Improving Trustworthiness of Computational Results: Opportunities for the NSF Office of Advanced Cyberinfrastructure to address recommendations from the National Academies Report on Reproducibility

The Working Group on Reproducibility and Sustainability[1, 2]

Final Version DRAFT, April 28, 2022

1. Introduction

Reproducible[3] computational results require planning and activities throughout the scientific process, above and beyond directly producing the results for publication. Data, computational environments, and computational steps must be clearly described, reviewed and made accessible in the future, in order to make reproducibility possible (Figure 1).

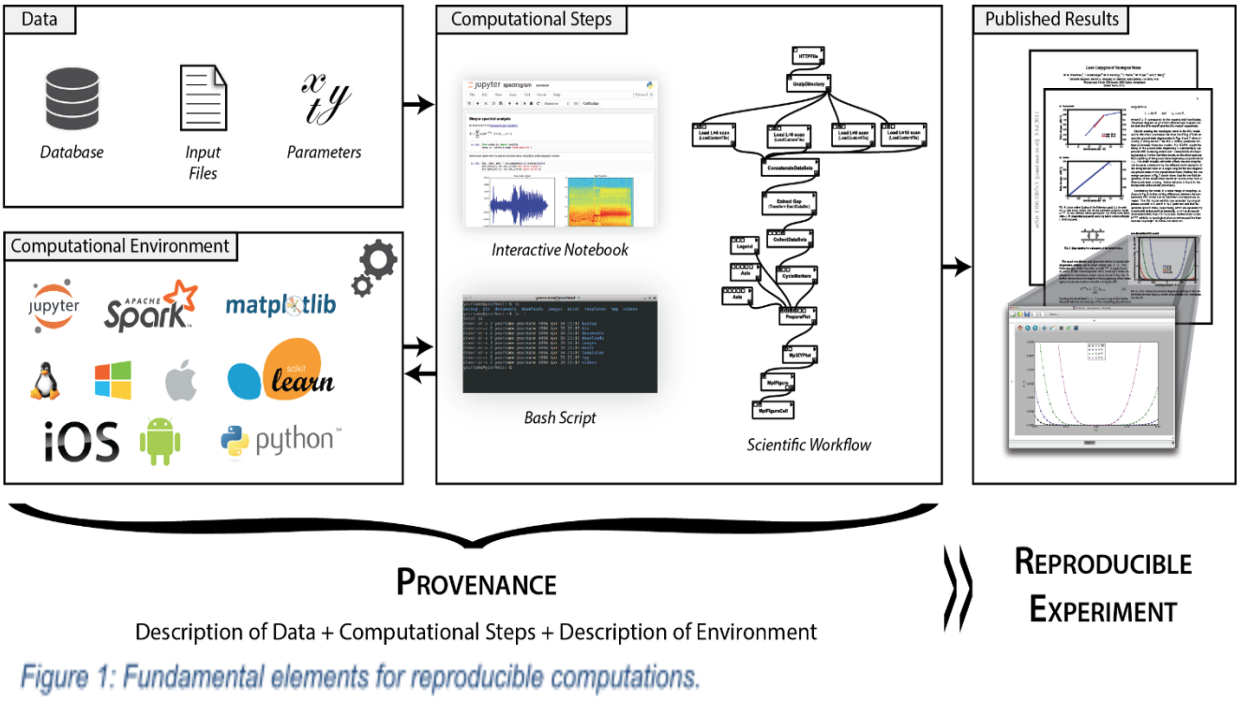

Figure 1: Fundamental elements for reproducible computations.

[1] Working Group Members: Wolfgang Bangerth, Juliana Freire, Patrick Heimbach, Michael Heroux (chair), Ivo Jimenez, Ellen Rathje, Hakizumwami (Birali) Runesha, Victoria Stodden

[2] The Working Group on Reproducibility and Sustainability for the Advisory Committee on Cyber Infrastructure (ACCI) for the Office of Advanced Cyberinfrastructure (OAC) for the US National Science Foundation provides this document of opportunities for how NSF, and OAC in particular, might address the recommendations found in the National Academies report on Reproducibility and Replicability in Science [1]. (See appendix for a list of relevant National Academies recommendations).

[3] The definitions of reproducibility and replicability are found in the Glossary at the end of this report.

The scientific community has been and continues to be concerned about the integrity and trustworthiness of scientific results, computational results being a particular subset. In a pair of studies by *Nature* [2], more than 80% of surveyed members stated that the current status of reproducibility was a crisis in their fields. While these studies were conducted in the mid-2010s, there is little evidence that the situation has fundamentally changed.

