FINANCIAL CONDUCT AUTHORITY

Feedback Statement

FS25/5

Summary of Feedback Received on the Engagement Paper proposing AI Live Testing

September 2025

Contents

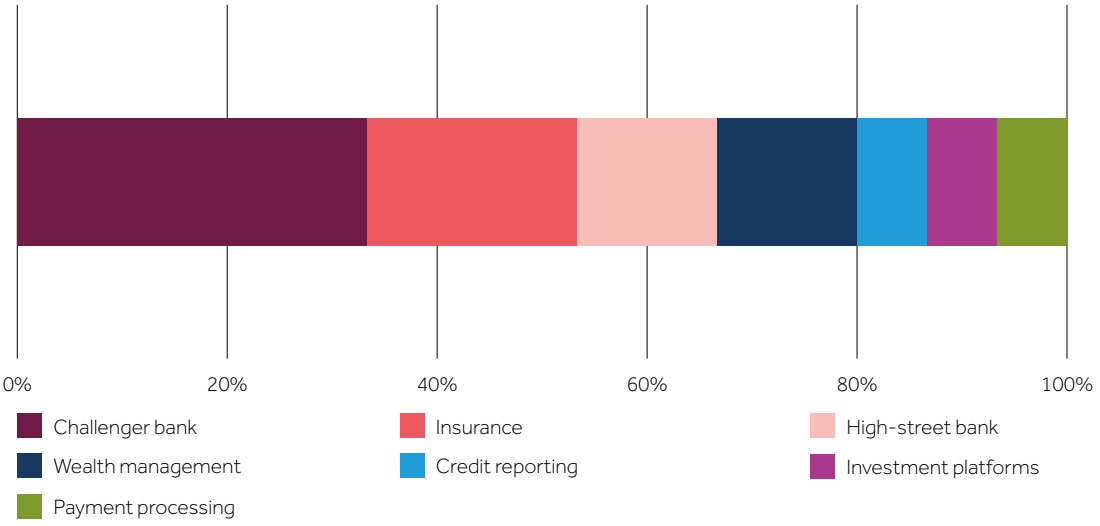
Chapter 1

Executive summary

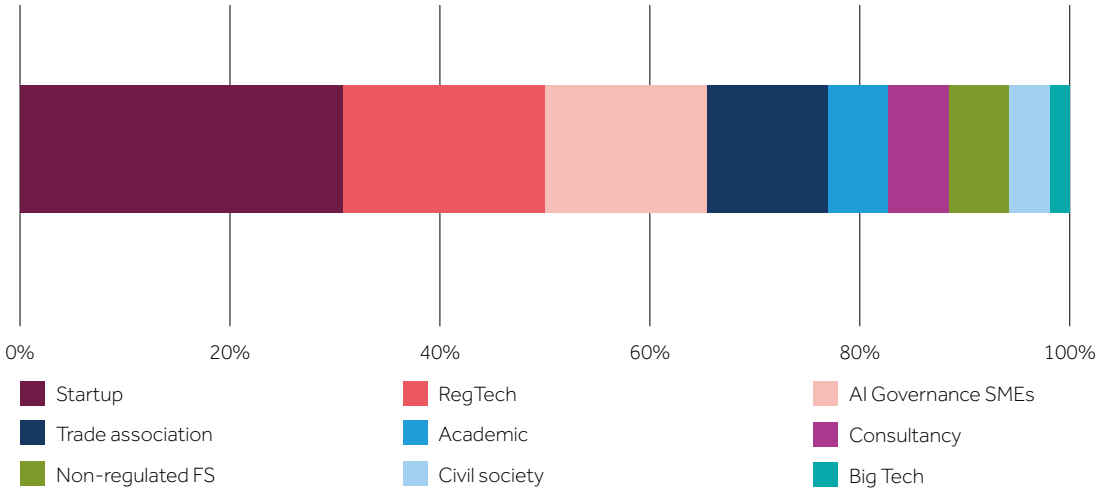
1.1 In April 2025, we published an Engagement Paper on our proposal for AI Live Testing. This aimed to help firms use AI safely and responsibly and achieve positive outcomes for UK consumers and markets. We received 67 responses to the Engagement Paper, representing a broad range of respondents and views (for a list of non-confidential respondents, see the end of this paper).

