ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

RSS RESPONSE TO FINANCIAL CONDUCT AUTHORITY PROPOSAL FOR AI LIVE TESTING

10 June 2025

1 Introduction

1.1.1 This is the Royal Statistical Society's (RSS) response to the Financial Conduct Authority's (FCA) proposal for AI live testing. The RSS is the membership organisation for statisticians and data professionals, with over 12,000 members. We represent the views and expertise of our members – our charitable objectives are to promote the use of data and evidence in decision-making and to promote public understanding of statistics and data.

1.1.2 The RSS has a significant interest in AI. Our membership includes data professionals who are building and deploying AI models, professionals with expertise in data ethics and governance and academics with expertise in computational statistics and machine learning. The RSS has recently launched an AI Task Force to bring these groups together with the aim of i) promoting the relevance of statistics and data ethics to AI; ii) building relationships with government departments and organisations who are making decisions on AI; iii) offering a source of statistical and data expertise to inform decisions.

1.1.3 Our response identifies five blockers to live market deployment of AI models:

a. **Data challenges**: High-quality, diverse datasets are essential but costly and time-consuming to curate; the risks of bias and overfitting persist.

b. **Deployment complexity**: Models often require significant adaptation to fit into existing systems, with ongoing monitoring needed to track data drift and performance degradation.

c. **Evaluation limitations**: Current metrics may not generalize across tasks; over-reliance on benchmark datasets can lead to misleading performance assumptions.

d. **Uncertainty and accountability**: Model performance can vary significantly with small changes; lack of behaviour transparency leads to a reluctance to delegate decision-making to AI.

e. **Lack of standardisation**: Few industry-wide best practices; while standards like ISO 42001 exist, adoption is slow and resource-intensive.

1.1.4 We are supportive of FCA's proposal and the aim to be exploratory without providing regulatory approval. However, there remains a risk that – in spite of disclaimers – organisations will treat the system as providing approval. We also propose a focus on model evaluation – this is especially important in multi-agent environments like financial services where there are serious systemic risks.

2 Question 1: *What are the primary blockers that you encounter prior to live market deployment of AI models? Are these related to technical issues, AI model-related, governance, regulatory or other?*

2.1.1 There are five blockers to live market deployment of AI models that we would highlight.

2.1.2 The first relates to **data challenges**. It can be difficult to curate datasets that are sufficient to train a model to perform a new task: current AI models have a huge number of parameters and are therefore able to learn incredibly complex properties of the data they are trained on. There is a need to train them on data which is large and varied to capture the full statistical range of the target distribution, and to limit the risk of "overfitting" to a small training sample of data. Often creating this data requires manual annotation from humans, which can be costly. Conversely, when models are trained on human developed outputs, we run the risk of building historical biases and mistakes into the model itself.

2.1.3 Second is **deployment complexity**. Deploying a model has many potential pitfalls; most published models are not ready to use "out of the box" for many applications. Besides the bespoke training datasets mentioned above, the models often need to be adapted to fit within the software pipeline of the company using them, on their own compute resources. As a model runs continuously, there also needs to be contingencies built into this deployment – approaches to handle potential downtime when the software crashes or unexpected errors arise, and to monitor how the model performs on new data. The phenomenon of "data drift" is well documented, where the data that a model interacts with in deployment evolves away from that it is trained upon, due to the dynamic and evolving nature of our world. As we travel further from the environment that the model is optimised for, performance can become increasingly erratic. Deploying such models means the developers need an in depth understanding of how to quantify "success" and to detect when we are meaningfully deviating away from it, typically though a strong statistical competency. The validity of any metric for evaluating this performance depends heavily on the application in question, meaning the developers need to

identify the key quantities they wish to track at an early stage. A close and ongoing relationship between statistics practitioners and the end users of the tool are therefore required from model creation and through to deployment.

