data were collected and their known limitations; Mitchell et al. (2019) propose **model cards** that report a model's intended use and its performance across groups. Review real outcomes after deployment, since bias often appears only in use.

Against overconfidence, measure calibration: compare stated confidence with actual accuracy on held-out data, and adjust or flag scores accordingly. Present outputs with honest uncertainty, and say explicitly where the model has not been tested.

Against over-reliance, design the work so that people stay engaged. Audit a sample of automated outputs by hand on a regular schedule. Rotate analysts so that they keep the skills to do the task without the tool. Make it easy to record disagreement with the system, and treat such disagreements as useful data.

These controls fit within broader frameworks. The AI risk management framework of the National Institute of Standards and Technology (2023) organises the work into four functions: govern, map, measure and manage. It asks organisations to identify where systems are used, measure their trustworthiness, and assign clear responsibility for managing risks.

Questions to ask before relying on an automated product

Before an automated output informs a decision, an analyst or manager can ask six short questions:

1. **Where did this come from?** Can every factual claim be traced to a named source that someone has checked?
2. **What was the system trained or tested on?** Does that resemble the situation it is being used for now, or has the market, population or period changed?
3. **How accurate is it, and for whom?** Is there evidence of accuracy for the specific groups, regions or products involved, not only on average?
4. **How confident should we be?** Has the system's stated confidence been compared with its actual record, and does the product show uncertainty honestly?
5. **What would we expect to see if it were wrong?** Are there independent sources, such as survey data, field reports or simple counts, that could confirm or contradict it?
6. **Who has checked it?** Is there a named person who reviewed the output and is prepared to stand behind it?

If the answers are unclear, the output should be treated as a lead to investigate rather than as a finding to act on.

A business illustration

A recruitment consultancy introduces an automated tool to rank candidates for client roles. In the first month, consultants find it saves hours. An internal audit then compares rankings with outcomes and finds that candidates who took career breaks are ranked consistently lower, regardless of experience. Investigation shows that the training data came from past placements, when clients had favoured continuous employment histories. The tool had learned a historical preference, not a measure of ability. The consultancy retrains the model without the employment-gap feature, sets up quarterly audits by group, and requires consultants to review every shortlist rather than accept the top five automatically.

At the same time, the consultancy uses a language model to draft candidate summaries. A spot check finds that some summaries attribute qualifications the candidates do not hold. The team changes the process so that summaries are generated only from the candidate's own documents, with each statement linked to the passage it came from.

Common mistakes

- **Testing only on average.** A model that is 95% accurate overall may be much worse for a minority group or an unusual case.
- **Assuming that removing sensitive variables removes bias.** Other variables can act as proxies.
- **Trusting the confidence score.** Unless calibration has been measured, a stated probability is not a reliable probability.
- **Checking only at launch.** Data, conditions and users change; risks must be monitored continuously.
- **Leaving no one accountable.** Every automated product needs a named person responsible for its quality.

Conclusion

Automated intelligence brings speed and scale, but it also brings characteristic errors: invented facts, unfair patterns and unwarranted certainty, made worse by people's readiness to trust a system that is usually right. None of these risks is a reason to avoid automation. Each is a reason to use it with discipline: grounding outputs in sources, testing performance across groups, measuring calibration, keeping people engaged, and holding a named person accountable for every product that reaches a decision-maker.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety* (arXiv:1606.06565). arXiv.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias*. ProPublica.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

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

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330).

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

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (Article 43). https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

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

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). https://doi.org/10.18653/v1/2020.acl-main.173

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), Article 115. https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM. https://doi.org/10.1145/3287560.3287596

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

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

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

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252