## 2. Goals and Impact

Social and Behavioral Science (SBS) has a crucial impact on economic, political, social and health-related policy. The sheer scientific difficulty of SBS research, coupled with questionable research practices ("HARKing", "p-hacking", positive publication bias, and underpowered studies) resulted in a flood of findings that cannot be reproduced and replicated (R&R), undermining confidence in SBS research [1,2,3,4,5]. Even when working with the same data in good faith, different studies may reach different conclusions because they use different methods or account for different variables [6].

Scoring R&R of research claims represents a bottleneck in research progress and practical application of research. Our goal is to automate the evaluation of SBS research and assist domain experts in assessing the credibility of studies. We propose to develop an informatics framework that provides a global, holistic view of research claims that spans specific research practices and the broader research context. The core of our framework will be an SBS knowledge graph that curates both *micro-level* features of claims, such as subject populations and statistical techniques, and connects these features at a *macro level* across studies using features from social dynamics and bibliometric data. Given a set of claims relevant to a predefined area of research, we will score claims to encode confidence in their R&R with full provenance pertaining to how scores were calculated.

**What is the problem? Why is it hard?** The core research problem is scoring scientific claims to predict their R&R. The two central challenges to computationally solving this problem are *understanding research choices* in individual articles and *synthesizing knowledge* by framing these research choices using connections to other studies in a field. The first challenge requires acquiring the requisite domain knowledge and applying this knowledge to identify the disparate features, such as datasets, analytical techniques, and experimental design, necessary to diagnose possible methodological issues. Addressing the second challenge requires compiling a global bird's eye view of research studies and identifying interconnections to assess latent social factors that impact claims, including publication bias, citation patterns, and critical responses to studies. As a machine learning problem, this work is likely to require high-quality validation of R&R (from TA1) and expert annotations (from TA2) and it is likely that there will be domain-specific semantic differences between claims from different SBS disciplines.

**How is it done today, and what are the limits of current practice?** Evaluation of SBS claims currently relies on expert scientists who undergo years of training to acquire the necessary skills. This approach is highly labor-intensive, prone to social and cognitive human biases, lacks transparency, is bespoke, rarely updated, and creates static results. Movements to create more transparent scientific evaluation, such as pre-registration of studies and open review frameworks are gaining traction at a slow but steady pace. Explicit studies of R&R include replication studies for specific claims, as well as meta-analyses and meta-research [7], but pervading research incentives may not reward such efforts, and the narrow focus of such efforts may perpetuate the research flaws of particular techniques. Automated methods that use text mining to build systematic reviews in SBS have been proposed [8, 9], but there exist no easily accessible tools to capture and formalize the claims that are only imprecisely described in the narrative of an article.

**What's new in your approach and why do you think it will be successful?** Our approach centers on building a *knowledge graph* (KG) [10, 11] of scientific claims that integrates information across multiple levels of granularity. Our system will capture fine-grained features about particular articles such as the datasets and subject populations, experimental design parameters, analytical tools, and statistical power of results using *weakly-supervised information extraction techniques* [12, 13] that can quickly be updated to adapt to new disciplines. However,

these article-level features are often insufficient to diagnose whether these choices will result in reproducible and replicable research. To provide a meaningful confidence score, we will also gather features representing social factors at play in SBS disciplines, e.g., using co-authorship and citation networks and social media posts, allowing our system to diagnose publication bias, venue quality, and relationships to existing research (whether exploratory or confirmatory). These features will, moreover, allow for codifying latent social factors affecting research, such as mentor-mentee relationships, author centrality, ranking, etc. [14]. This scalable, KG-based approach will fuel *network and graph analysis to score claims* by combining claim features with novel, unconventional indicators of R&R using the knowledge and relationships curated by the KG. These analytical techniques are readily explainable in the context of specific R&R features and relationships to the research discipline, allowing end users to quickly diagnose potential issues and validate predictions [15].

**What difference will it make? What impact will success have?** Uncertainty about the validity of our knowledge of social phenomena—from how individuals make decisions to how groups behave during a crisis—hinders our ability to effectively anticipate and manage social change. If successful, our work will provide researchers with a framework to analyze SBS findings. Our SBS knowledge graph analytics will support explainable scores that can specifically identify patterns ranging from incorrect statistical analyses for a particular experimental design to contradictory claims found in highly selective publication venues. The framework will allow policy makers to make policy decisions based on solid evidence from SBS research, help advance our understanding of processes in complex social systems and quickly assess how new claims fit within the broader network of SBS research.

**What are the risks and payoffs? How will they be measured?** High fidelity CS predictions present a challenge given the small amount of training data provided by TA1. We anticipate that many of the threats to R&R are nuanced issues at the confluence of several choices. Systems must excel at extraction across several feature classes, incorporate strong feature engineering, and use data-efficient learning techniques. We will work with domain experts to ensure that our explainable scores correctly capture nuanced perspectives on R&R.

**How much will it cost? How long will it take?** The project will last 27 months and will cost approximately $800K per year.

**What are the midterm and final "exams" to check for success? How will progress be measured?** We will measure progress by introducing metrics for each component of our system. Feature extraction will be evaluated for precision and recall relative to a manually annotated benchmark set. Bibliometric and social media features will be evaluated on the basis of coverage of claims and researchers in the TA1 studies. The knowledge graph will be evaluated based on the completeness of entities, attributes, and relationships with specific evaluations for entity resolution and link prediction. Gaming will be evaluated by model sensitivity to features that are easily manipulated, such as unconventional choices of statistical methodology and more conventional features, such as availability of experimental data. Explainability will be validated by experts in SBS research. Ultimately, the system will be evaluated on end-to-end performance based on the overlap with the TA1 R&R scores provided in training data.