In addition, most of the eleven recommendations from a previous report from the National Academies, entitled *Fostering Integrity in Research* [3] also intersect with reproducibility concerns, including the role of funding agencies, the importance of training, and increased rigor in the review of scientific artifacts.

The Office of Advanced Cyberinfrastructure (OAC) can play important roles in advancing the incentives and support for improved reproducibility of NSF-funded research. This document focuses on four types of opportunities:

1. **Research** – Specific research topics to consider for funding.
2. **Infrastructure** – Capabilities and services for NSF-supported scientists.
3. **Programmatic** – New elements of OAC activities.
4. **NSF-wide with OAC scope** - Broad NSF topics with strong OAC connections.

For the opportunities mentioned below, we will often draw upon the recommendations from the National Academies report [1] mentioned above. The number in parentheses indicate the related recommendations from the National Academies report (see Appendix).

1.1 Vision for Trustworthy Computational Science

Reproducible computational results are necessary for establishing trust in science. To qualitatively advance trustworthiness, scientists must incorporate reproducibility into the planning and activities from beginning to end of their scientific workflows, without gaps. Furthermore, as scientists build upon the work of others, and make their work available to others, provenance and replay must remain feasible.

We look toward a future for computational science where all computational results are reproducible, including those from pipelines across multiple teams. Effective and efficient reproducibility will enable qualitative advances in science and make possible a new level of demonstrable trust in scientific results and outcomes.

1.2 Why Focus on Reproducibility

Reproducibility is not the only important goal toward building trust in computational results. Replicability—the ability to obtain consistent results using different studies and data—is essential as well. Other goals, including interpretability, explainability, and transparency, are broadly valuable.

Even so, reproducibility is the foundation for all trustworthiness. Without the ability to reproduce results, pursuing the goals of interpretability, explainability and replicability are not meaningful. In addition, even though transparency is a lesser requirement than reproducibility, in our experience the best means to assuring transparency is through making sure results are independently reproducible, demonstrating the transparency of how the original results were produced.

1.3 Goals for this report

The Working Group on Reproducibility and Sustainability is affiliated with the ACCI. The emphasis of much of the content in this report focuses on the unique role of OAC, which interacts directly with the ACCI. At the same time, the opportunities we present here are relevant for computational experiments conducted under NSF funding in a much broader scope. We intend that the opportunities discussed in this report will be seen as compelling to NSF as a whole, especially as the role of computational results becomes even more important to science, and the importance of reproducibility in elevating trustworthiness continues to grow.

The overarching goal of this report is to provide a description of high-value opportunities for OAC specifically, and NSF generally, to foster and effect culture change in the scientific community, moving toward a future where the tools and processes that support reproducibility are fully integrated into our computational environments and the community culture is transformed to support reproducibility as essential.

1.4 Summary of Opportunities

The opportunities described in this report are summarized here for convenience.

Research	2.1 Support research in reproducibility essentials
	2.2 Support improved provenance capture and replay
	2.3 Support advanced reproducibility testing
	2.4 Support holistic approaches to advancing trustworthiness
Infrastructure	3.1 Enable standardized research delivery
	3.2 Promote community software stacks
	3.3 Establish a Research Software Engineer career track
	3.4 Establish a digital asset management plan
Programmatic	4.1 Establish reproducibility training and certification
	4.2 Elevate reproducibility priorities in project funding and review
	4.3 Support specific funding for reproducibility
	4.4 Start a working group on reproducibility policies, tools, practices
NSF-wide	5.1 Establish a reproducibility initiative

Figure 2: Summary of Opportunities for addressing recommendations from the National Academies Report on Reproducibility and Replicability. Detailed discussions in in subsequent sections.

1.5 First Steps

Reproducibility training and certification: Among the opportunities summarized in the above figure, we believe efforts to stand up a basic training and certification process for NSF-funded computational science teams are most promising as a starting point. While advanced reproducibility concepts, tools, and workflows have not evolved to be packaged in a training and certification program, fundamental ideas such as definitions of transparency, reproducibility, and replicability, and the relationships between these activities are well known and similar across many computational domains. Developing and deploying training toward certification can be accomplished in the foreseeable future, and can provide the first step toward a deeper appreciation of the role and importance reproducibility plays in trustworthy science.

Workshop on roles and responsibilities: The second step we see as timely and as a pre-requisite to progress elsewhere is planning and organizing a workshop on the roles and responsibilities for trusted computational results. Such a workshop could bring together key stakeholders in trustworthy science and facilitate a deeper understanding of how each member, institution, and organization can play a role to elevate the trustworthiness of results. We believe that such a workshop could accelerate our abilities to address the other opportunities discussed in this report.

