HR0011ST2025D-03
Unlearning Bias
Frequently Asked Questions (FAQs)

1. What specific types of biases is the government planning to target?
   A: social biases. Statistical biases are in scope as long as it is clear how those biases manifest in social biases.

2. How will the biases to be unlearned be identified? Are there any specific metrics, tools, or datasets that will be used to define and measure bias?
   A: it is up to the performer to define these things

3. Are there any constraints on the size or type of the model (e.g., large-scale models like GPT or smaller domain-specific models like LLaMA-7B)?
   A: no constraints

4. Are we allowed to combine multiple techniques (e.g., modular add-ons and conceptual space analysis) into a hybrid approach?
   A: yes

5. Should the proposed solution be entirely novel, or is the adaptation of existing methods acceptable?
   A: innovative solutions are preferred. If the solution proposed is incremental, a strong justification must be made for the impact of the solution

6. Is there a preference for methods that are generalizable across multiple domains or highly specialized for a single domain?
   A: no preference

7. Are there any specific biases that DARPA is interested in unlearning specifically? Can they provide any examples of biases or types of biases?
   A: Any social bias is within scope. Statistical bias that manifests itself in a social bias is also within scope.

8. Does DARPA consider model overconfidence a form of bias?
   A: Yes, but the performer must make it clear how such overconfidence can manifest itself in a social bias and how to overcome it.

9. What types of models or generative task is DARPA planning to use this tool on? LLMs? VLMs? Image Generation?
   A: All are within scope. Other architectures like KANs, LRMs, etc. are also within scope.

10. Regarding the models that DARPA wishes to unlearn certain biases from, do they generally have access to the model's internals?
    A: Not always, but performers can proceed with either assumption.

11. Can we assume that the end-user is able to determine biased vs non-biased model outputs? Or at least provide a few examples of definitive positive and negative cases?
    A: Yes

12. What is the level of expertise of a proposed end-user?

A: Any level of expertise is fine as long as the idea is innovative and demonstrates potential for impact.

13. What is currently preventing DARPA from doing this already?
    A: N/A

14. The topic description mentions "concept replacing" as a potential approach for unlearning bias. Given the examples in Zhang et al. (2024) focus primarily on text-to-image diffusion models, would DARPA be interested in solutions that demonstrate unlearning across other model architectures beyond diffusion models, provided we can validate the approach using similar evaluation metrics?
    Answer: Yes.

15. The solicitation emphasizes developing methods for "removing bias from a generative model without accessing the original training data." Could you clarify whether techniques that require limited access to model gradients (but not training data) would be considered compliant with this requirement?
    Answer: Yes. Such techniques are within scope.

16. Regarding the evaluation schema mentioned in the deliverables, beyond demonstrating that targeted bias has been removed, are there specific metrics or benchmarks DARPA would like to see for measuring the preservation of model performance on non-bias-related tasks?
    Answer: it is up to the performer which benchmarks they use or they may create their own as well.