- The majority of submissions (52) came from a wide range of non-regulated firms. These included consultancies, AI specialist firms, RegTech firms, universities and civil society organisations, startups, trade associations and large technology providers.
- Fifteen regulated firms responded. They included high street and challenger banks as well as insurance firms, wealth managers, investment platforms, payment processing firms and credit reporting firms.

Regulated firms



Non-regulated firms



1.2 This document summarises these responses. It does not set out the FCA's views or policy on the use of AI in financial markets or on the specifics of each individual response. Where feasible, we have used text from the actual responses, shortened to ensure flow and consistency.

Chapter 2

Summary of feedback

Support for AI Live Testing

2.1 Respondents welcomed our proposal for AI Live Testing, who felt it was a constructive and timely step toward increasing transparency, trust and accountability in the use of AI systems.

2.2 Respondents highlighted several key benefits and opportunities:

- **Real-world insights:** Live production testing was seen as a valuable mechanism for understanding how AI models perform under real-world conditions, including system integration, data variability, output quality and user experience.
- **Overcoming Proof of Concept (PoC) paralysis:** Many firms reported that AI PoCs often demonstrate technical merit but fail to progress due to concerns such as regulatory uncertainty and skills shortages. AI Live Testing was seen as a potential solution to this 'last mile' challenge.
- **From principles to practice:** Respondents noted a lack of guidance on how to operationalise and measure key AI principles such as fairness, robustness, safety and security. AI Live Testing could help bridge this gap by providing a structured, repeatable process for assessing the performance of AI systems under real-world conditions.
- **Creating trust:** Respondents emphasised that traditional assurance methods are insufficient in the face of rapidly evolving AI capabilities and that trust in AI must be intentionally designed and transparently demonstrated. AI Live Testing offers an opportunity to redefine AI assurance, not as static documentation, but as observable system behaviour under stress.
- **Addressing first-mover reluctance:** Some firms hesitate to use AI in sensitive areas without greater regulatory clarity. Respondents see AI Live Testing as a key enabler to overcome first-mover concerns and unlock responsible innovation. Other respondents said it could be a way for the market to distinguish between responsible AI use and higher-risk applications, encouraging a culture of safe experimentation.
- **Regulatory comfort:** Being given 'regulatory comfort', potentially through individual guidance or other tools, can substantially de-risk innovative AI use and encourage firms to bring beneficial products and services to market more quickly.
- **Collaboration:** Respondents noted that AI Live Testing is a welcome step forward between the regulator and the industry to jointly navigate the challenges. At the same time, they highlighted the importance of a shared, clear roadmap of the overall objectives for AI Live Testing and that engagement with the FCA and AI Live Testing should be carried out separately from FCA supervisory engagement.
- **Model metrics:** AI Live Testing can foster collaboration and help develop a shared technical understanding between firms and the FCA on complex AI issues such as model validation, bias detection and mitigation and ensuring robustness. Respondents felt this dialogue is crucial for navigating uncharted territory.

Next steps

2.3 We want to give firms the confidence and certainty to invest in AI systems in a way that drives growth and delivers positive outcomes for UK consumers and markets. We also want to work with industry to better understand AI risks and explore effective mitigation strategies, as part of our regulatory and supervisory approach to ensure a safe and responsible future for AI. This is why we have launched AI Live Testing, as part of the existing FCA AI Lab.

2.4 The application window for the first cohort of AI Live Testing was opened on 9 July and extended until 15 September 2025. We will start working with participating firms in the first cohort in October.

2.5 Thus far, the FCA has plans for 2 cohorts, each made up of approximately 5-10 participating firms. We may accept a higher number of firms depending on the range of use cases submitted and how representative they are of the UK financial services sector. The application window for the second cohort will open before the end of the year.

2.6 We want to be transparent with the wider industry about our work. For AI Live Testing, we have committed to publish an evaluation report at the end of the approximately 12-month process. Insights from AI Live Testing may also inform other areas of our work, such as our international engagement, and feed into our publications.

Chapter 3

Detailed comments

Respondents highlighted a number of key issues across the majority of submissions. The sections below provide more detail on respondents' views.

