NIST EO 14110 RFI response

**Technical Contact**

Susama Agarwala

[susama.agarwala@twosixtech.com](mailto:susama.agarwala@twosixtech.com)  
**Phone:** (202) 386-0750

**Fax:** (703) 543-9662

**Administrative Contact**

Shannon Gardner

[Contracting@twosixtech.com](mailto:Contracting@twosixtech.com)  
**Phone:** (571) 409-6705

# Two Six Technologies’ Response to the National Institute of Standards and Technology’s Request for Information regarding tasking under Executive Order 14110

Two Six Labs, LLC dba Two Six Technologies is pleased to respond to the National Institute of Standards and Technology (NIST)’s request for information (RFI) regarding assignments under Executive Order (EO) 14110. We believe that the NIST Artificial Intelligence (AI) Risk Management Framework (RMF) is a good first step towards addressing many of the issues, both societal and technological, caused by the current way in which AI technologies are developed and deployed.

We also recognize that there are fundamental differences in how AI technologies work, versus, say, how consumer or software products work that makes it difficult to monitor and manage their performance, and similarly, to ascribe responsibility for any errors made by the system.

Nonetheless, we believe that with the right types of tools to both validate and verify the outputs of AI systems, one may properly align incentives both for developers to build responsible AI and allow end users and deployers to responsibly adopt AI.

## Executive Summary

In this comment to NIST’s RFI on EO 14110, Two Six Technologies focuses on the area of transparency and what it means for AI users and deployers, including understanding errors in different contexts, what transparency means after a model enters deployment, and the role and limitations of red teaming efforts. The approach of this document is to consider a simple classifier situation. However, the principles discussed in this document generalize to all forms of AI, including generative models. The classification model is chosen only for its simplicity and explainability.

Two Six Technologies recognizes:

* That AI system safety and assurance is inseparably linked to the context in which it is actually deployed
* That AI systems and safety is not, and cannot be engineered to be, an absolute goal. Instead, it must be specified to the needs of the end users/ deployer

Based on this common understanding, we propose that

* Transparency not be a requirement placed solely on the developer, nor be restricted to the model and the underlying data.
  + Rather, in order for an end user/ deployer to have transparency into the potential harms of an AI model, they need to have access to tools that help them reason through their mission’s specific needs
  + Investments in such tools for the end users/ deployers must be made
* Evaluation of AI safety can only only happen at scale (due, to the stochastic nature of the technology). Therefore, transparency for the end user requires large scale analysis of deployment data
* In order to enact this, in part, the US government should require contracting partners assess model quality over the lifetime of the model’s deployment, in the context of the intended implementation

## Transparency: Beyond model cards and red teaming

While Model and Data cards were a revolutionary take on transparency in modeling, it is not a sufficient solution to provide the end user insight into whether or not the model is suitable for *their* mission. Model cards focus on five aspects of a model: model details, intended use, performance metrics, training data, and quantitative analysis. It also gives recommendations for ethical use of the model.

The issue with model cards, however, is that the AI developer does not know, and cannot know, all the different ways in which the product is going to be deployed, even within scope of the intended use case. One possibility to work around this is to require the deployment only be done under conditions that are statistically non-differentiable from the training data. While this satisfies a mathematical description of correctness and rigor, in reality, these use cases are vanishingly small and will prohibit a model from ever being used.

Therefore, we subscribe to a paradigm where *the end users and deployers are given a suite of automated tools to understand whether or not the model that they are about to use is suitable for their purposes*. In other words, we emphasize the need for a verification toolbox that can be made available to AI consumers to assess the technologies that they wish to use.

