ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

AI is Statistics

Why statistical thinking is vital for the effective, ethical and safe use of AI

2 March 2026

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

1. Introduction: AI is statistics

Alan Turing tried to answer the question of whether machines could think by considering what he called the imitation game. If there was a machine that could successfully converse with a human interrogator and convince it that the machine was human, then there is a meaningful sense in which we could say that machines think.

Looking at this question at the start of 2026 the answer looks obvious – the large language model (LLM) AIs that proliferate today all seem to quite convincingly imitate conversing with a human and it can certainly be hard to distinguish AI responses from human responses. But there is something different about how AIs converse – and that difference is statistics.[1]

LLMs are complex statistical models. They are trained on vast quantities of text, identify patterns and use these patterns to repeatedly predict the next word in a sequence. And what they are trying to predict is the next word in the sequence that a specific user would want to see – so, when they receive a new prompt from a user, the model takes in new data and makes new predictions until its response is accepted.[2]

This is important to emphasise: especially as the models get increasingly sophisticated and feel increasingly human-like. At a fundamental level what AIs are trying to do is recognise patterns and use them to predict the optimal response to an input. This is, in part, why an AI can sometimes invent a source to cite in favour of a certain position. It is fundamentally different to how humans think.

Why is the Royal Statistical Society making this (seemingly philosophical) point? Because as AIs get used in an increasingly wide range of situations – recruitment, health care, credit lending, insurance pricing, the justice system, education, decisions about welfare entitlement, etc – it is going to require everyone to think more like statisticians to use this technology effectively and ethically. Statistical thinking can help people question the quality of the data that an AI is using, what assumptions are driving pattern-recognition, what the impact of different prompts might be.

This paper makes a number of practical recommendations drawing on this insight into the statistical nature of AI.

To improve public understanding of AI:

Recommendation 1: *The Government should embed training on statistical thinking in its general AI skills training module, with an emphasis on how to evaluate, question and understand AI tools in order to use them effectively.*

Recommendation 2: *Statistical thinking and statistical evaluation – in the context of AI – should be a core part of digital literacy in schools.*

---

[1] For a summary of recent research here see Scientists Go Serious About Large Language Models Mirroring Human Thinking from *Towards Data Science*.

[2] It is, of course, not just statistics that AI depends on. It also draws heavily on linear algebra, calculus, numerical optimisation, and computer science. The Academy of Mathematical Sciences is planning a report setting out how the mathematical sciences as a whole underpin AI.

To improve evaluation of AI:

Recommendation 3: The government should invest in collaborations with industry and academia to develop new statistical approaches to AI evaluation.

Recommendation 4: Any high-stakes deployment of AI should include explicit evaluation of whether the system is operating on data that is similar to its training data, and should provide clear evidence of performance under distribution shift.

To improve the use of AI in policy-making:

Recommendation 5: The Government's AI Playbook should be updated to reflect the statistical nature of LLMs.

Recommendation 6: Regulators should adopt a common statistical framework for the evaluation of AI systems, including requirements for uncertainty quantification, subgroup analysis and robustness testing.

Recommendation 7: Any public-sector deployment of AI should include transparent, statistically grounded evidence of its performance, including distributional assumptions and limitations.

To improve how practitioners engage with AI:

Recommendation 8: Government and industry should invest in developing statistical methods for evaluating and interpreting LLM-derived features, ensuring that their use in modelling is robust, transparent and evidence-based.

Recommendation 9: Funding bodies should support research at the interface of statistics, mathematics and AI interpretability, with a focus on improving model monitoring, governance and safety.

2. Public understanding of AI

For most people, everyday engagement with AI now happens through systems powered by large language models: chatbots, search assistants and the many applications built on top of them. Recent applications include AI-assisted triage in healthcare, automated grading and feedback in education, and AI-driven support for public services such as benefits administration or tax guidance. These high-impact examples illustrate both the potential and the statistical limitations of LLMs in real-world contexts.

These tools feel conversational, responsive and often surprisingly insightful – their power is indicated by the increasing number of people who use them for companionship or therapy. The corporate and media reporting on AI, its benefits and risks often encourages the public to think of these new tools as intelligent beings.