AI deployment complexity

Complexity challenges

3.1 The complexity of AI deployment is a recurring theme in the responses to the Engagement Paper. As respondents noted, this is largely because AI models are not ready to be used 'out of the box' for many applications. In addition to bespoke training datasets, AI models often need to be adapted to fit within the software pipeline. As a model runs continuously, there also need to be contingencies built into deployment. Firms also need to have an agreed approach to handle potential downtime when the software crashes or unexpected errors arise. This is in addition to monitoring how the AI model performs on new data. Respondents highlighted the phenomenon of 'data drift' as a live issue, where the data that a model interacts with evolves away from that it is trained on. Performance can become increasingly erratic as the AI model is used further from the environment it was originally optimised for.

3.2 Practically, this means that AI developers need an in-depth understanding of both the use case and also how to quantify, validate and ensure ongoing model performance. Respondents noted this depends heavily on the use case. Developers need to identify the key model characteristics and intended outcomes early on and then ensure these outcomes are achieved throughout the lifecycle of the AI deployment.

Explainability challenges

3.3 Respondents said deployment is made more complex because AI models often lack explainability. This can make it difficult to ensure AI decision-making processes are understandable, verifiable and auditable. Some respondents also noted the AI model's lack of explainability could make oversight and governance more difficult. This can lead to hesitancy to use AI systems.

3.4 Some respondents argued that AI Live Testing could provide an ideal environment to explore scenarios where robust, reliable, demonstrable and consistently beneficial outcomes might provide sufficient assurance. This could be helpful even if detailed step-by-step explainability of the AI's internal workings is limited.

High consumer impact

3.5 Many respondents felt AI systems that directly interface with consumers, or make significant decisions affecting consumers, should be a priority for AI Live Testing.

3.6 Respondents also noted that testing and evaluating AI solutions specifically designed to identify, support and protect vulnerable consumers is critical. They suggested that the focus here should be on ensuring that AI deployment delivers genuinely fair, accessible and beneficial outcomes without inadvertently creating new forms of exclusion or harm.

3.7 A small number of respondents argued in favour of AI Live Testing developing methodologies for measuring the fairness and accuracy of Large Language Model (LLM) outputs, particularly those involving vulnerable customers and the risk of providing incorrect guidance or taking inappropriate automated actions.

AI model performance and evaluations

Validating and evaluating AI systems

3.8 Almost all respondents highlighted that AI systems are probabilistic, as they do not always produce the same results when given identical inputs. They also noted a key operational business challenge is determining the level of testing needed to ensure the model operates as expected, including the extent of ongoing monitoring.

3.9 There was some disagreement on the practical feasibility of adequately evaluating and validating AI models. A few respondents noted a conflict between wanting to use more capable machine learning models and the ability to adequately evaluate and validate them. This includes ensuring models will not lead to harmful outcomes for consumers and markets. Respondents also noted that, while model validation techniques can be useful, actual model behaviour can be very different from previous tests, including the validated behaviour.

3.10 Respondents said that deciding when an AI model is ready for live deployment can be subjective and challenging. For example:

- Which metrics define the right outcome and what range of accuracy is appropriate? For example, is 90% accuracy in predicting an outcome sufficient compared to 95%? And how should this differ across use cases?
- Often, technical teams will benchmark the AI against an existing human process. If the model can perform 'equally or as good as' a human, does that suggest it is ready to be deployed? Or does AI model performance need to be significantly better than humans?

3.11 Respondents asked how dependent the AI model metric is on context. A related question is whether the AI model still holds up with real customer behaviour – or whether that is true for lab conditions only. And what happens if market conditions shift? Can the model adapt fast enough? Equally important is the risk of relying too much on historical data.

3.12 Respondents also said it is currently unclear how financial services rules and regulations read across to AI model design, validation practices and monitoring. A large number of respondents noted this is particularly challenging in high-impact areas (defined by respondents as credit decisions, financial advice or fraud detection, ie anywhere where

the AI model could directly affect consumers' access, cost or protection levels). In these areas, minor deviations can lead to disproportionate outcomes in AI system behaviours. More generally, respondents noted that existing compliance frameworks are not generally adaptable to AI-specific risks.

3.13 Some respondents pointed out there is a lack of explicit guidelines, whether on standard AI model validation techniques or, more specifically, how financial services rules and regulations apply to AI models. Without clear direction, firms err on the side of caution or tend to rely on internal governance bodies to approve a model as fit for release. The result can be too focused on worst-case scenarios.