2. OAC Research Opportunities

Presently, many research teams are uncertain about what information is required to assure future transparency and reproducibility of their results. The goal is particularly challenging when the nature of the computation does not permit deterministic computations. Parallel execution, stochastic methods, and lack of control over the versioning of the external software and hardware environments represent some of the largest challenges.

Reproducible research relies on effective and efficient methodologies for capturing the required information needed for future efforts to inspect and reproduce computational results. In order to adopt rigorous reproducible workflows, research is needed in these primary areas.

2.1 Support research in reproducibility essentials (4-1)

Opportunity: Support efforts to explore, catalog, generalize, and standardize the experience of existing reproducibility efforts in order to define the essentials a research team needs to capture in order to assure future reproducibility and ensure trustworthiness.

Present Status: Many scientists and scientific communities [4] have developed local approaches and localized standards for assuring the trustworthiness of their computational results. There are many such approaches based on the experience of seasoned computational scientists. In addition, numerous conferences and journals [5], [6] have established expectations for artifacts, tools, and processes to better assure the trustworthiness of published results. Artifacts include artifact descriptors, reproducibility challenges, complete containerized

environments to encapsulate the software and data used to perform a computational experiment, and more.

Discussion: While there are many approaches to addressing reproducibility challenges, there is also insufficient coordination between communities to learn from or leverage each other's work. Furthermore, there is little research into the fundamentals of reproducibility and the role it plays in the trustworthiness of computational results, and the scientific discovery these results enable. This is in contrast to the lab sciences where experimental setups are recorded in lab books and how this record is to be structured is taught in college classes with a standardized curriculum.

Opportunities for exploration and further advancement of reproducibility methodologies also include characterizing discipline-specific concerns. For example, in high-performance computing, re-execution of a computational experiment may not be feasible due to lack of access to the platform, excessive cost of re-running the experiment, or changes in the computer software or hardware configuration.

Risks and Requirements: There are many distinct efforts to identify localized reproducibility solutions and strong community identity around some of these efforts. Furthermore, there are many levels of understanding, description, and needs for reproducibility rigor. All of these factors mean that defining and articulating essentials will require significant multi-community engagement efforts with public forums for such engagements to occur.

In addition, some scientific communities do not yet have a history and community understanding of reproducibility requirements and may require more incentive to prioritize reproducibility efforts relative to producing new scientific results. In communities where competition for funding is strong and review of results has less rigor, expecting an increased focus on reproducibility may require special shepherding in the short term [7].

2.2 Support improved provenance capture and replay (4-1, 6-3)

Opportunity: Support efforts to provide effective and efficient provenance capture environments and tools that enable computational scientists to more easily provide reproducible results.

Present status: Many provenance capture and replay tools exist (Reprozip, Popper, CK, WholeTale, Code Ocean, etc.). These tools have user communities and are evolving to improve usability. A few tools, for example, ReproZip, are used across many communities. However, the broad spectrum of scientific applications and experiments has diverse requirements, and often, no single tool is able to cater to all requirements. One must mix and match tools to make experiments reproducible. There are also gaps that require new tools and approaches to be addressed.

Discussion: While there are many usable provenance capture and replay tools, we believe more work is required. Many tools already have a user group and support for adding features and helping users. However, we are unaware of any tools and environments that have reached a level of success so as to be available in a turnkey and sustainable way. To qualitatively improve the situation, we need to gain a better understanding of reproducibility essentials and learn about common needs across communities, and about the special needs of some communities. Funding for capture and replay is needed in order for tools to keep pace with our understanding of effective and efficient reproducibility methodologies.

Risks and Requirements: Because there are many existing tools that address only part of the NSF community, funding existing efforts without expanded expectations would not fundamentally change our present situation. Instead, the proposed work should be expressly focused on expanding the potential user base and lowering accessibility barriers, and ideally result in a concerted effort to build an ecosystem of tools to support reproducibility.

2.3 Support advanced reproducibility testing (4-2)

Opportunity: Fund efforts to establish trust-building methodologies that go beyond traditional reproducibility approaches.

Present Status: Computational results obtained by repeating a computational experiment can vary, even for the same input conditions. Changes in algorithms, the use of updated third-party software, and coding errors are all possible sources. Furthermore, many scientific computations involve floating-point computations whose results can vary due to changes in the order of operations as generated by compilers, or may rely on stochastic methods.