Yet beneath the surface they remain statistical engines, and consequently, statistical thinking and statistical evaluation shed light on several aspects of AI. For example, statistical tools can tell us about the properties of data, including its quality, provenance and the biases it contains. It helps explain how weaknesses in the data used to train or prompt a model will translate into biases or inaccuracies in the model's behaviour. And it is relevant to thinking about why the data is chosen, how it is sampled, cleaned and represented – and how these inputs shape what an AI system is capable of saying or doing.

On the other hand, statistical evaluation can be very helpful to assess whether the output of AI models is SAFE: Secure, Accurate, Fair and Explainable; and, in general, whether it is reliable and trustworthy.

People do not need to be experts in statistics to effectively use AI – but a level of statistical thinking and statistical evaluation does help with understanding what these systems are, what they are not, how to interpret their limitations, and how to assess them. Concepts such as sampling error, uncertainty, model fit, bias, robustness and explainability are no longer academic abstractions but practical tools for making sense of AI outputs. Equally important is the statistical mindset that provides a habit of and structure for enquiry when engaging with AI: what data is a result is based on? How reliable it is? How are results in line with scientific knowledge about a problem? How might it vary over different iterations or with slight changes in inputs? What assumptions lie beneath it? As AI becomes woven into everyday life, these skills become an even more essential part of digital literacy.

It is important therefore that government invests in strengthening people's statistical literacy in the context of AI. The government recently released a suite of AI training modules. We were disappointed to see that these were entirely focused on using particular companies' AIs – and there was no general training from a neutral source, and no guidance on how to contextualise or interrogate the outputs of these tools. These statistical skills also need to be embedded in education around AI from when it is first encountered in schools. The recent curriculum review emphasises the importance of digital literacy and this is welcome – but it makes no mention of statistical literacy in this context and we believe that this needs to be a core part of digital literacy.

Recommendation 1: The Government should embed training on statistical thinking in its general AI skills training module, with an emphasis on how to evaluate, question and understand AI tools in order to use them effectively.

Recommendation 2: Statistical thinking and statistical evaluation– in the context of AI – should be a core part of digital literacy in schools.

3. Evaluation

Modern AI chatbots can produce answers that are fluent, confident and persuasive – even when they are completely wrong. This is not a sign of malfunction but a reflection of the underlying statistical machinery. Most tools based on large models currently have no natural way to express uncertainty or to signal that they are operating outside their zone of competence. Statistical methods provide precisely the tools needed to understand and communicate these limits.

This is especially important in high-stakes environments: by those we mean where decisions impact on people's health, liberty, reputation and finances.

Statistical approaches such as Bayesian estimation help quantify how confident a model truly is in its outputs. Techniques like Monte Carlo dropout, concordance measures and ensemble modelling allow us to see how much a model's answer changes when the model is perturbed, mimicking adversarial attacks, and evaluating robustness. Explainable tools such as Shapley values, attention scores and gradient based importances can help to assess which variable features "drive" the outputs. Prediction intervals can communicate a range of plausible answers rather than a single definitive result. These are well-established statistical principles; understanding how to apply them to models of the size and complexity of modern AI is essential if we want to move from polished-sounding answers to trustworthy ones.

A further challenge is that many AI models work well only when their inputs resemble the data they were trained on. When the data shifts – because the population changes, or the circumstances differ, or the inputs are simply unusual – performance can degrade quickly and unpredictably. Statistics again provides the language to diagnose this. Sensitivity analysis helps us see how performance changes when inputs vary. Influence functions show which training examples most heavily affect a model's decisions. Out-of-distribution detection can highlight when a model is being asked to operate in unfamiliar territory. Bootstrapped stability testing can reveal whether reported performance metrics are robust or fragile. Again, in the context of modern models, which are trained on massive datasets for which there is generally no available public description, there is a need for work to extend these statistical tools to provide the monitoring we need to ensure the tools perform as required.

Beyond all this lies the fundamental distinction between correlation and causation. Many AI systems learn correlations extremely well but lack the ability to reason about cause and effect, which is essential for decision-making (especially where individuals are impacted such as in healthcare, education, welfare or justice). Statistical causal inference tools, such as counterfactual analysis and causal discovery methods, provide ways of assessing whether a model has learned meaningful causal relationships or merely surface patterns. Approaches rooted in causal rules and in mechanistic interpretability such as the do-calculus, have the potential to test if the conclusions drawn by an AI system are sound and reliable.