3.14 Some respondents noted that AI models should be evaluated not just for their performance but for their impact on financial markets more generally. They also suggested that an equally important question is: how can digital trust be maintained in financial markets? AI Live Testing, some suggested, could help firms and us jointly explore the downstream consequences of AI model decisions.

3.15 Some respondents went further, arguing that we should prioritise decision-critical AI systems (ie AI models that influence financial decisions or customer outcomes) as part of AI Live Testing.

3.16 Respondents also noted the significant cost and procedural complexities in undertaking validations for each AI model. These range from the computational cost of running tests to the procedural challenges of making tests repeatable relative to meaningful risks. Respondents also noted that many established financial institutions find integrating modern AI capabilities with existing IT infrastructure a substantial technical, practical and financial hurdle. Older systems are often siloed and lack the flexibility required for seamless data flow and interaction with AI systems – particularly those reliant on cloud computing.

3.17 A small number of respondents noted that, while AI models are often evaluated in a single setting, when they are used in the real world they are increasingly chained together in agentic workflows. So, AI validation approaches will increasingly need to include validation of interactions.

3.18 Similarly, a few respondents highlighted that AI systems do not understand context. This raises concerns about low probability/high impact events and their potential implications for consumer outcomes, market integrity and financial stability more generally. These are also difficult probabilities to test for.

Developing standards

3.19 Many respondents argued we should support the development of industry standards for AI design and use. Most respondents focused on process suggestions and did not specify what exactly these industry standards should focus on. For example, they suggested we could convene cross-functional working groups or training forums that build mutual understanding and learning opportunities. This could also provide guidance on how current regulatory expectations map to AI-driven systems. Submissions did not further specify where guidance would be helpful.

3.20 Respondents also noted that standardised AI evaluation techniques, including outcome assessment strategies, could be used for regulated firms' engagement with third party model providers (see below for more on third parties).

Adversarial testing

3.21 A significant number of respondents said we should emphasise developing and applying rigorous methods for stress-testing AI models. This includes against unexpected or adversarial data inputs, evolving market conditions and potential cybersecurity threats, to ensure their ongoing reliability and safety. Many respondents pointed out that all AI models have inherent fragilities, which go undetected during tests and are encountered at runtime. These fragilities can possibly also be exploited by adversaries.

3.22 Some argued that malicious actors and financial crime should be a priority for AI Live Testing. They argued it offers a critical opportunity to move from conceptual AI assurance frameworks to practical, threat-informed AI stress-testing.

Data

Data challenges

3.23 Respondents were clear that reliable, safe AI systems fundamentally depend on the quality, representativeness and ethical handling of the data used for the AI systems' training and operation. This also includes ensuring data accuracy, minimising biases in datasets, complying with data protection regulations and managing consumer consent effectively.

3.24 Some submissions highlighted the challenge of getting sufficient and high-quality data necessary for robust AI model testing and optimisation.

Synthetic data

3.25 A number of respondents suggested we expand collaboration on the use of synthetic data to help address data scarcity and privacy concerns in AI development. They also suggested actively involving consumers in the testing process to get deeper insights into real-world impacts.

Accountability ambiguity

Governance frameworks

3.26 Many respondents explained the challenges they face in designing and implementing effective governance frameworks to manage the risks of AI deployment in financial markets.

3.27 Some said that clearly defining who is responsible for the AI model's ongoing use, performance monitoring and maintenance post-deployment is often a challenge. Without clear accountability structures, organisations struggle with effective governance, risk mitigation and responsiveness to emerging issues. This is compounded by the fact that it is not always clear where risks will emerge, including how best to mitigate them.

3.28 Respondents pointed out there is currently a lack of an agreed governance and oversight framework, including the role of human-in-the-loop. Risk committees and legal and governance groups understand the technical issues of privacy, security, resilience, etc., while AI model accuracy is less well-defined. It is similarly important that risk committees and legal and governance groups are more risk averse.

3.29 Another key challenge is the time needed to assess AI models' overall risk and whether they are in line with corporate risk appetites.

3.30 A few respondents pointed to the Senior Managers & Certification Regime (SM&CR), arguing that the ambiguity surrounding what constitutes 'reasonable steps' slows governance processes down. This means senior managers may lack the comfort to proceed with using AI, fearing personal repercussions. This acts as a disincentive, particularly for more advanced AI models. One firm noted that, anecdotally, there is an expectation that an AI solution is perfect or near-perfect even if expected results exceed what a human could reasonable achieve.