At the same time, many scientific software teams use some form of “gold standard” file comparison for assuring that software modifications do not unexpectedly change output results for known input conditions. If output differs from the gold standard, the software team must determine the cause and either fix the error or determine that the change is acceptable (after a labor-intensive, manual inspection) and update the gold standard file for future comparison.

Relatively few software teams have testing methods that are reliable in the presence of these variations. Little research funding is provided to address this topic and community incentives for investing in it are typically low.

Discussion: Use of gold-standard files is a very common approach for detecting potential software problems. It provides a relatively low-cost technique to detect software coding errors, problems with third-party software dependencies, compiler changes, and more. At the same time, the approach assumes that the underlying software and hardware environment will produce deterministic results. Some gold-standard comparisons may support ignoring “noisy” low-order bits, but even then, the choice of what to ignore is typically *ad hoc*. New rigorous approaches to assure effective and efficient reproducibility testing are needed, especially in the presence of dynamic parallelism (where bitwise reproducibility of floating-point computations

may not be guaranteed), use of stochastic processes that are inherently not deterministic, computation on difficult-to-access platforms and software environments where computational scientists cannot readily repeat an experiment.

In addition to more sophisticated approaches for assuring reproducibility of a particular computation, we need more holistic approaches for explaining the results of computational pipelines, including automatically identifying root causes.

Risks and Requirements: Going beyond a gold-standard file may require domain-specific approaches that are not easily generalizable. Progress on this topic may be more about effective high-level approaches and methodologies than about a specific technique that can be broadly applied.

2.4 Support holistic approaches to advancing trustworthiness (6-6)

Opportunity: Support efforts to partner with social and cognitive science teams to characterize and advance reproducibility understanding and trust in scientific results, considering both technical and human factors.

Present Status: Sustained progress in reproducibility often depends on elevating the priority of assuring trustworthy results. The importance of repeatable and reproducible results is often determined within a community. While producing relatively untrustworthy results is bad for a team, investing too much in reproducibility efforts, relative to peers, can lead to producing results more slowly and becoming less competitive.

In most situations, the impediments to improving trustworthiness are not fundamentally technical. Methods, tools, and environments exist to greatly improve the situation, but the community as a whole lacks sufficient incentives to change the status quo.

Discussion: Computational science teams have evolved over time to include increasingly diverse skill sets. Applied mathematicians offer theoretical rigor to computational techniques, computer scientists assure the use of the best algorithms and data structures, and software engineers bring improved tools, practices, and processes. The impact of these roles is very positive. The addition of skills in the cognitive and social sciences can enable a deeper understanding and improvement of how scientists develop and use computational tools to do research.

Removing impediments to improving trustworthiness requires a holistic approach. Partnering with social and cognitive scientists to characterize and advance reproducibility enables consideration of both technical and human aspects. Many tools and processes exist to address technical challenges, but few efforts have incorporated the rigorous use of social and cognitive sciences [4] to understand how to improve the priority of trustworthiness within a community.

Research topics can include understanding what keeps people from using reproducible workflows, better ways to incentivize researchers to include reproducibility into their workflows and publications, and building the infrastructure to support recognition for good work in reproducibility. Additionally, NSF can support research that investigates why panelists do not more highly value reproducibility in proposals, despite the fact that NSF policies ask them to.

Risks and Requirements: Social and cognitive scientists have historically not interacted much with scientists from fields where computational results are produced. Bringing these communities together, and getting the computational sciences communities to understand the value of social and cognitive sciences findings, may be difficult. Building expanded computational science communities that include cognitive and social scientists will require relationship building.

3. OAC Infrastructure Opportunities

Presently, many published computational results are generated from software environments that are loosely managed, with little to no information captured about the software tools and versions of tools used. Furthermore, even if this kind of provenance is captured, there is insufficient ability to repeat or reproduce an experiment in the future because the software environment is unavailable to the author or others, or the costs of learning how to repeat the experiment are too high.

Infrastructure investments, from laptop environments to high-end supercomputing facilities, can provide important workflow capabilities that improve both effective and efficient reproducibility, improving the overall trustworthiness of computational results. The availability of curated and trustworthy software environments will enable communities to standardize their workflows more easily, to obtain trusted results across teams working in a particular community.

3.1 Enable standardized research delivery (4-1, 6-5)

Opportunity: Sponsor creation and standardization of tools, processes, and workflow management systems for capturing provenance for a variety of common software environments.

Present Status: Beyond basic scripting languages like Python and bash, most computational science teams cannot rely upon pre-installed, standardized provenance capturing and replay tools and processes across their computational environments.

