HR0011ST2025D-05
Model-Agnostic Detection of Bias (MAD-Bias)
Frequently Asked Questions (FAQs)

1. What specific types of bias (e.g., racial, gender, cultural, data imbalance) are expected to be addressed in this project?
   Answer: any subset of those is fine

2. Are there specific domains (e.g., human data, project description, policy) where the proposed technology should be prioritized for bias detection?
   Answer: no preference

3. Is the primary focus on detecting bias in structured data, unstructured data (e.g., natural language or images), or both?
   Answer: either one

4. How will this project account for biases that arise during data collection versus biases that arise during data representation? (e.g., AI model biases vs data biases)
   Answer: this is up to the performer

5. Will the government please elaborate on what types of geometric and topological data should be addressed to detect bias in datasets? Are there specific metrics or features that these methods will measure?
   Answer: this is up to the performer

6. Will the approach also include the identification of causal relationships contributing to bias, or will it be limited to statistical detection?
   Answer: this is up to the performer

7. What specific outcomes are expected in the proof-of-concept phase (e.g., bias detection rates, accuracy metrics)?
   Answer: it is the responsibility of the performer to propose the outcome and why it is meaningful when compared to existing methods, as well as what its downstream impact would be.

8. Should the live demo in Phase I use synthetic data, real-world datasets, or both?
   Answer: real-world data is preferred, but if the domain of interest has no reasonable datasets, synthetic data is fine.

9. If real-world datasets are to be used, are there any restrictions or recommendations for data sources?
   Answer: N/A

10. What are the expectations for scalability?
    Answer: this is up to the performer

11. Should the solution handle specific data sizes or adapt to particular use cases?
    Answer: no preference

12. Does the government expect the developed model/methodology to integrate with existing tools or software as a deliverable?
    Answer: it is up to the performer to identify existing tools to be integrated into.

13. Are there particular sustainability metrics (e.g., maintenance costs, computational efficiency) that the solution must meet?
    Answer: No

14. The topic description mentions "geometric and topological data analysis" as potential approaches. Would you please clarify if DARPA is specifically interested in methods that can analyze the topology of the training data manifold to detect inherent bias, independent of the downstream model architecture?
    Answer: such methods would be within scope but are not required.

15. In reviewing both the Jha et al. 2022 paper on LLM proper noun bias and the D'Incà et al. 2024 paper on open-set bias detection, we notice different approaches to quantifying bias severity. For the purposes of this topic, should proposals prioritize developing new bias metrics that are model-agnostic, or focus more on methods to detect previously unknown types of bias?
    Answer: both are within scope.

16. The topic emphasizes bias detection regardless of downstream AI model architecture. Would you please clarify if the scope should include developing methods to detect bias in the raw data distributions themselves, before any model training occurs?
    Answer: that is within scope as long as it is model-agnostic.