Request for Information (RFI)
DARPA-SN-17-57
Confidence Levels for the Social and Behavioral Sciences

Responses Accepted: Until 4:00 PM (Eastern) on August 24, 2017
Point of Contact: Adam Russell, Program Manager, DARPA/DSO
Email Address: SBSCL_RFI@darpa.mil

"Confidence Levels" for the Social and Behavioral Sciences

The Defense Advanced Research Projects Agency (DARPA) Defense Sciences Office (DSO) is requesting information on new ideas and approaches for creating (semi)automated capabilities to assign "Confidence Levels" to specific studies, claims, hypotheses, conclusions, models, and/or theories found in social and behavioral science research. These social and behavioral science Confidence Levels should rapidly enable a non-expert to understand and quantify the confidence they can have in a specific research result or claim's reliability, reproducibility, and robustness.

Background

Given the accelerating social and technical complexity of today's world—a world that is increasingly connected but still poorly understood—there is a growing interest in leveraging the social and behavioral sciences (SBS) for data, insights, theories, and research results that can help address critical complex national security questions.[i] For example, the National Academies is currently conducting a "Decadal Survey of Social and Behavioral Sciences for Applications to National Security," identifying a number of questions and challenges where SBS research could make significant contributions[ii]. Likewise, the Minerva Research Initiative has identified a series of topics and research priorities that explicitly seek to support SBS research and transition results for DoD use.[iii] Indeed, there are even more examples where SBS are specifically mentioned as potentially contributing to solutions for a wide range of national security challenges, including deterrence,[iv] stability,[v] trust and influence,[vi] and extremism.[vii]

At the same time that this demand signal has been increasing, however, there have been growing concerns about the credibility of many SBS results, with a number of recent studies and analyses revealing that different studies and conclusions may vary substantively in terms of their reliability, robustness, and reproducibility.[viii] Endeavors to alleviate these concerns include new calls and early efforts for advancing (in particular) reproducibility of SBS research, in part by identifying, promoting, and recognizing best research practices among different communities.[ix] Yet, while increasing reproducibility of SBS research may be a necessary step towards calibrating user confidence in SBS results, it is also unlikely to be sufficient in and of itself. For example, a study, result, or claim may be—strictly speaking—computationally reproducible[x] but still be best characterized as having a low Confidence Level. Low confidence may mean the claim or result is the product of early exploratory research, making the finding tentative or speculative as compared to a confirmatory study.[xi] Similarly, a study may claim diagnostic or predictive accuracy or generalizability that seem unlikely given known practical or theoretical limitations,[xii] and/or the study seems to involve Questionable Research Practices (QRPs).[xiii]

Current approaches to establishing Confidence Levels in SBS claims or results are often slow, as peer-review may take significant amounts of time and may often suffer its own limitations.[xiv,xv] Likewise, publications often lack mechanisms for rapidly incorporating dynamic changes in the literature, such as retractions, replications, or new findings, which presumably should alter a user's confidence in a specific study's conclusions. Finally, while the use of expert prediction markets to collectively evaluate specific SBS claims has recently shown promise[xvi] and highlights the continuing value of expertise in assessment, practical limitations remain and there are questions about the scalability and speed of such an approach more generally.

Taken in the context of growing numbers of journals, articles, and preprints, the current state of affairs results in an inability for most consumers of SBS research to evaluate the confidence he or she should assign to a particular SBS study or claim for their purposes. In some cases, this purpose might be deciding whether, and to what extent, to incorporate a cognitive bias as a reliable, robust, and replicable phenomenon into a model of group decision-making. In others, it might mean deciding whether and how to weight the potential influence of hurricane names on potential population behaviors, or effects of certain priming interventions on public health. In each case, the user needs to evaluate the claim's reproducibility, reliability, and robustness in order to calibrate the appropriate Confidence Level they might have for relying on it and—perhaps more importantly—the extent to which they should weight or privilege that claim within their own forecasts, models, analyses, or decisions.

Goals

DARPA hypothesizes that there may be new ways to create automated or semi-automated capabilities to rapidly, accurately, and dynamically assign Confidence Levels to specific SBS results or claims, in order to help SBS users better calibrate the confidence they should have in those results or claims. Such a capability should help non-experts identify and triage "interesting" but "novel" claims with a low Confidence Level (those which may or may not prove to be true or robust) from "proven out" and "reliable" claims with a high Confidence Level (those which may be readily adopted or used because their reliability, robustness, and limitations are well established and recognized).

