HR0011ST2025D-02

Mitigating Explicit and Implicit Bias Through Hybrid AI

Frequently Asked Questions (FAQs)

1. What specific types of biases (e.g., gender, racial, cultural) are being targeted in this proposal?
   Answer: any form of social bias.

2. Should the focus be on mitigating bias in training data, model design, or outputs—or all three?
   Answer: any subset of those.

3. What balance should be prioritized between mitigating bias and preserving accuracy, and how should this balance align with the major focus of this proposal?
   Answer: it is up to the performer to define the bias-accuracy trade-off

4. What types of symbolic representations are most relevant for this project (e.g., grammars, ontologies, logic-based systems), and how might they specifically address bias?
   Answer: any of those is fine. How they might specifically address bias is up to the performer to propose.

5. Which neural architectures are preferred or most applicable for this proposal (e.g., transformers, recurrent networks), and why are they suitable for mitigating bias?
   Answer: any model that has relevant applications, e.g., attention networks, KANs, LRMs…

6. Are there any preferred datasets or application domains (e.g., text, image, speech) for demonstrating bias mitigation, or should the proposal cover multiple domains?
   Answer: no preference

7. Should the proof-of-concept focus on re-training an existing AI model, or is it acceptable to propose methods that primarily modify or constrain the inference process (e.g., decoding)?
   Answer: both are in scope.

8. How should success in mitigating explicit and implicit bias be measured in this proposal, and are there specific metrics or benchmarks recommended for evaluation?
   Answer: It is up to the performer to define those metrics.

9. To better focus the proposed techniques on the biases most relevant to the project's goals and ensure alignment with the application domain, could you clarify the specific types of explicit and implicit biases that this project aims to address in safety applications? For instance, in counter-UAV or swarm UAV operations of military cases, biases could emerge in threat classification, anomaly detection, or target prioritization, potentially leading to false positives/negatives due to data imbalances,

adversarial spoofing, or operational constraints. Are there particular bias-related challenges in AI-driven defense systems—such as perception biases in sensor fusion, decision-making biases in autonomous threat assessments, or biases in multi-agent collaborative UAV operations—that you consider are most interested and critical to mitigate? Or any other preferred mitigation applications?
Answer: No preference

10. To better align our approach with the project's objectives, do you prefer enhancing existing neuro-symbolic methods (e.g., probabilistic flow circuits, grammar-constrained decoding, as mentioned in the solicitation) or exploring the development of entirely new techniques? Additionally, are there specific challenges or gaps in current approaches that you would like to see addressed?
Answer: We encourage entirely new techniques and for performers to take risks. The techniques mentioned may be adapted as well, but innovation is key.

11. To ensure the proposed solutions are practical, relevant, and testable within the project scope, could you clarify what datasets or domains are expected to be used for testing bias mitigation techniques? Are these datasets pre-defined, or is there flexibility to propose and introduce new ones?
Answer: No preference

12. What metrics or evaluation criteria will be used to measure the effectiveness of bias mitigation methods? Are there specific benchmarks or performance thresholds to meet?
Answer: No preference

13. To ensure our proposed bias mitigation techniques align with the project's objectives, could you clarify the specific types of AI systems that the project aims to improve? For example, is the focus primarily on generative models (e.g., LLMs, diffusion models), decision-making systems (e.g., AI-assisted policy recommendations, automated hiring), or recommendation engines (e.g., personalized content filtering)? Additionally, are there specific operational scenarios or applications where bias mitigation is most critical? Understanding this will allow us to tailor our approach to the system architectures, constraints, and real-world deployment considerations relevant to the project.
Answer: No preference

14. Your solicitation describes the need for neuro-symbolic methods that can mitigate both explicit bias (defined through symbolic descriptions) and implicit bias (manifested in model parameters/data). Could you clarify if DARPA has specific preferences for how the symbolic and neural components should interact? For example, should they operate sequentially, in parallel, or in a more integrated fashion?
Answer: Preferably, the symbolic and neural would be integrated in innovative ways, but we are open to other methods. Of course, integrated, sequential, and

parallel are not mutually exclusive and can all play a role at some point in the proposed solution.

15. The Zhang et al. reference demonstrates content removal from diffusion models. Should proposals address both the removal of biased content and the verification that such removal hasn't created new biases? Additionally, how important is it to demonstrate that the removal process hasn't degraded model performance on unrelated tasks?
Answer: That direction is within scope but is not required. It is important to at least empirically show the degradation or impact in model performance.

16. Regarding evaluation metrics, would DARPA prefer to see quantitative measurements of both bias reduction and preserved accuracy? If so, are there specific benchmark datasets or evaluation frameworks that should be considered?
Answer: Yes, and we defer to performers which benchmarks they want to use or they may also create their own.

17. The Jha et al. reference discusses impacts of proper nouns on model behavior. Should proposals treat proper noun bias as a distinct technical challenge from general concept bias? Additionally, how important is it to demonstrate generalization across different types of proper nouns (e.g., names from different cultures, brand names, location names)?
Answer: that is up to the performer. Any social bias is within scope.