

**HOUSE OF COMMONS
BUSINESS, INNOVATION, SCIENCE, AND TRADE COMMITTEE INQUIRY INTO
UK'S ECONOMIC STRATEGY FOR AI**

8 October 2026

Written Evidence to the Committee from Meta Platforms, Inc.

Thank you for the opportunity to provide oral evidence at the Committee hearing on 13 October 2026. In advance of that hearing, Meta Platforms Inc provides this written statement.

Our vision is to bring personal superintelligence to everyone: an exceptionally capable personal agent that understands you, your goals and what you care about and that can take actions to help you achieve your full potential. We are fortunate to live at an incredible moment in history and in the next few years people will use superintelligence to create and discover extraordinary new things. Our latest step on this journey is Muse, our personal AI agent which recently launched in the US and Canada and which we hope to bring to the UK soon. We're gratified by the reaction we have seen so far to Muse, which shows people's interest in using AI not simply to answer questions but to complete tasks on their behalf, returning meaningful time to them and helping them to achieve their personal goals.

For people and businesses in the UK, the benefits are already real. For example, working in partnership with a number of UK sight loss and low vision organisations, Meta AI in our Ray-Ban Meta glasses is helping a significant number of blind and partially sighted people across the country to complete everyday tasks, such as navigating to work, which were previously out of reach. Getting the policy right matters: building technology with safeguards in place is critical to not only preventing harm but ensuring that people can trust this technology enough to rely on it. At the same time, holding back progress across the board would delay these benefits for British consumers, businesses and public services without making anyone safer.

Realising this ambitious goal will require developing capable AI models that are reliable, robust, and secure. Every company has the responsibility to move only at the pace required to train and deploy its models safely and to take action at every stage to assess and mitigate risks of model development and deployment. At Meta, we are committed to fulfilling that responsibility, as reflected in our recent decision to delay the release of Muse by several months in order to address safety and security. Decisions about pace are already an integral part of how we build, as they should be for any company developing advanced technology.

I. AI Governance and Accountability

We approach safety holistically, taking a layered and risk-based approach across all our AI technologies. Central to this is our internal AI risk program, which involves multi-disciplinary engagement, including internal and, where appropriate, external experts from various disciplines—such as engineering, product management, compliance and privacy, security, legal, and policy—as well as company leaders.

Accountability at Meta is designed not only to be robust but also to be demonstrable, which is why we publish our [Meta Superintelligence Scaling Framework](#), which establishes the processes we use to evaluate our frontier models for catastrophic risks as well as the standards we use to decide whether to stop or slow development of models or withhold their release altogether. We released the latest version of our Framework earlier this month and have published earlier iterations for the past several years. Likewise, we routinely publish [preparedness](#)

[reports](#) that describe in detail our internal and external testing to evaluate our models prior to release as well as our findings, so that anyone can understand and check our work. It's why we continue to release models openly, so that governments and experts can do their own evaluation of the quality and safety of our models. And it's why we have operated a [bug bounty program](#) for more than 15 years to fund researchers who identify and responsibly report vulnerabilities to us so we can fix them before they can be exploited.

These efforts are part of a broader, comprehensive risk management program that is designed to review changes to our models and products as they are made, adopt appropriate mitigations to address any risks we identify, and release changes when we're confident it is both safe and beneficial to do so. This work goes well beyond catastrophic risks and involves continuous improvement. For example, with the release of our recent Muse Spark family of models, we [adopted a new approach](#) that translated our trust and safety guidelines across areas like content and conversational safety, response quality, and handling different viewpoints into clear, testable principles. We also trained the model on why something is safe - not just on the rules, but also the reasons behind the rules. This means the model is better equipped to handle novel situations that rules-based systems might have failed to anticipate, and helps guard against not only foreseeable risks but risks we don't foresee as well.

Taken together, this work reflects our efforts to ensure that the benefits of powerful new technologies are responsibly diffused across society, taking into account the novel benefits and challenges that our most advanced AI models present. Our goal is not just progress for the sake of progress, but instead innovations that give people new tools to improve their lives and achieve their potential.

II. Open Approach

AI should be for everyone, no matter who you are, where you live or what language you speak. This is why we have been committed for nearly two decades to open sourcing AI and other technology so not only Meta but the broader world can build on and benefit from them. To date, we have made more than 1,000 open source projects available.

We believe that this approach offers three key benefits: First, taking an open approach to model deployment is part of how we bring the benefits of AI to everyone. It removes the barriers of cost, scale and access that have historically concentrated AI's benefits among a few. Second, open source enables faster societal progress because other organisations can learn from and improve on our technology, and in turn allows us to learn from and improve on the technology that others develop as well. Third, open source has been a critical component of information security for decades because it opens technology to broad pressure-testing, which gives society confidence in the robustness of its protections. Indeed, the importance of this dynamic is reflected in the fact that open source protocols and encryption literally form the backbone of today's Internet.

