USC University of Southern California

Los Angeles, April 29, 2022

We write to provide high-level comments on the NIST AI Risk Management Framework from a perspective that is at the intersection between system safety engineering and AI.

We appreciate the emphasis of the framework on risk management, as risk is the factor that AI systems and therefore regulations should be designed to minimize.

Before providing specific comments, we would like to bring up three key points that could be incorporated into the document:

1) In safety engineering, there is a concept of safety-critical systems where their operation serves a mission such that failures have significant consequences. This criticality is something that must be taken into account for AI systems and their operational environments, and the risk management aspects are designed with respect to the criticality assessment.

2) With safety-critical systems, there is always at least a human operator who has received special training in how the system works, how to handle unexpected situations, and how to avert a potential failure of possibly high and catastrophic consequences[i]. The identification of the qualifications and training of operators is crucial in safety-critical systems, and so it should be for AI systems that are safety-critical systems.

3) No existing safety-critical systems in any sectors have been fully autonomous and without a human operator. In contrast, AI systems that can be considered safety-critical systems are already being deployed without full consideration of safe operations with human oversight. Careful consideration and rigorous oversight should be given to the use of AI in any safety-critical systems where any level of autonomy is allowed.

Section 5.1.2 on reliability is very focused on machine learning and machine learning models, and should be reconsidered to encompass all AI technologies. In addition, the traditional focus of reliability in safety-critical systems has been on repeatability and predictability of system behaviors. One possible path forward would be to restrict the use of AI systems to situations where the behaviors of the AI systems are highly predictable (i.e., situations where the behavior has been observed repeatedly to conform to expectations). Another possibility, and this is where AI technologies are being explored and sometimes deployed, is that AI systems can synthesize new behaviors on their own that are novel and never seen before. These AI systems can have this capability to synthesize new behaviors from first principles by reasoning about explicit knowledge relevant to the problem when confronted with a new or unexpected situation. In those cases, the definition of reliability should be reconsidered and the level of decision making

and autonomy permitted for the AI system would have to be concomitant to that new definition of reliability.

Section 5.2.4 on safety could be further refined in terms of two key concepts: failure reduction and consequences. Risk is the complementary notion of safety, and is typically assessed probabilistically based on the likelihood of failures. Design strategies to reduce the probability of failures should be inherent in AI systems, such as redundancy and contingency planning. In addition, once risk is assessed there needs to be an analysis of consequences of the possible failures in terms of human life, cost, and other important considerations.

We hope our input on improving federal support for artificial intelligence risk management is useful.

Respectfully,

Yolanda Gil and Najmedin Meshkati
University of Southern California

BIOS

Dr. Yolanda Gil is Principal Scientist and Senior Director of Strategic Initiatives in AI and Data Science at the USC Information Sciences Institute, and Research Professor in Computer Science and in Spatial Sciences. She received her M.S. and Ph. D. degrees in Computer Science from Carnegie Mellon University, with a focus on artificial intelligence and cognitive science. Her research is on intelligent interfaces for knowledge capture and discovery, which she investigates in a variety of projects concerning scientific discovery, knowledge-based planning and problem solving, information analysis and assessment of trust, semantic annotation and metadata, and community-wide development of knowledge bases. Dr. Gil collaborates with scientists in different domains on semantic workflows and metadata capture, social knowledge collection, computer-mediated collaboration, and automated discovery. She has edited over a dozen volumes and co-authored more than 250 refereed publications. She has held dozens of training sessions on scientific reproducibility, digital scholarship, and open science in conferences, universities, and government labs around the world. In 2019 she co-chaired the community report "A 20- Year Artificial Intelligence Research Roadmap for the U.S." She served in the Advisory Committee of the National Science Foundation's Directory of Computer and Information Science and Engineering. She initiated and led the W3C Provenance Group that led to a community standard that provides the foundations for trust on the Web. She is a Fellow of the Association for Computing Machinery (ACM), the Association for the Advancement of Science (AAAS), and the Institute of Electrical and Electronics Engineers (IEEE). She is also a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI) and served as its 24th President. Dr. Gil is an Advisory Board member for two of the NSF AI Institute awards.

Dr. Najmedin Meshkati is a (tenured, full) Professor of Civil/Environmental Engineering, Industrial & Systems Engineering; and International Relations at the University of Southern

California (USC); an Associate (ex-Research Fellow) with the Project on Managing the Atom at Belfer Center for Science and International Affairs at Harvard Kennedy School; and has been an Associate with the Mossavar-Rahmani Center for Business and Government at Harvard (2018-2020). He was a Jefferson Science Fellow and a Senior Science and Engineering Advisor, Office of Science and Technology Adviser to the Secretary of State, US State Department, Washington, DC (2009-2010). He is a Commissioner of The Joint Commission and is on the Board of Directors of the Center for Transforming Healthcare. He is a member of two boards of the NASEM (National Academies of Sciences, Engineering and Medicine): Board on Human-Systems Integration (BOHSI) and Gulf Offshore Energy Safety (GOES) Board. For the past 35 years, he has been teaching and conducting research on risk reduction and reliability enhancement of safety-critical complex technological systems, including nuclear power, aviation, petrochemical and transportation industries. He has been selected by the National Academy of Sciences (NAS), National Academy of Engineering (NAE) and National Research Council (NRC) for his interdisciplinary expertise concerning human performance and safety culture and served as member and technical advisor on two national panels in the United States investigating two major recent accidents: The NAS/NRC Committee "Lessons Learned from the Fukushima Nuclear Accident for Improving Safety and Security of U.S. Nuclear Plants" (2012-2014); and the NAE/NRC "Committee on the Analysis of Causes of the Deepwater Horizon Explosion, Fire, and Oil Spill to Identify Measures to Prevent Similar Accidents in the Future" (2010-2011). Dr. Meshkati has inspected many petrochemical and nuclear power plants around the world, including Chernobyl (1997), Fukushima Daiichi and Daini (2012). He has worked with the U.S. Chemical Safety and Hazard Investigation Board, as an expert on human factors and safety culture, on the investigation of the BP Refinery explosion in Texas City (2005).

---

[i] Meshkati, N. and Khashe, Y. (2015). Operators' Improvisation in Complex Technological Systems: Successfully Tackling Ambiguity, Enhancing Resiliency and the Last Resort to Averting Disaster. *Journal of Contingencies and Crisis Management (*JCCM*), 23*(2), 90-96.