CRA Computing Research Association

acm U.S. Technology Policy Committee

July 28, 2021

Scientific Integrity Fast-Track Action Committee (SI-FTAC)
Office of Science and Technology Policy
White House
1650 Pennsylvania Avenue NW
Washington, DC 20502

Re: Request for Information to Improve Federal
Scientific Integrity Policies (86 FR 34064)

Members of the Scientific Integrity Fast-Track Action Committee:

Collectively, the Computing Research Association and the Association for Computing Machinery's U.S. Technology Policy Committee represent more than 80,000 computing professionals in North America and more than 200 academic and industrial institutions engaged in computing research. We appreciate the opportunity to respond to this request for information, provide perspective on some of today's challenges to scientific integrity and highlight opportunities arising from the ubiquity of computing across modern research environments.

We respectfully submit the following observations, which address concerns raised in multiple topics in the RFI, but particularly those concerning Topic 1 "*The effectiveness of Federal scientific integrity policies in promoting trust in Federal science*," Topic 3 "*Effective policies and practices Federal agencies could adopt to address scientific issues and the scientific workplace*," and Topic 5 "*Other important aspects of scientific integrity and effective approaches to improving trust in Federal science*":

1. **Repeatability, reproducibility, and replicability of research: scientific results/data produced by computational artifacts.**

Virtually all of today's science – from life sciences to the physical, social, behavioral and economic sciences – relies upon computational artifacts. Computational artifacts are digital objects used or generated as part of a study or experiment. They can include: databases and spreadsheets that store data collected during an experiment; code to run experiments or simulations; code to produce visualizations of data; large datasets used to train machine learning algorithms; and scripts and software packages that are used to analyze results. Such digital artifacts create significant and multiple new challenges to repeating, reproducing, or replicating scientific results. For example, a 2017 study in the *Proceedings of the National Academy of Sciences* that attempted to reproduce the findings from a random sample of 204 scientific papers published in *Science* highlighted the extent of this problem. The study authors were able to obtain the artifacts used from just 44 percent of the papers and could reproduce the findings from just 26 percent.[1]

The computing research community has been confronting these issues for many years, and as technology has evolved, it has developed and refined best practices addressing many of them. These include guidance on ensuring reproducibility of results, including the use of source code and data repositories, software version control, recording random number seeds and run-time parameters, and the use of virtual machines and software containers (e.g., Docker). The ACM Task Force on Data, Software and Reproducibility is incentivizing reproducibility by introducing "Result and Artifact Review and Badging" for journal and refereed conference paper submissions.[2] This policy defines various badges that will be listed with ACM publications and in digital libraries to recognize papers that have been independently verified. Different badges are awarded for "Results Replicated," "Results Reproduced," "Artifacts Evaluated -- Functional", "Artifacts Evaluated -- Reusable," and "Artifacts Available."

Specifically, we recommend that Federal science agencies: a) consider both cultural and technical obstacles to reproducibility when developing scientific integrity policies; b) recognize the ubiquity of digital artifacts across the sciences, and c) monitor and adopt procedures and policies used in computer science research, tracking changes that are necessitated by continuing technological advances.

2. **Machine Learning contributes its own special challenges for reproducibility, explainability, and transparency of algorithms and datasets.**

While Machine Learning (ML) has been incredibly useful for research and in enabling scientific discovery, its use also creates new challenges for maintaining scientific integrity. As applied, almost all applications of ML effectively function as a black boxes, providing no understandable, explainable or transparent results. Together these and other challenges to repeatability in ML threaten the integrity of research.

ML requires datasets to "train" models for the problems researchers hope to solve. But those training datasets are often proprietary and thus unavailable for reproducibility studies. Even when the data and algorithms are available, they often require exorbitant levels of computational resources to analyze that are not available to most researchers. This is a particular problem because datasets may have biases in them that are challenging to detect or verify without access to the data. The problem can be compounded by the fact that some biases are inherent in what often are considered to be best practices. For example, choosing the largest possible

[1] Importantly, prose descriptions of computational artifacts do not necessarily reflect what the associated code actually does when it is executed. Other examples of hurdles that can face researchers attempting to reproduce the computational work generated by other parties include: the rapidly declining lifespans of both software and hardware before they become obsolete often prevent repeating a calculation or rerunning code under the identical conditions that existed when an experiment was first conducted; data management plans may rely on assumptions that certain public repositories are permanent when they are not; documentation of software configurations, libraries, and experimental setups are often insufficient for an independent party to replicate; and results that depend on proprietary code that is not maintained, easily available, or inspectable.

