HR001118S0047
SYSTEMATIZING CONFIDENCE IN OPEN RESEARCH AND EVIDENCE (SCORE) Program
Frequently Asked Questions (FAQs)
as of 7/18/18

30Q: Will the data from TA1 include actual PDFs of the published articles, previous R&R studies, and preprints, or just bibliographic references to such?

30A: As stated in the SCORE BAA in Section I.E.1, TA1 proposals will recommend the types and formats of data to be include in their Common Task Framework (CTF). TA1 proposals should identify the literature or data repositories to be leveraged and detail a data management plan (see Section I.I.3) with credible mechanisms to curate, store, and release data for TA2 CSs assignment and train/test data for TA3. While proposed data management plans are at the discretion of each TA1 team, DARPA will be considering the practicality of suggested data, such as published articles, previous R&R studies, preprints, and bibliographic references, and will fund proposals which are most likely to achieve SCORE objectives.

29Q: Will TA2 know which 200 claims are being replicated or reproduced by each TA1 performer?

29A: As the focus for TA2 is to assign accurate CSs to all TA1 claims, regardless of whether those claims will be empirically evaluated by TA1 or not, DARPA anticipates that the TA2 teams will not know which specific 200 claims are being replicated or reproduced by each TA1 performer.

28Q: Is it in scope for TA2 teams to request that specific metadata variables (e.g. sample size, statcheck tool output) be extracted by TA1 teams and included as part of the CTF of "raw" data? Or, is it expected that TA2 teams would extract such additional information themselves if they required it?

28A: As discussed in Section I.E.1 of the SCORE BAA, the goal of TA1 is to curate SBS research claims into a single CTF dataset. Hence, while specific metadata variables for that dataset may depend on the unique solutions proposed by selected TA1 team(s), TA2 proposals should not assume that any metadata will be extracted/provided by TA1. If TA2 teams anticipate requiring metadata variables to elicit accurate human expert CSs, then they should identify a clear and credible approach for collecting/aggregating metadata for the CTF dataset (see Section I.E.2).

27Q: Will there be a single metadata standard for the data outputs of TA1 performers? If so, how will this be decided, and will TA2 and TA3 performers have input into this?

27A: As stated in the SCORE BAA in Section I.E.2 and I.E.3, both TA2 and TA3 teams should identity anticipated requirements of/for data sets, structure, formats, and data management in their proposals. There will be a standard TA1 data output which the T&E team will be responsible for finalizing with the input of all the TA teams (see Section I.E.4 for more on the T&E team). In addition, all performers are encouraged to include a Data Management Plan (DMP) in their proposals which would include the necessary and sufficient scope of data that may be applicable to their goals. Note, the DMP does not count against the page-limit for Volume 1.

26Q: Will the claims that TA2 performers are assessing be drawn from the library of a single TA1 performer, or will TA2 performers be drawing claims from all TA1 performers?

26A: Per Section I.E.2 of the SCORE BAA, TA2 performers will draw claims from all the TA1 performers, not just a single TA1 performer.

▲▲▲▲ New Questions and Answers ▲▲▲▲

25Q: What is the timeline by which a decision is made about the white paper/abstract?

25A: DARPA policy is to attempt to reply to abstracts within thirty calendar days. Abstracts submitted for TA1 and TA2 have already been reviewed and feedback has been provided. As stated in Section IV.B.1.a, the official notifications were sent via email to the Technical POC and/or Administrative POC identified on the abstract coversheet.

24Q: Is the SCORE program institutionally limited? For example, is the eligibility of a PI contingent upon the approval of any white paper/abstract they may have submitted?

24A: As described in the SCORE BAA in Section I.D, the same person or organization may be on multiple proposals as long as those proposals are submitted to the same TA. However, proposers should provide sufficient approaches for managing potential conflicts of interest and/or firewalls among different teams/proposals in that TA and provide evidence that they have sufficient resources to mitigate any technical, cost, and/or schedule risk should they or their team members be on multiple proposals selected for negotiation. This is to avoid OCI situations between the Technical Areas and to ensure objective test and evaluation results (see Section III.D). As described in Section IV.B.1.a, DARPA will respond to abstracts with a statement as to whether DARPA is interested in the idea, but regardless of DARPA's response to an abstract, proposers may submit a full proposal.

23Q: Can TA3 submit abstracts before the November 1st due date (and expect to receive feedback before then)?

23A: Please refer to FAQ 18 for an answer to this question.

22Q: Will TA3 performers be able to attend TA1 and TA2 kickoff meetings before the TA3 start date?