Accordingly, DARPA is seeking responses to this RFI to help assess the State of the Art (SOA) in current capabilities that can speak to the challenges of evaluating the maturity and credibility of SBS claims, and to request new ideas, approaches, tools, or methods that could enable (semi)automated capabilities for assigning reliable SBS Confidence Levels (SBSCLs) to SBS research. Responses may address one, some, or all of the following questions:

1. **State of the Art**
   a. Beyond standards as used in a number of domains and industries[xvii], and relevant efforts like prediction markets[xviii], various reproducibility indices[xix], meta-analyses, or bibliometric, citation, and impact factors,[xx] what are current capabilities for assigning Confidence Levels to SBS results? What are their limitations and

challenges?

b. What are potentially relevant technologies for (semi)automating processes that may use—and fuse—highly variable data sources to assign SBS Confidence Levels? What are their current limitations and challenges? Examples might include machine reading, natural language processing, automated meta-analyses, statistics-checking algorithms, sentiment analytics, crowdsourcing tools, data sharing and archiving platforms, network analytics, etc.

c. What capabilities, if any, exist for qualitatively and/or quantitatively evaluating the confidence one should have in a SBS claim within the wider context of the research literature? What approaches or tools are used when evaluating the confidence one should have in a claim, study, or hypothesis, when that may be bound up with other studies or claims with different Confidence Levels?

2. **SBS Confidence Levels: taxonomies, definitions, applications**

a. What might a semi- or fully-automated system for assigning Confidence Levels to SBS claims look like? How would it function?

b. What might be a proposed SBSCL taxonomy? How would Confidence Levels be defined in this taxonomy? What would a low or high SBSCL reflect, and what might a SBSCL spectrum look like? What factors and weightings might be most important for SBSCLs? What might be candidate quantitative SBSCL scales, as well as qualitative SBSCL scales?

c. How might SBSCLs include "internal" evidence, such as data inherent to the study or the claim itself (e.g., missing outcome measures or insufficient power of a study, or causal claims based on experimental vs. observational research, etc.)? How might SBSCLs include "external" evidence about the wider context of that research (e.g., evidence of a claim being replicated in new populations, a study being demonstrably reproducible, a theory being tested across conditions to evaluate generalizability, amount of peer criticism or controversy, etc.)?

d. What is the correct focus of application for SBSCLs? Should a SBSCL be assigned to an individual claim (e.g., variable X affects Y in conditions Z), in the assumption that claim-level Confidence Levels could be aggregated to a SBSCL at a more general level? An effect size level? Or to an overall study (e.g., Russell 2017)? A topic (e.g., priming)? An author? An organization? A theory? A discipline? How might a SBSCL reflect some or all of these different applications at once?

e. What might be the most efficient, intuitive and effective methods for conveying SBSCLs to users? Could a qualitative SBSCL be mapped onto a quantitative mathematical Confidence Level, and if so, how might one validate the robustness of this mapping? What visualization tools or approaches might best communicate SBSCLs to different SBS users and communities? How might SBSCLs be best integrated—visually or otherwise—into different end-products (studies, news reports, analyses, meta-analyses, decisions, etc.)?

f. How could SBSCLs be defined in ways that align with potentially different users with different needs and applications? For example, an analyst may simply want some general background on different social systems, where a claim asserting differences among individualist/collectivist cultures may have one Confidence Level, while the same claim would have a different Confidence Level if that analyst sought to use it as the basis for a cognitive model of individuals in an agent-based simulation for policy decisions.

3. **Data sources and signals**
   a. What are potential data sources for SBSCLs? Are there techniques that could automatically identify the most discriminating signals for SBSCLs, potentially eliminating a source of bias, or uncovering undervalued information? What previously unconventional data sources might be newly leveraged and aggregated for SBSCLs? Examples might include public criticism,[xxi] popular press coverage, sentiment analysis and social media/blog posts, crowd-sourcing,[xxii] automated evidence of QRPs, evidence of publication bias or file drawer problems in a topic or field, relative rate of accepted posters or talks at conferences, social networks of reviewers or co-authors, retraction rates, structure of incentives, journal transparency[xxiii], ratio of replications to novel findings, pre-registration rates for the authors, topics, disciplines, etc.
   b. How might additional signals or new data and sources be used to update SBSCLs once determined? What weight should be given to new evidence?
   c. What might be the impact of firewalls and lack of open access literature on SBSCLs?

4. **Validation**
   a. How might SBSCLs be validated? How should validation be scored? What level of accuracy should be considered sufficient validation?
   b. Could prediction markets be used to validate SBSCL algorithms, and if so, what might a validation plan look like? What level and amount of SME input would be required? Are there potential or upcoming SBS replication efforts that might serve as validation opportunities for SBSCL algorithms or capabilities?
   c. How could SBSCLs avoid being gamed as a "sum" score (e.g., a high score in several categories could mask some critical failure in another)?
   d. How might SBSCLs address challenges that face other industry standards, such as the criticism that they "blur" several aspects of technology and product readiness into a single number? How can we highlight the relative contributions of the various weighted factors of readiness throughout the lifetime of a claim, model, or theory?

Submission Format

Responses may address one or more of the questions outlined in this RFI. DARPA encourages responses that describe integrated solutions that address some or all of the questions, but responses to specific questions are also acceptable. Respondents are encouraged to be as succinct as possible, while also providing actionable insight. Page limits for each section are indicated below.