[2] Boisvert, Ronald F.,"Incentivizing Reproducibility," *Communications of the ACM*, Vol. 59, No. 10., p5 https://dl.acm.org/doi/pdf/10.1145/2994031

dataset will inherently bias against groups underrepresented in the dataset.[3] Further, the algorithms themselves used by the ML systems on the training data are also often proprietary, inhibiting the ability of others to inspect the code and understand how results are derived.

To successfully address challenges around ML and training data, it is critically important for the AI research community that key components of its recently released *20-Year Community Roadmap for Artificial Intelligence Research in the U.S.*[4] be robustly supported. These elements must include, specifically: AI-Ready Data Repositories, AI Software and AI Integration Frameworks, an Open Knowledge Network, and AI Testbeds. Support also is vital to the basic research outlined in the AI Roadmap, and also for data collection, preparation, and maintenance activities that are conducted for the “common good” of the research community, but often do not receive the same degree of recognition (and hence are traditionally less incentivized) within the community. Federal agencies likewise should be encouraged to engage with the community to make their own data more available for AI and ML research.

With respect to the important issue of the privacy of personal and health-related information, the use of synthetic data in machine learning research appropriately is receiving increased attention given its potential to solve several key problems, including: the high cost of collecting and labeling very large datasets needed to train and test such algorithms: the inherent difficulty -- maybe even impossibility -- of safely anonymizing data so that sensitive personal details are never exposed; and, ultimately, the reproducibility of important categories of machine learning research.

These key problems must be solved for everyone to compete on a level playing field that is built on synthetic data protocols. At the same time, however, the usefulness of techniques developed with synthetic data hinges both on the quality of the simulation and the effectiveness of transfer learning (a branch of machine learning aimed at adapting concepts learned in one (artificial) situation to the real world with its real-world potential for unpredictability). This is an active area of research deserving of additional Federal support, along with theoretical techniques for privacy-preserving data mining, differential privacy, and secure multi-party computation.

There is also a need for hypothesis-driven research in ML. Current ML embeds many of the risks of data mining. Lack of integration of hypotheses and failure to seek root causes of failures exacerbate the risks of using ML in security- and safety-critical domains. Without hypothesis-driven research, systematic failures in data labeling may be impenetrable to analysis with data-only ML approaches.[5]

[3] Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

[4] Y. Gil and B. Selman. A 20-Year Community Roadmap for Artificial Intelligence Research in the US. Computing. Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI). Released August 6, 2019. p. 88-93. arXiv:1908.02624 https://cra.org/ccc/wp-content/uploads/sites/2/2019/08/Community-Roadmap-for-AI-Research.pdf

[5] Moriano, P., Pendleton, J., Rich, S., & Camp, L. J. (2017, October). Insider threat event detection in user-system interactions. In Proceedings of the 2017 International Workshop on Managing Insider Security Threats (pp. 1-12).

3. <u>The distribution of false, misleading, or inaccurate information with the intent to deceive is an existential threat to the United States.</u>

Relevant to topic #4, *Effective policies and practices Federal agencies could adopt to improve the communication of science and technological information*, the spread of mis- and disinformation in the current media environment threatens to undo even the best efforts to communicate accurately about Federal science efforts. In a 2020 Quadrennial Paper[6], CRA's Computing Community Consortium outlined the threat:

> In the 21st Century information environment, adversarial actors use disinformation to manipulate public opinion. The distribution of false, misleading, or inaccurate information with the intent to deceive is an existential threat to the United States – distortion of information erodes trust in the socio-political institutions that are the fundamental fabric of democracy: legitimate news sources, scientists, experts, and even fellow citizens. As a result, it becomes difficult for society to come together within a shared reality; the common ground needed to function effectively as an economy and a nation.
>
> Computing and communication technologies have facilitated the exchange of information at unprecedented speeds and scales. This has had countless benefits to society and the economy, but it has also played a fundamental role in the rising volume, variety, and velocity of disinformation. Technological advances have created new opportunities for manipulation, influence, and deceit. They have effectively lowered the barriers to reaching large audiences, diminishing the role of traditional mass media along with the editorial oversight they provided.
>
> The digitization of information exchange, however, also makes the practices of disinformation detectable, the networks of influence discernable, and suspicious content characterizable. New tools and approaches must be developed to leverage these affordances to understand and address this growing challenge. Tools must be developed for security agencies, educators, journalists, civil society organizations, and citizens at large to make sense of, and counter, information pollution. These solutions must incorporate better understandings and models of the "demand side" of the disinformation ecosystem—the consumers of the content—as much as the detection, attribution and characterization efforts support recognition and interdiction on the "supply side," where it originates. Development of such tools and approaches will require collaboration of computer and computational scientists with cognitive and social scientists to better understand this ecosystem and model vulnerabilities in a comprehensive way. As a research topic, the disinformation landscape is a socio-technical ecosystem; research approaches need to meet the new challenges of such a landscape, including adversarial actors and platform companies whose product decisions shape the nature of the threat and its diffusion. Critically, all disinformation solutions must respect ethical principles that balance the privacy and autonomy of individuals online with the societal benefits of understanding and mitigating the threat.