Without statistical evaluation, we cannot reliably determine how capable AI systems are in a given context, a risk that is amplified by their surface fluency, which suggests a competence they do not always possess. With statistical evaluation, we gain a clearer understanding of when a model can be trusted, when it should be used cautiously, and when it should not be used at all.

Without statistical evaluation, the assessment of AI quality and reliability inevitably becomes subjective, and audit based, rather than objective and risk based. Statistical evaluation allows different AI requirements, such as Security, Accuracy, Fairness and Explainability, to be measured with consistent metrics, that can also be integrated in a single assessment score, and for which statistical tests can provide objective thresholds.

There are wider methodological challenges that AI raises. At present, evaluation of large language models is fragmented, model dependent and inadequate. Benchmarks are often narrow and not reflective of the tasks for which AI systems are actually being used. The datasets used for evaluation are unrepresentative, and evaluation results are difficult to compare, and are not model agnostic. Yet as more organisations seek to use AI systems – and as use cases range from advising to creating, recommending or deciding – a proper approach to evaluation is vital. Priority access for the Foundation Model Taskforce is welcome, but it does not match the scale of need. The UK has faced a similar problem before: in the early 1990s, rapid growth in medical research outpaced the methods for assessing cost and effectiveness, prompting the creation of the Health Technology Assessment Programme. That investment underpinned the UK's global leadership in healthcare evaluation, and there is now a comparable opportunity to build world-leading capability in AI evaluation.

The need for such capability is evident from real examples. In healthcare imaging, evaluation must check accuracy, generalisability, and the balance between sensitivity and specificity; in education, generative models must be assessed for provenance, factual accuracy and responsible use of sources. More broadly, evaluations must reflect the overall system in which AI is deployed and must adapt as models evolve; something current one-off, pre-deployment checks cannot provide. Bias assessment remains difficult, and communicating evaluation results clearly to different audiences is still underdeveloped. Statistics has a central role to play in addressing these gaps, providing rigorous methods for uncertainty estimation, bias quantification and effective communication. This work should be driven by collaborations

between industry, government and academia – to define evaluation goals and set and solve research challenges.

Recommendation 3: The government should invest in collaborations with industry and academia to develop new statistical approaches to AI evaluation.

Recommendation 4: Any high-stakes deployment of AI should include explicit evaluation of whether the system is operating on data that is similar to its training data, and should provide clear evidence of performance under distribution shift.

4. Implications for policy

It is important for policy-makers to understand the statistical nature of AI: both when using AI in the policy development process and in other aspects of their jobs, and when making decisions about the regulation, funding and use of AI.

There are a few ways that statistical thinking can improve the use of AI in policy development processes. First, it is important to consider data quality: AI outputs are only as reliable as the data feeding the model, and policymakers must be equipped to assess how representative, timely, and unbiased the underlying data are. Statistical thinking is vital to help policy professionals assess the quality of data inputs to make a judgement on the reliability of outputs. Second, understanding that AI excels at spotting correlations and patterns is important in thinking about appropriate uses of AI. For instance, there is interest in using AI to simulate focus groups: AI is likely to be very good at identifying patterns, but what it cannot do is reflect that humans tend to have unlikely and unpredictable sets of views, which a pattern analysis might struggle to capture. Furthermore, simulations based on biased data will never generate the views of populations which are not well represented in it, leading to the amplification of existing inequalities.

The government has developed guidance in the form of an AI Playbook to set out how government should use AI. This could be strengthened in a number of ways by foregrounding the statistical nature of AI. First, the Playbook's section on the limitations of AI could go further by requiring all departments to quantify and communicate statistical uncertainty in model outputs. This is an essential step given that AI systems produce probabilistic, not deterministic, results. Second, the new material on AI quality assurance and managing risk would be more effective if it introduced structured, evidence-based evaluation standards (such as pre-deployment validation, calibration checks and ongoing drift monitoring) to ensure models are assessed with the same rigour applied to other forms of statistical evidence. Third, procurement guidance could be strengthened by requiring suppliers to provide statistical detail about model training data (its representativeness, bias risks, class balance, and known limitations) so that departments can assess the evidence base behind vendor claims. Finally, governance structures such as AI boards and system inventories should explicitly embed statistical oversight, ensuring that decision-makers have access to the expertise needed to interrogate model performance, monitor drift and reliability over time, and ensure models remain fit for purpose in real-world conditions.