At the same time, we recognise that open-weight models may be exploited in ways that APIs cannot, including through fine-tuning refusal behaviour away. These considerations are reflected in our assessments for open-weight models, which consider the unique risk vectors of such releases and plausible deployment contexts of such models. The most recent update to our Meta Superintelligence Scaling Framework specifically commits us to consider the ways in which a model will be used in its specific deployment context as a part of our risk assessment and mitigation efforts.

III. Meta's Approach to Frontier AI Safety

Our [Meta Superintelligence Scaling Framework](#) takes an outcomes-led approach to assess and manage the capabilities of our most advanced models in a systematic, evidence-based manner. It does so by structuring our evaluation and mitigation approach, including risk thresholds, around an identified set of catastrophic outcomes that we work to prevent. More specifically, the Framework requires us to assess whether a frontier model would substantially contribute to one or more threat scenarios which the Framework identifies as potentially sufficient to realise a catastrophic outcome. These threat scenarios represent causal pathways to outcomes, which are determined through systematic threat modelling exercises.

We leverage external expertise to validate and strengthen our evaluations, whether that is assisting in threat modelling, co-designing internal evaluations, or running additional evaluations. For example, we have engaged experts from Deloitte and from the UK-based firms Faculty and Apollo Research.

We also continue to invest in testing, measurement, and alignment techniques in order to maintain the robustness and reliability of our evaluations and models as AI capabilities advance, helping us to predict and shape how our models will perform once deployed.

Our work does not stop once a model is deployed. Following the deployment of our models into our AI offerings, we continue to monitor live traffic to try to detect any material changes in behaviour of the model or our mitigations so we can remediate any identified issues, as appropriate. Additionally, effective risk management requires preparation not only for ongoing operations but also for unexpected tail events. To do this, we maintain a comprehensive global incident response programme, including identifying incidents from both internal and external sources, and reporting critical incidents as indicated under our Framework.

We recently underlined our commitment to robust frontier AI governance practices by signing onto the White House Accord on Super Intelligence: a set of principles that set a floor for safe development, including robust internal controls, independent external evaluation of whether those controls are operating as intended, and independent Board committee oversight.

We continue to evolve our standards by making these specific, verifiable commitments. Over the coming months, we are also making additional changes to our Framework and to our overall approach to AI governance. We will establish a new AI committee of our Board of Directors, which will review future changes to our Framework and independently ensure that our operations conform to the standards we have set.

IV. Model Containment

Recent events have demonstrated the capabilities of today's AI models to find and exploit software vulnerabilities in ways which require proportionally stronger containment during testing. This is an industry-wide challenge that labs, including Meta, are actively working to address, including through improved containment during testing, better separation between evaluation and production environments, and ongoing research into how to evaluate safely without reinforcing unwanted behaviours in the model.

To that end, we have extended our Framework's loss of control requirements to cover AI containment during training and evaluation. Before any reinforcement learning training run, we will determine whether the capabilities of the model being trained could potentially compromise its environment. If so, we will require specific mitigations to be in place prior to beginning the run:

- Models with cyber capabilities comparable to models currently considered frontier will run in sandboxes approved through a standardised internal process that vets them for vulnerabilities. Sandboxes with high-severity vulnerabilities are mitigated or excluded before the run begins. Environments are also checked for features that invite reward gaming, such as exploitable grading or impossible tasks.
- Model outputs, including full trajectories and chain of thought, are written in real time to secure logging systems that prevent them from being modified or deleted after the fact.
- In addition to our on-call teams tracking models during training, automated systems will monitor all rollouts of high-risk models capable of halting an individual training or evaluation rollout when they detect high-severity misbehaviour or attempts to escape containment. These monitors will also be stress-tested to help establish that they would succeed in detecting unauthorised actions by our models.

We are also implementing independent verification requirements for test environment isolation and scenario review before evaluations begin. Finally, we are contributing to industry-wide efforts to help improve security and evaluation safety practices, model alignment, and how models behave in adversarial testing scenarios. We plan to continue to strengthen how evaluations are conducted in a secure manner going forward.

V. Conclusion

Superintelligence will be among the most important technologies in history, and its opportunities and challenges will likely be greater than any we have seen in our lifetimes. Every transformative advance has been accompanied by concern that people would be left behind, and on each occasion a greater number of people have come to share in increased prosperity, health and freedom. We believe this same outcome is achievable for superintelligence and we are committed to the work required to secure it, including for people and businesses across the UK. We look forward to discussing these important issues with the Committee.