Discussion: Perhaps one of the biggest impacts OAC can have on improving the abilities of computational scientists to create reproducible and trustworthy results is to establish common toolsets and processes that are available across computational environments. If we can count on core provenance gathering and replay tools, we should see a large increase in the creation of reproducible workflows and greater adoption of reproducibility in the development and review of computational results.

OAC can use its unique position in NSF, the US, and international scientific communities to lead in this effort. Particular activities can include:

1. *Provide standard tools, systems, and processes for personal computing platforms.*
2. *Support standard tools, systems, and processes on leadership computing platforms.*
3. *Provide provenance capture plug-ins that simplify capturing the state of the software environment in which new and existing software products run. Information captured would include meta-data, software stack details, etc. This collection should be automated as much as possible.*
4. *Provide training on the use of these tools, systems, and processes.*

Risks and Requirements: Developing and maintaining a provenance capturing and replay toolkits will require investment in people and infrastructure, as part of a sustained effort. Presently staffing for this kind of work comes from research software engineers (RSEs). While this is generally a good match in skill sets, the role of RSEs in the computational science community is still emerging and unstable at some institutions.

3.2 Promote community software stacks (6-3, 6-5)

Opportunity: Collaborate in establishing open source community research software stacks.

Present Status: Standard scientific software stacks such as NumPy and scikit-learn exist in the Python community. Within the high-performance computing (HPC) community the Extreme-scale scientific software stack (E4S, https://e4s.io) and the math libraries suite xSDK (https://xsdk.info) are emerging as standard stacks for scientific computation, making available a large and growing collection of open-source, reusable libraries and tools broadly used by the HPC community.

Discussion: Curated, high-quality and standardized scientific software stacks can play an important role in promoting software reuse, rigorous version management, and provenance, and dramatically reduce the amount of code a computational science team needs to write since it can rely on functionality from the scientific software stack.

Coordinating with the US Department of Energy (DOE), other US agencies, and international partners on curated research software stacks, documentation portals, testing infrastructure, and software quality policy can lead to cross-community portability and amortization of resource costs. Further efforts can include establishing funded support for transitioning research software into these ecosystems.

Some of the first opportunities exist in collaborating between DOE and NSF on software for leadership computing facilities, especially as the HPC community transitions to heterogeneous architectures, where software adaptation and support for GPUs and similar accelerators need heavy investment.

Risks and Requirements: Effective scientific software stacks require a sustained investment in stack development and support within user communities. Coordination across US agencies will require investment and support beyond the funding of individual NSF software products, and may perhaps be best done using a software portfolio management approach. Individual scientific application teams are understandably reluctant to adopt third-party software dependencies on products that are not demonstrably sustainable and portable. Any new ecosystems will need to foster trust and make a long-term commitment to providing user support.

3.3 Establish a Research Software Engineer career track (6-6)

Opportunity: Promote and establish research software engineers as permanent members of computational science teams.

Present Status: Research software engineers (RSEs) have emerged as critical members of the computational science ecosystem. At the same time, RSEs still face uncertain career stability because their funding typically comes directly from individual research grants. In contrast, other critical roles such as IT and administrative, are usually sustained positions covered by institutional overhead, even though a large element of these roles is focused on research project support.

Discussion: The RSE role has emerged as an essential and universally-recognized job category in the past decade. Teams that provide a budget for RSEs typically see a strong gain from the dedicated software skills RSEs bring to the team, as a complement to the domain science focus. The challenge for many RSEs is that, despite their strong interest in the RSE career, uncertainty around long-term funding can force them to transition to another more stable career path over time. OAC can lead efforts to stabilize RSE career paths, especially at their leadership computing facilities, setting an example for the broader computational science community.

Risks and Requirements: Because RSEs are in fact funded through research grants, institutions will need to carry the risk of long-term RSE support in between particular funding grants. Certainty of the sustained investment in the RSE role needs to be established.

3.4 Establish a digital asset management plan (6-5)

Opportunity: Establish a digital asset management plan for NSF projects.

Present Status: Data management plans (DMPs) are widely recognized and used with NSF, DOE, and other funding agencies. These plans are submitted as part of a proposal and must contain basic elements about data storage and retention, to assure that data collected by the funded project will have integrity and be available to the research team and the broader community over time. DOE also selectively requests software productivity and sustainability plans from proposals for funding that will require significant software development for sustainable use.

While DMPs are a required element for NSF proposals, in our collective experience, reviewers seldom take into account the quality of a DMP as a critical element in assessing the quality of a proposal. Similarly, we have not seen that there is follow-up as part of project reviews to determine if a project team has followed its DMP.

