Department of Commerce
Bureau of Industry and Security

Re: *Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters*; BIS-2024-0047

Introduction

We are AE Studio, a bootstrapped 150+ person data science, dev, and design consultancy. We reinvest profits into impact-focused foundational research, like neurotechnology and AI alignment, with the mission to increase human agency.

Our company works with frontier models across a wide array of clients within and beyond the tech sector, from the largest corporations to startups outsourcing their first technical projects to us, across a wide range of applications. To survive in our industry, we excel in our client work, with the best technical talent keeping up to date with the cutting edge of AI capabilities.

We are thankful to the Bureau of Industry and Security for their work concerning the future of AI. We believe AI has a highly uncertain future, and in some possible scenarios, perhaps inevitably, advanced AI models will be immensely capable and essential for national security and for the flourishing of our economy.

But the possible risks need to be managed. We should not entirely trust our future, the next nuclear technology, to the private sector, yet we should not also preemptively stifle essential innovation. We must defensively accelerate, and America must win the AI race, should there be one.

Monitoring the state of the art in the AI labs of America is the only way the US will be able to ensure we both: lead on the technology, and mitigate risks from the technology. Doing so with minimal friction such that we perfectly balance safety with winning, is our task.

Overview

Enhancing Reporting Requirements for Dual-Use Foundation Model Developers

We present several proposals to strengthen the effectiveness and reliability of reporting requirements for entities developing dual-use foundation models. Our focus is on recommendations that the Bureau of Industry and Security (BIS) can implement within its current authority, prioritizing ideas that require minimal additional resources.

Key Recommendations

1. Establish a Protected/Anonymous Reporting Channel: Create a secure mechanism for employees of entities developing dual-use foundation models to report concerns. This channel (ai_reporting@bis.doc.gov) would allow staff to disclose:
   a) Potential inaccuracies or misleading information in company reports to BIS
   b) Safety and reliability issues related to dual-use foundation models
   c) Activities or risks that may impact U.S. national security

   Ideally, this platform should be both protected (prohibiting retaliation) and anonymous. If protection is unfeasible, an anonymous system would still be valuable. Entities should confirm in their reports that employees are aware of this mechanism and that company policies do not hinder its use.

2. Implement Regular Employee Interviews: Conduct quarterly interviews with staff from entities producing dual-use foundation models. BIS would select interviewees from various teams to gather diverse insights. These conversations would cover model capabilities, safety and security concerns, and predictions about AI advancements that could pose new safety and security risks.

3. Require Capability Forecasts: In addition to red-team testing reports, companies should provide their best estimates of when they or others might develop dual-use foundation models with specific security-relevant capabilities. This would assist the U.S. industrial base and defense sector in making informed predictions about future AI advancements and their defense implications.

4. Mandate a Summary Form for Non-Experts: Alongside detailed reports, require entities to submit a concise Summary Form accessible to non-technical audiences. This form would highlight the most critical defense-relevant information in a clear, easily digestible format.

5. Modify Notification Conditions: Require entities to inform BIS of significant capability improvements that present immediate security risks within 5 days of discovery. This ensures BIS is promptly notified of crucial advancements between quarterly reports.

We suggest applying these recommendations to all entities developing dual-use foundation models. However, if this proves impractical, focusing on the top 10 entities would still yield valuable information for U.S. industrial and national defense interests.

Detailed Explanations and Rationales

1. Insider Reporting Mechanism

Overview: BIS should maintain a channel for employees of dual-use foundation model developers to report safety, security, and compliance concerns anonymously and, ideally, with protection from retaliation.

Justification: Insiders are often the first to identify safety and security risks in dual-use foundation models. Their insights can help BIS understand the impact of these models on industrial and national defense interests. Insider reports can:
- Reveal potential inaccuracies or violations in company reports
- Provide diverse viewpoints when company leadership and staff disagree on safety objectives or security concerns

Evidence suggests some entities have used restrictive agreements to prevent former employees from voicing concerns. A reporting mechanism would provide a clear pathway for insiders to share information despite such pressures.

Legal Basis: The Defense Production Act (DPA) authorizes the President to obtain necessary information from any person to support national defense. BIS has implemented similar confidential reporting tools for export control violations and boycott compliance.

2. Regular Insider Interviews

Overview: BIS should conduct regular interviews with employees or contractors from entities developing dual-use foundation models, focusing on safety, security, capability forecasts, and voluntary commitments.

Justification: Interviews provide a structured way to gather insider information, complementing the reporting mechanism. They ensure regular communication and timely updates on potential risks and advancements.

Feasibility: This process can be efficient:
- Employee selection: Approximately 3 hours per entity to review staff lists and choose interviewees
- Conducting interviews: 50 hours quarterly for 10 employees from each of the top 5 developers
- Minimal participant burden: Questions align with employees' regular work, requiring little to no preparation

Legal Basis: BIS has the authority to conduct interviews under the DPA and has used similar methods in section 232 investigations.

3. Capability Forecasts

Overview: Entities should provide estimates of when dangerous AI capabilities might emerge, covering areas such as CBRN threats, cyber capabilities, persuasive abilities, autonomous actions, AI R&D contributions, and novel WMD development.

Justification: Understanding industry projections for defense-relevant capabilities is crucial for preparing the U.S. industrial base and anticipating AI impacts on national defense.

Legal Basis: BIS has the authority to request such information under the DPA.

4. Summary Form

Overview: Entities should submit a concise Summary Form alongside detailed reports, providing key information in an accessible format for non-technical audiences.

Justification: This form ensures that critical defense-relevant information is clearly presented, improving the clarity and robustness of reporting requirements. It allows both technical experts and defense leaders to quickly grasp essential information.

Feasibility: The form would primarily cover topics entities already consider in their work, minimizing additional burden.

Legal Basis: BIS already has the authority to send questions to covered U.S. persons about dual-use foundation model safety, reliability, and national security concerns.

5. Amended Notification Conditions

Overview: In addition to quarterly reports, entities should notify BIS within 5 days of discovering advancements that could lead to imminent national defense or security threats.

Justification: Rapid AI progress could render quarterly reporting insufficient. Some breakthroughs might occur without developing new models, such as unlocking new capabilities in existing systems.

Legal Basis: This recommendation modifies existing reporting requirements to include provisions for sudden or rapid AI progress relevant to U.S. national defense interests.

Additional Suggestions

1. Compute Cluster Reporting: Require entities to report plans for constructing or obtaining services of cutting-edge compute clusters or data centers.

2. Cyber Attack Reporting: Mandate reporting of cyber attacks (successful or near-miss) that compromise model security, weights, or algorithmic secrets.

3. Security Level Reporting: Require entities to report current and planned security levels based on the RAND Corporation's recent report outlining five security levels for frontier AI organizations.

Conclusion

These recommendations aim to enhance the robustness and effectiveness of BIS's reporting requirements for dual-use foundation model developers. We've focused on suggestions that BIS can implement within its current authority and with minimal additional resources. These measures will contribute to better understanding and preparedness for AI-related impacts on national defense and industrial capabilities.