22A: As stated in the SCORE BAA in Section I.D, the kickoff for TA1 and TA2 will occur in Month 1 of the program. The kickoff for TA3 will occur in Month 7 of the program. Since TA3 teams will not start work until Month 7, they will not attend the earlier kickoff for the TA1 and TA2 teams. Relevant information from the TA1/TA2 kickoff will be relayed to the TA3 teams selected for award. In addition, and as described in the SCORE BAA in Section I.F, the TA3 kickoff will coincide with the Month 7 PI meeting, which will allow all teams to meet and discuss their approaches.

21Q: Will there be an adjudication process of TA3 evaluation? For instance, if all TA3 algorithms are in high agreement, but all disagree with the reference judgement, will there be a process to review these claims to check for human judgement errors?

21A: As described in the SCORE BAA in Section I.E.3, TA3 teams will be evaluated in terms of their CSs overlap with the best performing TA2 CSs. Per Section I.E.2, TA2 CSs will be evaluated by their ability to predict the outcomes of TA1's empirical evaluations of the degree of reproducibility and replicability of a representative subsample of studies and claims. Therefore, as the accuracy of the best performing TA2 teams will have been determined prior to Phase 2, during which TA3 algorithms will be evaluated in terms of their increasing overlap with TA2 CSs (Section I.E.3), no adjudication of this kind for TA3 algorithms is anticipated. If special circumstances arise where this kind of adjudication process may be deemed critical for program success in Phase 2, DARPA may consider it. However, proposers should not assume such a process will be used for evaluating performance under SCORE.

20Q: Does TA3 have a different proposal due date than TA1 and TA2?

20A: Yes, the technical areas (TAs) have different proposal due dates. As stated in Part I of the SCORE BAA, proposals for TA1 and TA2 are both due on August 1, 2018, at 4:00 p.m. EST. Proposals for TA3 are due December 12, 2018, at 4:00 p.m. EST.

19Q: Will proposals to develop approaches that apply Confidence scores in multiple scientific disciplines, not only social and behavioral sciences, be considered for this program?

19A: Approaches that apply Confidence Scores (CSs) across multiple scientific disciplines will be considered, provided they are initially focused on the Social and Behavioral Sciences (SBS), as stated in the SCORE BAA in Section I.A. The reason for this requirement to include SBS research is addressed in Section I.B, namely that SCORE is intended to help address critical complex national security challenges in the Human Domain, such as enhancing deterrence, supporting stability, increasing trust and influence, reducing extremism, and enhancing "social-behavioral modeling." In the SCORE BAA in Section I.E.1, TA1 team(s) are instructed to identify and justify the SBS topics/literatures to be curated as well as any non-SBS topics/literatures that might be used to assess the generalizability of TA2/TA3 methods. Similarly, in Section I.E.2, TA2 team(s) are instructed to offer suggestions for SBS topics or literatures best suited for approach, including identifying extensions to other topics/literatures, including non-SBS research.

18Q: Can abstracts for TA3 be submitted before the official due of November 1?

18A: Yes, abstracts for TA3 can be submitted before November 1, however feedback on TA3 abstracts will not be provided until after the TA3 abstract deadline has passed. As described in the SCORE BAA in Section I.C abstracts must be received by DARPA no

later than the due date and time listed in *Part One: Overview Information*. Abstracts received after this time and date may not be reviewed.

17Q: Is it possible to speak directly with the DARPA/DSO PM leading this effort during the source selection process?

17A: All questions regarding SCORE should be submitted to the SCORE@darpa.mil inbox and will be posted as FAQs.

16Q: What is the delineation of duties between TA1 and TA3? Specifically, will TA1 be responsible for feature generation (e.g., inferring features like "study design" or "number of subjects" from research articles, and curating those for TA3 algorithm developers)? Is it in scope for TA3 performers to extract or create additional features (e.g., number of peer-reviewed articles published per author per study)?

16A: As discussed in the SCORE BAA in Section I.E.1, the TA1 team(s) are responsible for provisioning data for the program by curating SBS research claims into a single common task framework (CTF) dataset for TA2 and TA3 teams. The result of the TA1 effort will be a CTF of “raw” data, including journal articles, previous R&R studies, abstracts, and preprints. The TA1 team(s) therefore will not be responsible for feature generation. In Section I.E.3, the SCORE BAA specifies that the TA3 team(s) will be responsible for all aspects of algorithm development for automatically assigning CSs to the TA1 curated data. This is expected to include feature generation from the curated data, as well as identifying, extracting, and/or creating additional endogenous/exogenous features to inform their algorithms (see *Table 1: Technical Goals by Phase* in the SCORE BAA, Section I.F, page 16-17).

15Q: Can an offeror submit a full proposal to TA1 or TA2, and if not selected, submit to TA3?