Holistic management of digital assets is commonly done by many computational science teams, as part of their efforts to assure the trustworthiness of their computational results. But describing or assessing the processes, tools, and policies around these digital assets are not required as part of NSF funding requests and projects.

Discussion: In order to provide a complete and holistic approach to provenance capture, NSF can expand its scope to go beyond a data management plan toward a comprehensive digital asset management plan. Some things to consider as part of the effort are:

1. *Consider collaboration and integration of archival services such as Zenodo, Software Heritage, NSF-funded, institutional repositories, and others listed at http://re3data.org/.*
2. *Address the prevalent use of GitHub, GitLab, and related commercial products as persistent resources when they are not.*
3. *Develop standards and policies for digital asset management.*

Risks and Requirements: Increased requirements for managing digital assets will add upfront costs to research efforts. Initially, this may lead to delays and reduced output from research teams, or require additional funding. The adoption of improved quality practices needs to be introduced incrementally so that new approaches are assimilated effectively and efficiently.

4. OAC Programmatic Opportunities

Most funding strategies for incentivizing reproducible research focus on imposing increased standards for reproducibility and transparency, sometimes with additional funding. However, this approach intermingles objectives for producing science results with objectives for making the work transparent and reproducible.

To foster experience in reproducible research, NSF can provide specific funding for reproducible and confirmatory work to remove ambiguity and enable direct funding for reproducibility and transparency efforts, and direct assessment of these efforts.

4.1 Establish reproducibility training and certification (6-6)

Opportunity: Establish programmatic elements that enhance awareness of reproducibility concepts and provide opportunities for participating in reproducibility activities.

Present Status: Conferences and workshops sponsor reproducibility challenges, badges, and review processes. For example, the Supercomputing Conference series has an artifact review

committee whose only job is to assess the completeness of the artifact descriptor appendix of submissions. In these situations, authors and reviewers necessarily learn some of the key concepts of reproducible computational science.

Discussion: NSF requires specific training from its funded institutions on the responsible and ethical conduct of research (RECR). The information obtained from this training is considered fundamental to conducting any NSF-funded research. A similar approach can be developed to better assure that all NSF-funded research teams have an awareness of the fundamental concepts in reproducible research. Toward this end, the following activities can be considered:

1. *Create reproducibility training materials and provide training modules and events.*
2. *Create REU opportunities to participate in reproducibility efforts for publications, conferences, and workshops.*
3. *Sponsor creation of a curriculum to teach core skills in scientific reproducibility, and, in particular, training modules that can be inserted into existing courses.*
4. *Consider promoting the role of a reproducibility librarian, a person who could train faculty and students in reproducibility, a kind of bootcamp.*
5. *Consider creating a Reproducibility Carpentry, similar to Software and Data Carpentry.*

Risks and Requirements: Increased training in reproducibility concepts will add upfront costs to research efforts. Incremental introduction of this training will be important.

4.2 Elevate reproducibility priorities in project funding and review (6-6, 6-9)

Opportunity: Establish a long-term plan to increase and assess the rigor of digital asset management plans.

Present Status: Presently there is only one type of digital asset management plan, the Data Management Plan (DMP), required for NSF proposals. The scientific community has conveyed that good DMPs are desired, and this desire for reproducible research is already encoded in a number of NSF policy documents. As mentioned in Opportunity 3.4, our experience is that the review of DMPs is not sufficiently rigorous at this time.

Discussion: OAC can lead an effort to introduce a comprehensive Digital Asset Management Plan (DAMP) and increase the importance of these plans. NSF can progressively improve trustworthiness by considering the following:

1. *Provide training and guidance for proposal teams to plan for and construct an effective DAMP.*
2. *Provide training and guidance to NSF program managers and review committee members for reviewing and assessing DAMPs.*
3. *Introduce a required DAMP for relevant proposals.*
4. *Make DAMPs a part of the publicly-available information for NSF proposals, alongside titles and abstracts.*
5. *Include DAMP assessment as part of the review process.*
6. *Ensure that DAMP formats provide content as both human and machine readable.*

Risks and Requirements: Increased digital asset management expectations, infrastructure, processes, and training will add upfront costs to research efforts. Incremental introduction of DAMP scope and rigor will be important, requiring awareness and planning. Engaging the cognitive and social science communities, as mentioned in Opportunity 2.4, could help assure success.

4.3 Support specific funding for reproducibility (6-6, 6-8, 6-10)

Opportunity: Support research proposals that seek funding requests for improving reproducibility.

