JOINT COMMENTS FOR AI RISK MANAGEMENT FRAMEWORK: SECOND DRAFT

BY **EMPOWER AI** and **AMERICAN INSTITUTE OF ARTIFICIAL INTELLIGENCE**

This is a joint submission from Empower AI and American Institute of Artificial Intelligence. We appreciate the opportunity to respond to the NIST risk management framework.

Empower AI was built specifically for government missions, to solve its toughest challenges and elevate the full potential of the workforce. We leverage deep technical expertise and decades of experience solving complex challenges in Health, Defense, and Civilian missions. Our solutions give government leaders a direct path to meaningful transformation, equipping them with the insights and tools necessary to make critical decisions faster and move their missions forward.

American Institute of Artificial Intelligence (hereafter American AI) generates knowledge and software to build the backbone of America's artificial intelligence.

Our Interest in AI Governance:

At Empower AI, our highest priority is to be responsible technology innovators, for our customers, employees, and the nation. Before a single line of code was developed for our Empower AI Platform®, we committed to achieve or exceed the highest standards in governance and ethics (G&E), including industry standards, regulations, and all frameworks and models developed by the U.S. government.

American AI developed one of the most comprehensive AI ethics and governance model in 2016 which was shared with the original team of the OECD that designed the first AI ethics and governance model. American AI has authored several books on artificial intelligence strategy and its application and importance in fields ranging from policy and government to finance and supply chains.

Our Comments:

Our comments add to the goal of the NIST risk management framework: to address risks in the design, development, use, and evaluation of AI products, services, and systems.

We have identified nine opportunities to further enhance the framework. Some of those gaps are conceptual, while others are technical.

1. **Redefine the Concept of Lifecycle to Stage Transition:** The concept of lifecycle as used in AI (and the framework) is a remnant of legacy static technology that was developed, used, and then retired. When it comes to AI, systems are not retired, they transform. For example, Google (the search engine) is essentially an AI system, but at this stage, it is inconceivable that the Google search engine would one day be retired – this being consistent with the learning capabilities of the AI systems, where the system will

continuously transition to the next stage of intelligence. AI systems, therefore, should be viewed as stage transitions and not have the traditional born-use-retire pattern. We propose that the concept of *lifecycle* in the NIST methodology should be revised and even replaced by a concept of *stage transition*. The concept of stage transition is derived from changes in one or more of the following: 1) *Problem domain:* the nature of the problem domain changes; 2) *Features:* problem representation is improved by adding or removing new features (variables) that were not considered before; 3) *Data:* Adding new data to existing features to improve the problem representation; and 4) *Model:* Better or more efficient models can be produced. Hence, an AI system's lifecycle is composed of managing changes across these four areas of stage transition and not necessarily retiring the system.

2. **Incorporate System of AI Products vs. Single Product Focus:** The current NIST framework focuses on a single product and assumes that a single artifact is at the center of AI development. In industrial applications, AI products and services function as a series of interconnected AI and non-AI applications. These interdependencies give rise to incremental risks. Therefore, we believe it is essential to approach risk in terms of individual risk (as NIST has currently captured), process risk (risk in the *process* of which a certain artifact functions), and systemic risk (risk related to the entire broad systems). For example, an automated trading algorithm has risks that are part of its performance. However, the algorithm may work in concert with other algorithms and may increase or reduce risk related to the entire process (process risk). Finally, the algorithm contributes to the systemwide risk for the whole market (systemic).

3. **Add Legal to the Characteristics of Trustworthy Systems:** The NIST framework identifies the key characteristics of trustworthy systems as *valid and reliable, safe, fair and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced.* None of the attributes listed include legal. A system can meet all the above requirements and yet be illegal to build or use. For example, an AI system that meets all the above conditions may not meet the legal requirement of a particular country or region. Because laws vary from location to location, for a system to be reliably trustworthy, it must meet the legal definition. We recommend that NIST incorporates language that says, "a trustworthy system must abide by the local, national, and international laws in which it operates."