15A: Provided that offerors who submitted to TA1 or TA2 have received notice that they have not been selected for award, they are permitted to submit a full proposal to TA3. However, given the different requirements for each TA, proposers are encouraged to focus on the TA that they feel best suited to address. As a reminder, per the SCORE BAA in Section 1.D, each proposal should only address a single TA and no person or organization may be a performer on more than one TA, either as a prime or sub-contractor. However, the same person or organization may be on multiple proposals if those proposals are submitted to the same TA.

14Q: How will the validity of the expert assessments be established in the SCORE program?

14A: As described in the SCORE BAA in Section I.E.1, TA1 will be responsible for empirically validating the reproducibility and replicability (R&R) of a representative subsample of claims in order to evaluate the accuracy of TA2 CSs expert methods. To achieve this, TA1 will reproduce, replicate, or jointly reproduce and replicate a

representative sample of research claims for comparison to the TA2 expert predictions of R&R. This process will determine the validity of the TA2 expert assessments methods, with a target of at least 80% accuracy in Phase 1 (for more, see FAQ 9 below). As stated in the SCORE BAA, Section I.E.2, TA2 team(s) must identify one or more approaches for acquiring expert assessments, such as the use of prediction markets, expert surveys, online games, and/or other innovative and technically compelling approach(es). Accordingly, TA2 proposers are encouraged to consider the merits of different possible methods in order to select and justify their approach(es).

13Q: The BAA refers to "empirical evaluations" as a responsibility of the TA1 teams. How will these evaluations be carried out?

13A: As described in the SCORE BAA in Section I.E.1, TA1 team(s) are required to propose a principled plan for the empirical evaluation of a representative sample of SBS research claims from their curated datasets. This plan should include the anticipated number of replications, reproductions, and joint replications/reproductions. Empirical evaluations will involve the TA1 teams reproducing and/or experimentally replicating the representative sample of claims in order to evaluate whether TA2 CSs accurately predict the degree to which each claim is reproducible and/or replicable. As each TA1 team may have a different process and approach for achieving the required number of empirical evaluations, the exact criteria of what constitutes an empirical evaluation will depend upon specific solutions proposed by TA1 teams selected for award, and will be finalized by the SCORE T&E team in conjunction with TA1 teams and DARPA during the first month of Phase 1 (see *Table 2: Program Events by Month* in the SCORE BAA, Section I.F, page 17).

12Q: The BAA refers to "held-out TA1 test sets" on page 13. What do these tests refer to and what will they be used for?

12A: As described in the SCORE BAA in Section I.E.1, TA1 team(s) will be responsible for collecting a dataset of social and behavioral science (SBS) research articles to which the TA2 teams will assign confidence scores (CSs). In Phase 1 and Phase 2, TA3 teams will receive a subset of the dataset with corresponding CSs, which they may use in training their algorithms (i.e., "training data"). Another subset of the dataset will be "held-out" (i.e., "testing data") and presented to the TA3 teams without CSs. TA3 accuracy will be evaluated only using the held-out sample of data.

11Q: Will the TA2 Expert methods and the TA3 Algorithm methods be required to produce two confidence score (CS) sub-scores, one for the reproducibility of a study and the other for the same study's replicability?

11A: Per the BAA Section I.D., the exact definition and approach for CSs for the SCORE Program will be determined in Month 1 of the program. Further, the Technical Area descriptions in the BAA (see Section I.E.1 and I.E.2) invite proposers to propose

potential criteria and CS scoring approaches, which may include two (or more) sub-scores. Finally, per Section I.E.3, the SCORE Program seeks to develop algorithms that may be tailored to different users, some of whom may prioritize reproducibility or replicability according to their needs. Hence, DARPA anticipates that TA2 and TA3 teams should consider methods that allow for their CSs to be decomposable into different reproducibility and replicability sub-scores based upon a variety of different factors and signals.

10Q: How is a "research claim" defined in the scope of the SCORE program?

10A: For the purposes of the SCORE BAA, a "research claim" refers to the key original result of a research article; however please note that, per the BAA Section I.D., the exact definition of "research claim" for the Program will be finalized by the SCORE T&E team in conjunction with TA1 and TA2 and DARPA during the first month of Phase 1 (see *Table 2: Program Events by Month* in the SCORE BAA, Section I.F, page 17). Further, please note that Sections I.E.1 and I.E.2 of the BAA invite proposers to nominate specific definitions of what they believe a "claim" should constitute in order to best achieve the SCORE Program vision.

9Q: How is "accuracy" defined in the scope of the SCORE program?