2.1.4 Third is **evaluation limitations.** The metrics used to train and evaluate AI models can be misleading, or limited to a small range of tasks. There are challenges with only a small number of benchmark datasets used to compare a large number of models, potentially leading to industry wide overfitting on these data sets; choosing the “best model” based on performance on a publicly available dataset risks deploying a model that performs poorly when used more generally . Models are trained to predict the next token, label an image, or similar, but the apparent capabilities of foundation models have led to them being applied on much more advanced tasks, often without this being an explicit training objective. Some datasets allow evaluation on these advanced tasks, but performance is often extrapolated (either consciously or unconsciously) by assuming competence in one skill translates to another.

2.1.5 Fourth is **uncertainty and accountability**. Performance metrics are heavily dependent upon which version of a model is used, and uncertainty intervals are rarely reported and often misunderstood. Absence of such uncertainty makes informed risk quantification near impossible. Changes in which folds of the data are used for testing or different perturbations of the weights can cause large changes in performance, leading to very different potential risks in deployment.[1] Furthermore, decisions still need to be made by humans, and the logic for making them is often obscured. There is a cultural reluctance to give such agency to such unpredictable and novel models.

2.1.6 Finally, there is a **lack of standardisation.** There is little widespread industrial standardisation of best practice in the industry – ISO standards (e.g. 42001) are very helpful here but attaining them is onerous and many companies do not have them. It is likely that take-up will accelerate and become more common (eg, see Anthropic recently achieved 42001) but this approach can be seen as slowing progress in an industry which moves quickly. There may be a role for professional standards (such as the RSS's Advanced Data Science Professional) which seek

[1] More detail is provided in Accounting for Variance in Machine Learning Benchmarks (2021).

to encourage developers to adopt and deploy these practices at an individual level outside of such standardised frameworks.

3 Question 2: In your opinion, would the FCA proposal for AI Live Testing address potential AI deployment challenges? Are there particular areas we should focus on as part of AI Live Testing? This could be either certain types of AI models, AI evaluation techniques, outcome assessment strategies or particular financial services sectors. Is there more we could do?

3.1.1 We are broadly supportive of the FCA proposal, and think that it would help address some of the deployment challenges that we have identified: especially around supporting organisations to address complexities around deployment and manage some of the uncertainties we have highlighted. We would like to see a greater emphasis on model evaluation. Our view is that this is absolutely crucial to aid understanding of how models work when deployed together in a market. Models are often evaluated in a single setting, whereas when they are used in the real world they are increasingly chained together in agentic workflows. It is important that their interactions are evaluated and understood – this is especially important in the context of financial services where there are serious risks.

3.1.2 We would also urge the FCA to give further thought as to how they might clarify that the live testing environment does not amount to regulatory approval for a tool. The current disclaimer reads: "AI Live Testing is designed to be exploratory in nature and does not seek to cover AI auditing, certification, regulatory compliance with other frameworks or corporate governance questions." There could be a more explicit statement that this does not amount to regulatory endorsement. Furthermore, removing any consideration of these significantly important aspects of model use risks devaluing their contributions.

3.1.3 We also support the idea of peer review of model deployment: this could be very positive for deployment best practices. We believe that inviting participants to improve their prospective AI tools through collaborating with impartial domain experts is likely to lead to improved products.

4 Question 3: Is there any other feedback you would like to share with us?

4.1.1 We have a couple of additional points. The proposal references work stating 75% of companies using AI, but only 10% are using it for external customers. We would suggest that this may partly be influenced by the current lack of trust in the technology; companies and

individuals often feel reluctant to give too much control to AI algorithms because their behaviour can seem unpredictable or erratic. There is difficulty within the field on how best to evaluate these models, but there is a further communication gap when describing the strengths and weaknesses of such models to a wider audience.

4.1.2 Greater clarity over the scope of live testing would be beneficial. The current framing leaves some ambiguity about what this will mean in practice. For example, it could be read as proceeding with deployment at an accelerated pace to see what issues arise, which we feel would be reckless. Furthermore, if models are being deployed during this scheme, we would question who is assuming liability for any costs or issues that arise. As there is no mention of utilising digital twins for this testing (despite existing FCA sandbox initiatives), our overall impression is that this is more of a voluntary, light-weight review process. In such a scenario, we wonder if live testing is the best name for the proposal, or whether there is a clearer alternative.