There are also implications for regulation. The UK's current regulatory approach relies heavily on existing regulators interpreting and applying current legislation to new AI technologies. This provides flexibility, but it also places significant demands on regulators' ability to interpret complex, statistical systems. Because AI systems are fundamentally probabilistic, their oversight requires an understanding not only of what they do, but of how reliably they do it, where they fail, and how those failures vary.

In regulated sectors, organisations are expected to attest to the safety and effectiveness of the systems they use. Yet evaluation metrics are often inconsistent, opaque, or presented without information about sampling, uncertainty or variation across subgroups. A single accuracy figure, absent measures of uncertainty or robustness, can give a misleading picture of reliability. Regulators therefore need statistical capacity if they are to judge whether the evidence presented is meaningful and comparable.

In public policy itself, AI use cases are too often proposed or deployed without transparent, statistically grounded evaluation. Live facial recognition is a clear example: performance claims are sometimes presented without information on sampling, confidence intervals, error distributions, or performance across demographic groups. Without a statistical foundation, public claims about accuracy risk being incomplete or misleading.

A strong statistical foundation is therefore essential if policy-makers are to assess AI responsibly and if regulatory oversight is to be credible.

Recommendation 5: The Government's AI Playbook should be updated to reflect the statistical nature of LLMs.

Recommendation 6: Regulators should adopt a common statistical framework for the evaluation of AI systems, including requirements for uncertainty quantification, subgroup analysis and robustness testing.

Recommendation 7: Any public-sector deployment of AI should include transparent, statistically grounded evidence of its performance, including distributional assumptions and limitations.

5. AI practitioners

For practitioners working directly with data, the emergence of large language models opens up new possibilities, but also new responsibilities. LLMs offer powerful ways to extract semantic information from unstructured text, generate synthetic data for augmentation, assist with data cleaning, or support data linkage across systems. But integrating these capabilities into statistical workflows requires careful thought. Practitioners must assess how LLM-derived features compare with established methods, how stable and unbiased they are, and how their use affects downstream models. Just as statisticians have developed principled approaches to dealing with complex data types such as time-series, images, shapes and genomic sequences, the same rigour is now required for features derived from LLMs.

There is also a growing need to understand what models are doing internally. Modern AI systems use high-dimensional representations – embeddings, attention patterns, latent spaces – which are themselves statistical objects. Recent research by OpenAI[3] has demonstrated that forcing models to explain their thinking in English in fact leads to deception, with models learning to create explanations to satisfy humans rather than to accurately record their behaviour. To really understand how these models work, where they might fail and how their behaviour may change over time, we need to learn to interpret the symbolic language in which their processing is recorded. Statisticians and mathematicians have an

[3] GPT-6: The AI That Thinks in Its Own Secret Language by DevBoost Lab (Medium)

essential role in analysing these structures, developing interpretability tools, and building monitoring systems that can detect drift, bias or unexpected behaviour.

The relationship between statistics and AI is becoming deeper and more central to practice. For AI to be deployed responsibly, practitioners need access to statistical tools, statistical insight and statistical scrutiny.

Recommendation 8**:** Government and industry should invest in developing statistical methods for evaluating and interpreting LLM-derived features, ensuring that their use in modelling is robust, transparent and evidence-based.

Recommendation 9**:** Funding bodies should support research at the interface of statistics, mathematics and AI interpretability, with a focus on improving model monitoring, governance and safety.

From past to present...

The image of the wheatsheaf first appeared in our original seal. Being the end product of the harvesting and bundling of wheat, it was a pictorial way of expressing the gathering and analysis of data: the foundations of statistical work.

It also implied that statistical practice comprises more than the collection of data: it consists of active interpretation and application as well (threshed for others, if the rural analogy is sustained). Rigorous data gathering is still at the heart of modern statistics, but as statisticians we also interpret, explain and present the data we collect.