3.31 Another theme highlighted was the need to bridge AI governance with operational resilience. AI technology has become a fundamental part of processes that interact with customers and respond to market dynamics. So, respondents suggested, operational resilience processes need to be updated to ensure that:

- AI model reliability and degradation risks are considered as part of resilience planning.
- AI failure testing is built into business continuity assessments.

3.32 Many responses noted that building senior leadership confidence in AI reliability and value proposition has a long lead time. They also suggested there is considerable market skepticism about AI reliability in financial contexts, particularly for high-stakes decisions.

3.33 Many respondents argued that AI system materiality thresholds should be established, based on AI model decision criticality and autonomy levels instead of statistical risk scores. They suggested the relevant governance frameworks, including operational resilience approaches, could be developed on the basis of these AI system materiality thresholds.

AI model choice

3.34 Respondents consistently raised the issue of the choice of model, including whether to develop in-house or source from third parties. This is particularly relevant given the general lack of model transparency, particularly from third parties. For example, how can a purchasing firm know whether the model will operate as expected? What can be done to create incentives to model providers to be as transparent as possible? What levels of transparency are required from a regulatory perspective?

3.35 According to some respondents, this is a particularly acute issue when complying with data protection requirements. The 'black box' nature of AI systems remains a key challenge. As more models from AI third party suppliers are onboarded, the underlying logic of how the AI arrives at decisions is partly or fully unknown to the deployer. Respondents said this can create challenges in meeting regulatory requirements to evidence how AI has reached a decision. Third party suppliers may be unable or unwilling to provide details on algorithms, known biases in training data or margins of error. Respondents said that in such cases it becomes difficult for regulated firms to fully explain the decision-making process. Similarly, it becomes difficult to assess and mitigate conduct, data privacy and third party risks from AI models without knowing their internal workings.

3.36 Respondents also pointed out that the use of third party models carries regulatory risk, if the FCA considers there has been insufficient testing or the regulated firm did not do enough to fully evaluate the underlying AI model.

3.37 Equally, respondents raised concerns about working with small AI vendors. Many FinTechs pioneering in AI are new or small-scale. While their innovation is welcome, regulated firms have concerns about operational resilience, governance standards and the maturity of these providers. These concerns create barriers to collaboration.

3.38 Respondents pointed out that AI models are increasingly embedded into third party products, but are lagging in their ability to interrogate the provenance, safety and controls of embedded AI. They said this is particularly true where large models are continuously updated without user visibility.

3.39 Some respondents noted that limited third party transparency increases the challenges of AI model evaluations and validation techniques. As a result, firms currently tend to rely on internal models only as these can provide the required levels of insight into the AI model's limitations, risks and behaviours.

3.40 Respondents also suggested that ensuring third party AI models are safe and perform as intended is a major pre-deployment step and has direct relevance to AI governance and regulatory compliance. With this in mind, a few respondents suggested that AI Live Testing should actively explore what firms' responsibilities are when vendors are unwilling or unable to provide sufficient transparency. This includes around algorithmic decisions, known biases, margins of error, safety and data lineage.

3.41 Respondents pointed out that a lack of AI model transparency has implications for liability. This uncertainty can inhibit AI adoption, particularly where outcomes may directly affect consumers.

Chapter 4

Suggestions for next steps

4.1 Whilst there was strong support for AI Live Testing, respondents also provided views on how the FCA could further help firms' safe and and responsible AI adoption. Respondents' views are summarised below under the following themes: (a) AI model evaluation and validation, (b) AI scenario testing, (c) fairness and (d) collaboration.