Present Status: Specific funding for improving reproducibility is not available for NSF computational science projects. However, in other domains, very high-profile efforts have explored the reproducibility of published results, leading to fundamental changes in the way research is conducted.

Discussion: OAC can support confirmatory research activities. Possibilities include sponsoring events to reproduce published results to assure the trustworthiness of research outcomes through confirmatory activities. These activities can be especially suitable for undergraduate and graduate students.

Risks and Requirements: Confirmatory activities do not have a strong tradition in computational science, beyond specific conferences and efforts within a given team to better assure the correctness of their own results. There are cultural challenges that will require consideration, including the willingness of a team to have its results reviewed in this way, and for community members to see the value of committing their time to confirmatory activities.

4.4 Start a working group on reproducibility policies, tools, practices (6-5)

Opportunity: Commission a working group to prioritize, customize and advance the key opportunities described in this Opportunities Report, in particular addressing the items listed in Recommendation 6-5 of the National Academies report.

Present Status: The Working Group on Reproducibility and Sustainability has provided an initial, high-level description of opportunities across all recommendations from the National Academies report.

Discussion: The Working Group on Reproducibility and Sustainability can be extended, or a new working group formed, to go into further detail expanding and proposing concrete advice to OAC to address the items listed in Recommendation 6-5.

Risks and Requirements: No unusual risks or requirements are known.

5. NSF-wide Directions with Strong OAC Components

While OAC cannot drive NSF-wide initiatives, we think the following opportunity is synergistic with OAC efforts. Furthermore, the National Academies report findings are consistent with this opportunity. We understand that this opportunity is outside the scope of OAC, but highlight it in case NSF-wide initiatives might become feasible.

5.1 Establish a reproducibility initiative (6-5)

Opportunity: Establish an NSF-wide Reproducibility Initiative similar to the AI Initiative.

Present Status: There is widespread recognition that a reproducibility crisis exists in at least some scientific communities. Furthermore, the public trust in science to inform public policy is notably deficient. Widely-publicized retractions of results and conclusions, inadequate education on the nature of scientific results, and poor reporting of these results have led to a muted impact of science on society.

Discussion: NSF rightly promotes the excitement and potential for science and its role in society. At the same time, NSF can also elevate the quality and trustworthiness of scientific results by conducting a high-profile campaign to address reproducibility challenges that erode the trustworthiness of science. In particular, NSF can consider the following:

1. *Develop an initiative to advance the reproducibility of NSF-funded research.*
2. *Fund a comprehensive set of R&D, infrastructure, and collaborative activities via research calls.*

Risks and Requirements: It is outside the scope of the ACCI Working Group on Reproducibility and Sustainability to suggest such a broad initiative. We do not have the ability to see the full scope of NSF requirements and priorities so our description of the opportunity may not be urgent relative to other priorities or comprehensive. At the same time, we do see tremendous value in pursuing a high-level reproducibility initiative if it is consistent with other NSF goals.

References

[1] *Reproducibility and Replicability in Science*. National Academies Press, 2019.

[2] "Checklists work to improve science editorial," *Nature*, vol. 556, no. 7701. Nature Publishing Group, pp. 273–274, Apr. 19, 2018, DOI: 10.1038/d41586-018-04590-7.

[3] E. National Academies of Sciences and Medicine, *Fostering Integrity in Research*. Washington, DC: The National Academies Press, 2017.

[4] E. M. Rogers, *Diffusion of Innovations, 4th Edition*. Simon & Schuster, 2010.

[5] M. A. Heroux, "Editorial: ACM TOMS replicated computational results initiative," *ACM Transactions on Mathematical Software*, vol. 41, no. 3. Association for Computing Machinery, pp. 1–5, May 01, 2015, DOI: 10.1145/2743015.

[6] "Reproducibility Initiative • SC21." https://sc21.supercomputing.org/submit/reproducibility-initiative/ (accessed Mar. 30, 2021).

[7] P. E. Smaldino and R. McElreath, "The natural selection of bad science," *R. Soc. Open Sci.*, vol. 3, no. 9, Sep. 2016, DOI: 10.1098/rsos.160384.

Send correspondence to Michael Heroux (maherou@sandia.gov)

Appendix

NSF-related Recommendations from the National Academies Report on Reproducibility

These eight recommendations were called out in the National Academies report to have a significant intersection with NSF's mission. Of these eight, six of them have a potential direct interest for NSF OAC.

