ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

RSS EVIDENCE ON DATA REGULATION IN THE AGE OF AI AND OTHER DATA-INTENSIVE TECHNOLOGIES

07 September 2026

This response has been submitted via an online form. Please find a copy of the evidence we submitted below. Full details about the call for evidence can be found here.

1 Accessing and using data

The Royal Statistical Society (RSS) is a membership organisation for statisticians and data scientists, and we advocate for the importance of statistics and data. Representing over 10,000 members, we champion the role of statistics and data in society, working to ensure that policy formulation and decision-making are informed by evidence for the public good.

The value of data access and reuse depends on confidence in data quality, provenance and governance. AI systems are highly sensitive to the characteristics of the data used to train and operate them, including its representativeness, timeliness, completeness and potential biases. We believe it is therefore important that regulators strengthen expectations around documenting data provenance and known limitations.

We have longstanding concerns about barriers to data sharing and linkage across the public sector, which limit the effective use of administrative data for analysis and official statistics. Done well, data linkage could improve public services and people's interaction with those services, reducing duplication and ensuring that information is updated consistently. Improvements in linkage must, however, be accompanied by strong safeguards. Key ethical considerations include proportionality, function creep (where systems initially justified for one purpose start being used for other purposes), and ensuring that linkage does not extend beyond well-defined and justified purposes. Technical considerations are equally important; these include accuracy of matching, management of uncertainty, prevention of duplication, and interoperability between systems. Well-designed identifiers can reduce data-entry errors and improve coverage, but public trust depends on robust governance arrangements, auditability and transparency.

Data access requirements on public datasets should promote competition, innovation and the widest possible public benefit. Policies that increase barriers to access risk favouring well-resourced organisations over smaller businesses, researchers, charities and civil society organisations, potentially reducing innovation and limiting the broader social and economic value generated through data reuse. Open and proportionate access conditions help create a level playing field, enabling a diverse range of users to develop products, services, research and insights from public data. The value of public data often arises from secondary and unforeseen uses, making it important that access arrangements support experimentation, research and innovation rather than narrowly defined commercial applications.

2 Data quality, accuracy and downstream impacts

Data quality is a foundational requirement for the deployment of AI and other data-intensive technologies to be useful. Poor quality, incomplete, unrepresentative or outdated data can lead directly to unreliable outputs and harmful outcomes. However, high-quality input data does not automatically produce fair or accurate outputs. Assessments of system performance should therefore include uncertainty quantification, confidence intervals, robustness testing, subgroup performance analysis, sensitivity to changing inputs, and evidence of performance under distribution shift.

Existing data protection concepts of fairness and accuracy remain important but may require additional interpretation in the context of AI. In particular, regulatory approaches should recognise that many AI outputs are probabilistic rather than deterministic. Evaluation of AI should therefore consider not only whether outputs are accurate on average, but also how uncertainty is communicated and how performance varies across different groups and contexts. Risks to fairness may only emerge after deployment; we therefore advocate for continuous monitoring and evaluation rather than relying on one-off pre-deployment assessments.

3 Governing data use across organisations

Determining accountability and responsibility

Questions of responsibility and accountability are becoming increasingly complex, as AI systems are often developed by one organisation, deployed by another and then relied on by individuals who have limited or no sight of how its outputs are produced. Responsibility for data collection, processing, model development, deployment, and fine tuning may therefore be distributed across organisations. This fragmentation of responsibility creates regulatory ambiguity, leaving potential gaps in oversight and enforcement.

It is therefore essential that data governance and AI governance function in alignment. To give an example, for a government department commissioning a research agency to process social research data, the existing controller/processor framework remains broadly fit for purpose. Where the research agency processes data on behalf of the government department, the department remains the data controller and the agency the processor, regardless of whether AI tools are used in the research process. However, the use of AI introduces additional governance considerations – through the government's AI and data ethics framework – which extend beyond traditional data protection roles. While data governance concerns how data is collected, managed and protected, the AI ethics framework introduces consideration of the broader ethical and societal implications of using the data and the technologies built upon it. Data and AI governance should therefore be closely aligned, and perhaps even integrated, to reduce the bureaucratic burden. There is a strong case for recognising the responsibilities of AI developers and providers, particularly where AI solutions are outsourced, though this does not necessarily require the creation of a new formal role alongside data controllers and processors.