Format specifications for responses include 12-point font, single-spaced, single-sided, 8.5 by 11 inches paper, with 1-inch margins in .doc, .docx, or PDF format (and, as applicable, .ppt or .pptx). Respondents are responsible for clearly identifying proprietary information. Responses containing proprietary information must have each page containing such information clearly marked with a label such as "Proprietary" or "Company Proprietary." DO NOT INCLUDE ANY CLASSIFIED INFORMATION IN THE RFI RESPONSE.

A. Cover Sheet (1 page): Provide the following information.
   1. Response Title
   2. Technical point of contact name, organization, telephone number, and email address
   3. Indicate the RFI question(s) addressed by the response

B. Technical Description (5 pages)

C. Bibliography/References (1 page)

D. Graphic Overview Slide (Optional): If desired, include a single PowerPoint slide that graphically depicts the main ideas of the response.

Submission Instructions

All responses to this RFI must be emailed to SBSCL_RFI@darpa.mil. Responses will be accepted any time from the publication of this RFI until 4:00 PM (Eastern) on August 24, 2017. Early responses are encouraged.

Contact Information

All technical and administrative correspondence regarding this RFI should be emailed to SBSCL_RFI@darpa.mil. Emails sent directly to the Program Manager may result in delayed/no response.

Disclaimers and Important Notes

This is an RFI issued solely for information and new program planning purposes; it does not constitute a formal solicitation for proposals. In accordance with FAR 15.201(e), responses to this RFI are not offers and cannot be accepted by the Government as such. In addition, responses do not bind DARPA to any further actions related to this topic including requesting follow-on proposals from respondents to this RFI. Submission is voluntary and is not required to propose to a subsequent Broad Agency Announcement (BAA) (if any) or other research solicitation (if any) on this topic. DARPA will not provide reimbursement for costs incurred in responding to this RFI.

Respondents are advised that DARPA is under no obligation to acknowledge receipt of the information received or provide feedback to respondents with respect to any information submitted under this RFI.

DARPA will disclose submission contents only for the purpose of review. Submissions may be reviewed by the Government (DARPA and partners); Federally Funded Research and

Development Centers (FFRDCs); and Scientific, Engineering and Technical Assistance (SETA) support contractors.

[i] E.g., http://nsiteam.com/operational-relevance-of-behavioral-social-science-to-dod/

[ii] See http://sites.nationalacademies.org/DBASSE/BBCSS/SBS_for_National_Security-Decadal_Survey/index.htm

[iii] See http://minerva.defense.gov/Research/Research-Priorities/

[iv] E.g., https://www.nap.edu/catalog/18622/us-air-force-strategic-deterrence-analytic-capabilities-an-assessment-of

[v] E.g., http://nsiteam.com/stability-model-stam-assessments/

[vi] E.g., https://community.apan.org/wg/afosr/w/researchareas/7676/trust-and-influence/

[vii] E.g, http://nsiteam.com/violent-extremism-radicalization/

[viii] E.g., http://www.nature.com/news/reproducibility-1.17552

[ix] E.g., http://www.nature.com/news/reproducibility-1.17552

[x] Noting that there is still little consensus on formal distinctions among different kinds of reproducibility and replicability (e.g., https://www.nap.edu/catalog/21915/statistical-challenges-in-assessing-and-fostering-the-reproducibility-of-scientific-results), we use the term reproducibility here in its most minimal sense of computational reproducibility: can another researcher reproduce a study's results if given access to the data, meta-data, and analytic method and tools, including the code?

[xi] E.g., https://en.wikipedia.org/wiki/Research_design#Confirmatory_versus_exploratory_research

[xii] E.g., http://cs.stanford.edu/people/ashton/pubs/limpred.pdf

[xiii] https://www.timeshighereducation.com/blog/grey-zone-how-questionable-research-practices-are-blurring-boundary-between-science-and

[xiv] E.g., https://arxiv.org/ftp/arxiv/papers/1205/1205.1055.pdf

[xv] http://retractionwatch.com/2017/04/20/new-record-major-publisher-retracting-100-studies-cancer-journal-fake-peer-reviews/

[xvi] E.g., http://www.nature.com/news/the-power-of-prediction-markets-1.20820

[xvii] E.g., https://standards.ieee.org/

[xviii] E.g., http://mason.gmu.edu/~rhanson/SUM13.pdf and https://osf.io/yjmht/

[xix] E.g., https://replicationindex.wordpress.com/

[xx] E.g., http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002541

[xxi] E.g., http://www.nature.com/news/publicly-questioned-papers-more-likely-to-be-retracted-1.14979

[xxii] E.g. http://www.sciencemag.org/news/2017/06/great-paper-swipe-right-new-tinder-preprints-app

[xxiii] E.g., http://www.the-scientist.com/?articles.view/articleNo/32427/title/Bring-On-the-Transparency-Index/