RECOMMENDATION 4-1: To help ensure the reproducibility of computational results, researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results in order to enable other researchers to repeat the analysis, unless such information is restricted by nonpublic data policies. That information should include the data, study methods, and computational environment:

- the input data used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;
- a detailed description of the study methods (ideally in executable form) together with its computational steps and associated parameters; and
- information about the computational environment where the study was originally executed, such as operating system, hardware architecture, and library dependencies (which are relationships described in and managed by a software dependency manager tool to mitigate problems that occur when installed software packages have dependencies on specific versions of other software packages).

RECOMMENDATION 4-2: The National Science Foundation should consider investing in research that explores the limits of computational reproducibility in instances in which bitwise reproducibility is not reasonable in order to ensure that the meaning of consistent computational results remains in step with the development of new computational hardware, tools, and methods.

RECOMMENDATION 6-3: Funding agencies and organizations should consider investing in research and development of open source, usable tools and infrastructure that support reproducibility for a broad range of studies across different domains in a seamless fashion. Concurrently, investments would be helpful in outreach to inform and train researchers on best practices and how to use these tools.

RECOMMENDATION 6-5: In order to facilitate the transparent sharing and availability of digital artifacts, such as data and code, for its studies, the National Science Foundation (NSF) should

- develop a set of criteria for trusted open repositories to be used by the scientific community for objects of the scholarly record;
- seek to harmonize with other funding agencies the repository criteria and data management plans for scholarly objects;
- endorse or consider creating code and data repositories for long-term archiving and preservation of digital artifacts that support claims made in the scholarly record based on NSF-funded research; these archives could be based at the institutional level or be part of, and harmonized with, the NSF-funded Public Access Repository;
- consider extending NSF's current data-management plan to include other digital artifacts, such as software; and
- work with communities reliant on nonpublic data or code to develop alternative mechanisms for demonstrating reproducibility.

Through these repository criteria, NSF would enable discoverability and standards for digital scholarly objects and discourage an undue proliferation of repositories, perhaps through endorsing or providing one go-to website that could access NSF-approved repositories.

RECOMMENDATION 6-6: Many stakeholders have a role to play in improving computational reproducibility, including educational institutions, professional societies, researchers, and funders.

- Educational institutions should educate and train students and faculty about computational methods and tools to improve the quality of data and code and to produce reproducible research.
- Professional societies should take responsibility for educating the public and their professional members about the importance and limitations of computational research. Societies have an important role in educating the public about the evolving nature of science and the tools and methods that are used.
- Researchers should collaborate with expert colleagues when their education and training are not adequate to meet the computational requirements of their research.
- In line with its priority for "harnessing the data revolution," the National Science Foundation (and other funders) should consider funding of activities to promote computational reproducibility.

RECOMMENDATION 6-8: Many considerations enter into decisions about what types of scientific studies to fund, including striking a balance between exploratory and confirmatory research. If private or public funders choose to invest in initiatives on reproducibility and replication, two areas may benefit from additional funding:

- education and training initiatives to ensure that researchers have the knowledge, skills, and tools needed to conduct research in ways that adhere to the highest scientific standards; that describe methods clearly, specifically, and completely; and that express accurately and appropriately the uncertainty involved in the research; and
- reviews of published work, such as testing the reproducibility of published research, conducting rigorous replication studies, and publishing sound critical commentaries.

RECOMMENDATION 6-9: Funders should require a thoughtful discussion in grant applications of how uncertainties will be evaluated, along with any relevant issues regarding replicability

and computational reproducibility. Funders should introduce review of reproducibility and replicability guidelines and activities into their merit-review criteria, as a low-cost way to enhance both.

RECOMMENDATION 6-10: When funders, researchers, and other stakeholders are considering whether and where to direct resources for replication studies, they should consider the following criteria:

- The scientific results are important for individual decision making or for policy decisions.
- The results have the potential to make a large contribution to basic scientific knowledge.
- The original result is particularly surprising, that is, it is unexpected in light of previous evidence and knowledge.
- There is controversy about the topic.
- There was potential bias in the original investigation, due, for example, to the source of funding.
- There was a weakness or flaw in the design, methods, or analysis of the original study.
- The cost of a replication is offset by the potential value in reaffirming the original results.
- Future expensive and important studies will build on the original scientific results.

Glossary

The following definitions are from the National Academies report on Reproducibility and Replicability [1], page 42:

Reproducibility is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis.

Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

Additional terms used in this report:

Transparency is providing sufficient details about the data, environment and computational steps used to produce a computational result.

Interpretability is the extent to which it is possible to predict what will happen, given changes in problem details.

Explainability is the ability to describe why the particular results were obtained in a way that is understandable.