AI and the challenge of anonymisation

AI presents a serious challenge to ensuring that personal data remains genuinely anonymised. In the context of health data, for example, maintaining public trust in our ability to do this is essential if we are to maintain the consensus in favour of the use of anonymised data for health research. Traditional anonymisation approaches are being challenged by AI systems which can identify individuals even where direct identifiers have been removed, from subtle patterns, correlations across variables or inferred characteristics. Re-identification risk is increasingly dependent on the interaction between data, the models used to analyse it, and the wider data environment. This means re-identification risk should be treated as a dynamic, context-dependent and fundamentally statistical question, requiring ongoing evaluation as technologies and data ecosystems evolve.

Rethinking automated decision-making

Regulation around automated decision-making will need to adjust to remain fit for purpose in the age of AI. A useful example is the government's proposed rollout of AI tutors to support disadvantaged pupils.[1] These systems infer a pupil's knowledge, abilities and learning needs, and uses those inferences to personalise learning. The AI model infers a student's knowledge from limited information and uses that to guide subsequent teaching. Errors in this process, whether through incorrect explanations, misjudgement of ability, predictive analytics identifying 'at risk' pupils or inappropriate sequencing of material, risk compounding over time as the system continues to adapt.

Over an extended period, an AI tutor may develop a detailed profile of a pupil that might be used to make decisions about sets, intervention programs, exam entries or progression pathways. While teachers may formally be the decision-makers, meaning that decisions would not be considered automated, it is reasonable to expect that AI-generated profiles and recommendations would carry significant weight in practice. This raises important questions about accountability and transparency. The AI system is effectively making a series of consequential judgements, through its profiling, predictions and recommendations, yet it may provide little information about the uncertainty, limitations or reliability of those outputs. Teachers are therefore asked to rely on evidence generated by systems whose reasoning they cannot fully scrutinise.

Where decision-making is heavily informed by AI, it becomes problematic if decision-makers do not have access to, or meaningful understanding of, the data and processes underpinning the system's assessments. Although teachers are experts in evaluating pupils through classroom observation, participation and assessed work, they may be unable to assess the validity of an AI-generated inference about a pupil's capabilities or future performance. To ensure that existing protections around automated decision-making remain effective, we believe that organisations deploying AI systems should be required to communicate the uncertainty, variability and limitations associated with AI-

[1] https://www.gov.uk/government/news/edtech-and-ai-companies-invited-to-help-build-safe-ai-tutoring-tools-for-disadvantaged-pupils

generated outputs. Human decision-makers can only provide meaningful oversight if they have sufficient information to assess the reliability of the evidence being presented to them.

Maintaining public trust

Public trust in AI-informed decisions will depend on confidence that they are fair, accurate, understandable and subject to meaningful and expert human oversight. Effective mechanisms for challenge, review and appeal also require human decision-makers to have access to information about how an AI output was generated. Without this information, it may not be possible to scrutinise AI-informed decisions or identify errors.

4 Transparency and rights in complex data environments

Transparency should go beyond explaining how data is being processed. In the context of AI, where data is processed in increasingly opaque ways, we believe that users and decision-makers need meaningful information about the reliability of AI outputs to serve the principle of transparency. This should include information on uncertainty, confidence measures, known limitations, performance across different groups, assumptions about deployment environments and evidence of robustness. Where systems are operating outside the conditions for which they were tested, this should be communicated clearly.

In practice, there are limits to explainability for many complex AI systems. Where complete and concise explanations are not feasible, we believe that organisations should be required to provide meaningful information about performance, uncertainty and risk.

5 Effectiveness of data frameworks in regulating AI

Current data protection frameworks remain necessary but are unlikely to be sufficient for regulating AI. Traditional data protection frameworks are primarily concerned with how information about individuals is collected, processed and shared. AI systems, however, generate statistical inferences about people – this creates risks that are not fully addressed by rights over the underlying data. Individuals have rights relating to the data used by a model, but they have fewer practical means of challenging the inferences, predictions or profiling that the model produces. This raises important questions, which must be addressed by regulation, around what levels of reliability are acceptable and whether mechanisms for contesting decisions remain effective when outputs are inherently probabilistic.

Most regulatory debates focus on data governance processes, and we recognise that this is the focus of this call for evidence. However, the gaps in AI evaluation and regulation concern fundamentally statistical questions. Questions of reliability, uncertainty, variation between groups, distribution shift, robustness, and reproducibility cannot be addressed through data governance alone. Effective AI regulation must therefore extend beyond governing data inputs to scrutinising model outputs and real-world performance. Public trust depends not only on compliance with legal requirements, but on confidence that AI systems are accurate, reliable and produce fair outcomes.