The Quadrennial Paper also delineates the research challenges that need to be addressed to mitigate the threat of mis- and dis-information and calls on Federal science agencies to help:

[6] Bliss N., Bradley E., Garland J., Menczer F., Ruston S., Starbird K., & Wiggins C. (2020) An Agenda for Disinformation Research. https://cra.org/ccc/resources/ccc-led-whitepapers/#2020-quadrennial-papers

- initiate dedicated interdisciplinary research programs at the National Science Foundation;
- create public-private partnerships to support accessible research infrastructure;
- foster cross-agency collaboration, especially between NSF, Department of Defense S&T, Department of Homeland Security S&T Directorate, and the Intelligence Community's S&T efforts, to support the transition of promising research outcomes and secure the integrity of the information ecosystem;
- form cross-agency/cross sector partnerships -- with engagement from the media and industry, among others -- to support education and workforce training initiatives; and perhaps most crucial:
- encourage active, transparent, and good faith participation of the platform companies, whose algorithms and product decisions shape the spread and amplification of disinformation online.

4. **Computational expertise is sorely lacking in most agencies not related to the military or intelligence communities.**

Computation and digital artifacts play an ever-increasing role in citizens' everyday lives, the routing business and government transactions, and both the civil and criminal justice systems. While the legislative branch has made significant strides to institutionalize and fund its consistent access to competent technical expertise, many federal agencies and the judicial branch, in particular, often still lack sufficient such expert input to conduct their mission with scientific integrity and assure the full and fair administrative and judicial process often required by law. Technical fellowship or liaison programs across government should thus be encouraged and supported to a substantial degree and as a high federal priority. Programs like the Jefferson Fellowship program at the Department of State, or the AAAS Science Policy Fellowships ought to be seen as exemplar efforts, broadened and increased, with a special focus on infusing computational expertise throughout government.

5. **The ACM Code of Ethics and Professional Conduct can provide SI-FTAC with useful input about inspiring and guiding ethical conduct across scientific disciplines beyond computing**.

Revised in 2018 after a three-year and highly collaborative international process, ACM's benchmark Code of Ethics and Professional Conduct (ACM Code) has guided the work of professionals in all aspects of computing for almost 75 years. Relevant to Topic 5 "*Other important aspects of scientific integrity and effective approaches to improving trust in Federal science,"* to the extent that ethical professional conduct in science (not simply in computing) will foster both integrity in its practice and public faith in scientifically grounded products and policy, ACM and CRA commend the ACM Code to SI-FTAC to be shared with all scientific professionals. It also may be productively "mined" in the context of this proceeding for fundamental precepts of general applicability and potential public benefit.

The Code's Preamble states in relevant part:

Designed to "inspire and guide the ethical conduct of all computing professionals . . . and anyone who uses computing technology in an impactful way. . . [t]he Code includes [25] principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle.” While “not an algorithm for solving ethical problems,” the Code “rather serves as a basis for ethical decision-making. . . . Questions related to these kinds of issues can best be answered by thoughtful consideration of the fundamental ethical principles, understanding that the public good is the paramount consideration. The entire computing profession benefits when the ethical decision-making process is accountable to and transparent to all stakeholders. Open discussions about ethical issues promote this accountability and transparency.”

CRA and ACM respectfully submit that the ACM Code, appropriately extrapolated from and used to inform targeted messaging to all scientific professionals can help SI-FTAC achieve its goals in this proceeding across all applicable scientific fields and endeavors.

ACM’s U.S. Technology Policy Committee and CRA, and their thousands of expert members, look forward to assisting the SI-FTAC and OSTP throughout their work on this crucial set of issues. Please contact Peter Harsha of CRA (harsha@cra.org) and Adam Eisgrau of ACM (eisgrau@acm.org) with any questions concerning these comments, or for assistance on any computing-related technical matter within the scope of this proceeding, or other matters with which ACM’s and CRA’s expert members may be of assistance.

Respectfully submitted,

Nancy

Nancy Amato
Chair
Computing Research Association

Alec Yasinsac
Vice Chair
ACM U.S. Technology Policy Committee