9A: Per the SCORE BAA, Section I.E.2, "accuracy" refers to the degree to which the TA2 expert CSs are able to correctly predict the degree of reproducibility and replicability of the subset of research claims that are empirically evaluated by the TA1 teams. That is, how accurate is a TA2 team's method(s) in assigning higher or lower CSs in regards to a research claim's R&R when compared to TA1's empirical evaluation of that same claim? Does a TA2 team's method(s) accurately assign a low (high) CS when a claim turns out to be less (more) reproducible and/or replicable? The exact accuracy measures to be used will be finalized by the SCORE T&E team in conjunction with the TAs and DARPA during the first month of Phase 1 (see *Table 2: Program Events by Month* in the SCORE BAA, Section I.F, page 17). Similarly, while proposers are free to offer suggestions for defining and quantifying the "degree of reproducibility and replicability," these will also be finalized by the SCORE T&E team in conjunction with the TAs and DARPA during the first month of Phase 1.

8Q: How is "overlap" defined in the scope of the SCORE program?

8A: Per the SCORE BAA, Section I.E.3, "overlap" refers to the manner in which TA3 algorithms will be evaluated in terms of their degree of similarity to the best performing TA2 CSs, ultimately seeking to achieve 95% confidence in an algorithm's overlap with those TA2 CSs. The exact definition and metrics to measure the degree of overlap will be

finalized by the SCORE T&E team in conjunction with the TAs and DARPA during the first month of Phase 1 (see *Table 2: Program Events by Month* in the SCORE BAA, Section I.F, page 17).

7Q: How is "confidence" defined in the scope of the SCORE program?

7A: For the purposes of the SCORE BAA, "confidence" is the probability that a given research claim will be reproduced and/or replicated. The exact definition of confidence will depend in part on the SCORE Program's performers and their proposed solutions, and – per the BAA - will be finalized by the SCORE T&E team in conjunction with the TAs and DARPA during the first month of Phase 1 (see *Table 2: Program Events by Month* in the SCORE BAA, Section I.F, page 17). In the BAA, the terms "predictions," "probabilities," and "forecasts" are used interchangeably when talking about assigning "Confidence Scores" to SBS research claims.

6Q: How are "experts" defined in the scope of the SCORE program?

6A: For the purposes of the SCORE BAA, the term "experts" is meant to refer to individuals and/or collections of individuals who are able to most accurately and reliably assign CSs to the TA1 datasets. While a common definition of an expert often depends on subject matter expertise and/or experience in a given field of social and behavioral (SBS) science research, the SCORE Program remains agnostic about this particular definition. Consequently, the exact definition of "experts" is at the discretion of TA2 proposers and their specific methods to best achieve the SCORE Program vision.

5Q: Where can the Proposers Day slides be found?

5A: The Proposers Day slides can be found on the DARPA/DSO Opportunities page. Please see link provided http://www.darpa.mil/work-with-us/opportunities

4Q: What is the scope/definition of social and behavioral sciences (SBS) for the purpose of this BAA?

4A: SCORE is interested in a diverse range of SBS research in order to ascertain the efficacy of this program across multiple fields. As such, the specific area of SBS research to be assessed is left to the discretion of each proposer and their solution(s) designed to best position DARPA to achieve SCORE Program goals. Per the BAA Section I.E, proposers are invited to identify and justify their proposed SBS literatures, areas, and disciplines. Please note that the SBS research areas for SCORE will be selected in conjunction with the T&E team and DARPA during the first month of Phase 1 (see SCORE BAA, Section I.E.1, page 9).

3Q: What level of automation should there be in the algorithms? That is, how much human interaction/intervention would be tolerated or recommended?

3A: The specific degree of automation is at the discretion of the proposer and their specific solution they believe will best achieve SCORE Program outcomes and vision. However, proposers may wish to detail the extent to which their algorithms may be fully automated or may require some degree of human intervention, and consider justifying this approach as the strongest option for helping DARPA realize SCORE goals.

2Q: Is it mandatory to submit teaming profiles?

2A: No, it is not mandatory. The teaming profiles are intended to help offerors who would like to team with other persons/organizations. Collected teaming profiles will be shared to allow offerors to self-select teams. Please note that DARPA will not select teams, and specific content, communications, networking, and team formation are the sole responsibility of the participants.

Neither DARPA nor the DoD endorses the information and organizations contained in the consolidated teaming profile document, nor does DARPA or the DoD exercise any responsibility for improper dissemination of the teaming profiles.

1Q: Are offerors allowed to submit abstracts against multiple Technical Areas (TAs)?

1A: Yes, DARPA will accept and provide feedback on abstracts against multiple TAs from the same person/organization. However, for clarity, please note that – per the SCORE BAA, Section III.D – proposers can only be on multiple proposals if they are submitted against the same TA.