4. **Make Model Excellence a Required Standard:** Two of the most critical aspects of developing AI are: 1) proper representation of the problem in the model i.e., the model captures the problem domain in its entirety such that it solves the problem exhaustively; and 2) the model used to solve the problem is the best and most efficient in terms of both training efficiency and application effectiveness. These two considerations are important because they create customer orientation and protect customer interests in the AI field. In other words, some companies that claim to develop AI solutions may not offer the best model-problem consistency or may not apply the best techniques that lead to the fastest

process for the best algorithm selection, training efficiency, and performance. We recommend that NIST makes model excellence (i.e. the developer will strive to select the most efficient model that will properly represent the stated problem, will be most efficient to train, and perform best) as part of the framework. This would also imply that developers of AI should disclose to the clients, both government agencies and the private sector, why they chose a particular model, the comparative statistics on the use and performance of other models, and training efficiency.

5. **Include Evolution and Dynamics of Systems:** Changes in problem domains and their representation are essential to AI systems. For example, a trading algorithm developed on trading data from the last 30 years may not capture the unique dynamics of current-day high inflation and rising interest rates dynamics because we have not seen such dynamics in the last three decades. For this reason, we recommend NIST require developers to disclose how the selected model will adapt to future states and trajectories of the problem domains, implying that as a best practice, developers will assess the future changes in the problem domain and provide an evaluation of how their recommended solution will adapt to those changes.

6. **Give Feedback to Social Systems:** The NIST standard properly identifies the three forms of biases: systemic, computational, and human biases. The current approach, however, raises an important issue. Since the data representing social biases represent existing social truths, a system designed to evade those biases will no longer reflect true social biases. Since there will be a split between the real social constructs and the system's constructs, the system will represent a reality that is untrue at a broader level. This reality can lead to a false impression that social evils are cured when, those biases may very well continue to be part of the social fabric. This reality can lead to a false impression that social evils are cured when those biases may very well continue to be part of the social fabric. We recommend that in such cases where data is altered to reflect "what should be" vs. "what is", the developers of the system issue two required annual reports for public consumption: 1) An ongoing report that shows how the society would have made that (biased) decision vs. how the system made the (unbiased) decision; 2) what specific social biases the system overcame due to its design excellence. That way, the system will become a force for good where it will not only do the right thing but also create social awareness about the right things. The alternative is dangerous because while it solves the ethical issue of a given enterprise, it allows society to fester with biases. We have proposed a way for AI to become a source of positive change in society.

7. **AI Ethics frameworks are Mission Centric:** Since both Empower AI and American Institute of AI deal with AI that is applied in both civilian and defense agencies, it is our recommendation that relativeness of mission should be considered as an important determinant of AI ethics for NIST. This implies that no single ethical framework can fulfill the mission requirements of all agencies. For example, the AI ethics framework for FBI can be different than for CIA or NIH or DoD. Since missions are unique to the agencies, the mission

should drive the appropriate ethical framework for the specific agency. We suggest NIST should incorporate language that clarifies the plurality of AI ethical frameworks based upon agencies' missions.

8. **Include Other Lifeforms**: Since AI, both in defense and civilian operations, is often deployed to work with nonhuman animals (for example, home cleaning robots with pets, law enforcement dogs, farm animals, etc.), we suggest that the language of Section 4.2 (Safe, characteristic of trustworthy systems) should add, "nonhuman animals." After the proposed amendment, the Safe standard will read: AI systems "should not, under defined conditions, cause physical or psychological harm or lead to a state in which human life, health, property, **nonhuman animals**, or the environment is endangered". While animals are included under human property, due to the nature of the AI systems, we suggest that *nonhuman animals*, especially service animals, should be included as a separate reference consistent with legislation, and court decisions, and the stated vision of Animal Legal Defense Fund.

9. **Add Reference to Geopolitical Concerns:** The rise of geopolitical conflict, more capable adversaries, and great power competitors require us to reconsider our existing frameworks. We strongly suggest that NIST incorporate language requiring developers to consider the risks associated with involving and dealing with adversaries, their companies, and supply chains. Both data and application value chains should be assessed for risks related to data, technology, methodology, software, intellectual property, or models coming from or ending up in the hands of geopolitical adversaries. While such a requirement will not be necessary during normal times of a globalized economy, we believe that the US policy of limiting the access and involvement of potent adversaries (particularly in AI) in American business and the technological race between geopolitical rivals, the concern rises to a level where it should be formally reflected and adopted in the NIST standard.

We appreciate the opportunity to comment.

Sincerely,

Paul Dillahay
President and CEO
Empower AI

Dr. Al Naqvi
Founder and Chief Executive Officer
American Institute of Artificial Intelligence