1. AI model evaluation and validation:

- **Performance evaluation:** Some suggested we should develop standardised performance benchmarks in finance for AI. We could supplement this with anonymised benchmark data – eg performance ranges observed for similar use-cases. Respondents argued that a regulator-brokered benchmark has credibility and can meaningfully inform AI model performance evaluations.
- **Data:** Some respondents also noted we could facilitate data availability and data sharing to address AI bias concerns.
- **Outcome-focused monitoring:** Some respondents argued there should be less focus on the AI model and more on outcomes, ie who gets accepted, denied, flagged, or scored – and why? In other words, we should encourage firms to track outcome drift, not just model drift.
- **AI model assurance:** Some respondents said we should do all of the above and move into comprehensive AI assurance assessment and validation. This could include longitudinal post-deployment monitoring, mechanisms for drift detection, bias monitoring and user feedback, with firms submitting anonymised performance data to an FCA-monitored AI Incident Reporting Repository.
- **Environment, Social and Governance (ESG):** Some respondents suggested incorporating environmental metrics within AI Live Testing would encourage innovation directed at green initiatives. This could also provide the opportunity to assess the lifecycle carbon footprint of AI systems and encourage responsible AI model design choices. In addition, integrating environmental considerations with AI Live Testing would signal that technology must account for its ecological impact. Setting these benchmarks early in the development cycle would encourage innovation and environmental responsibility in finance.
- **AI model materiality thresholds:** As not all AI deployments carry the same risk, some respondents suggested graduated explainability and transparency requirements, depending on the AI use-case's risk profile.

2. Scenario testing:

- **Stress-testing under real-world pressure:** A small number of respondents argued we should simulate how the AI model performs in a crisis, including situations such as market volatility, regulatory change or data shift. Respondents asked if the AI model can adapt without reinforcing old biases or crashing entirely.

- **Safe harbour:** A small number of firms argued in favour of creating a safe harbour for disclosure with firms that share AI Live Testing findings, including failures, to receive proportionate regulatory comfort.
- **Transition from testing to operational deployment:** Some respondents said that the shift from controlled test environments to real-world deployment scenarios presents substantial challenges. Simulating realistic operational conditions to accurately predict long-term or downstream impacts is inherently difficult, making it challenging to establish precise, actionable metrics for evaluation and monitoring.

3. Fairness and bias mitigation:

- Some respondents suggested we should clarify our regulatory expectations, so that firms better understand what they can do without introducing unintended discrimination or regulatory risk. AI Live Testing can help define what responsible bias mitigation looks like – especially in areas like vulnerability detection, product suitability and onboarding.

4. Collaboration:

- **Shared learnings:** Most respondents agreed we should continue to actively promote a culture of experimentation and shared learning around AI in financial services. There were repeated calls for publicising success stories from AI Live Testing, including publication of anonymised case studies. As one respondent put it, 'firms are often more convinced by peer examples than theoretical advice'.
- **International standards:** A small number of respondents recommended that we consider integrating AI Live Testing with international efforts such as the National Institute of Standards and Technology's (NIST) AI RMF or Singapore's AI Verify to support global interoperability.
- **Research:** Other respondents highlighted the importance of thought leadership and evidence-based approaches. To ensure AI Live Testing delivers enduring value beyond compliance validation, a small number of respondents recommended anchoring the programme in cross-disciplinary, evidence-based research, including partnering with UK universities. Embedding thought leadership into the core design of AI Live Testing would reinforce its credibility, accelerate research-to-practice translation, and strengthen the UK's position as a leader in safe and responsible AI use in financial services.

4.2 We will consider these suggestions in more detail as we take forward our work supporting firms' safe and responsible AI adoption. Based on the responses, we are proceeding with AI Live Testing.

Annex 1

List of non-confidential respondents

The following respondents agreed to the publication of their names:

4admin

Adclear

Advai

AdvisoryAI

Automwrite

Behavox

Chip Financial

Datambit

Dynamic Planner

EV

ExoBrain

Features Analytics

Freetrade Limited

i-confidential

Innovate Finance

Kroo Bank

LTIMindtree

Model Office

Moneybox

mycomplaints.ai

Natwest Group

Open Insurance Services

Palindrome

Personal Investment Management & Financial Advice Association (PIMFA)

Plenitude Consulting

Qxplain

Royal & Sun Alliance Insurance

The Royal Statistical Society

Safe Intelligence

Strathclyde University Financial Regulation Consortium

Sumsub

Techno-Regulatory AI Sandbox (TRAIS) project (RAI UK)

The Investing and Saving Alliance (TISA)

UK Finance.

FINANCIAL CONDUCT AUTHORITY

12 Endeavour Square London E20 1JN
Telephone: +44 (0)20 7066 1000
Website: www.fca.org.uk

Pub ref: 2-008501

All our publications are available to download from www.fca.org.uk.

Request an alternative format

Please complete this form if you require this content in an alternative format.

Or call 0207 066 1000

Sign up for our **news and publications alerts**