In order to illustrate the need for this approach, we begin with a simple thought experiment on classifiers that we use as a working example for our recommendations. While the arguments we pose here are overly simplistic and restricted to classifiers for expositional ease, the same principles they illustrate

1. that measures of fairness are not intrinsic to the model but are heavily dependent on user need and
2. why it is important to incorporate model evaluation and verification transparency into the deployment stages of AI

apply to machine learning models in general. As NIST has pointed out in [previous reports](https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights), the issue of algorithmic bias is far deeper than bad training data, and therefore the approaches to AI fairness need to look beyond collecting better data and model transparency. In the thought experiment below, we argue that issues of algorithmic bias are sometimes fundamental mathematical issues that cannot be resolved with better engineering. Rather, the end user must make tradeoffs to use the algorithm that is the best tool for the job. Therefore, it is imperative that the end users and deployers have access to easy to understand tools that allow them gain transparency into the model impact on their domain, a means of verifying or at least reasoning about post deployment model performance as well as a means of reasoning about potential harms that may come out of deploying a given model.

### Fairness in Classifiers: A thought experiment

Imagine a situation where a classifier is trained on a carefully curated and representative US dataset for deployment of services. For ease of this thought experiment, assume that there are only two classes of people that the technology has to consider, red people and blue people. Cognizant of equity concerns between red and blue people, and desirous of building a fair model, the developers carefully construct a dataset and train the model to have a 1% error rate in both classes. They specify the necessary conditions for the model to work, how to handle missing data, what values are in and out of bound, etc. and report all of this in a model card.

Suppose this model was built with distribution of services in mind. Two users purchase this technology for deployment within the intended use. One is a community mental health program that wants to target at-risk (blue) communities, and the other is a police department that wants to provide better protections for the same blue communities. Both end users consider the analysis of model cards and determine that the 1% error rate for identifying red and blue faces is within acceptable bounds for their purposes.

However, unbeknownst to the police department, the area it is responsible for has 60% red people and 40% blue people, while the (different) area that the mental health service is interested in is comprised of 90% red people and 10% blue people (see confusion matrix below). Due to the large number of variables that go into this classification, it is not obvious to either end user what the distribution of their constituents are.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mental Health | | Predicted | |  | Police department | | Predicted | |
|  |  | Red | Blue |  |  |  | Red | Blue |
| Actual | Red (.90) | .891 | .009 |  | Actual | Red (.60) | .594 | .006 |
| Blue (.10) | .001 | .099 |  | Blue (.40) | .004 | .396 |

Therefore, when used by the police department, therefore, when used by the police department, when the algorithm predicts that a neighborhood is blue there is a 2% chance that it is wrong, and a 7% chance that a neighborhood that is predicted to be not at risk (red) is actually at risk (blue). The numbers are quite different for the mental health service. For them, 9% of the communities predicted to be at risk aren’t, while only 1% of the communities predicted to be not at risk actually are.

This leads to very different outcomes for both agencies that could not have been predicted by use of model and data cards alone.

* Because the actual distribution of red and blue communities for the police station happened to be closer to 50/50, it was far better able to distribute resources misallocating only 2% of its resources, as opposed to the 10% misallocated by the mental health group.
* On the other hand, the same distributional inequalities between the two served communities caused the police department to not reach 4 out of every thousand communities it needed to, as opposed to 1 out of every thousand missed by the mental health group.
* Since the classification of the communities is not directly observable without a thorough sampling and measurement of the deployment populations, it is impossible to know apriori how the model will fail for each of the agencies in this thought experiment.
* Finally, even assuming that the police department and the mental health service provider served the same community, they may have different internal preferences regarding misallocation of resources vs. not reaching individuals. Specifically, if both agencies served an area with 10% at risk communities, then a cash strapped community mental health service may not like the 10% misallocation rate, while the police department is happy to pay that cost to only miss 1 out of every thousand of at risk communities.

In short, the end user/ deployer has many private concerns and trade offs that they need to make for their specific use cases for any given AI technology. Furthermore, it is far too often the case that the ground truth classifications of the deployment scenario is unknown by the end user. Therefore, It is unreasonable to expect the AI developer to be able to predict and foresee these needs and how they will affect the performance of the model upon deployment. Therefore, we emphasize the need for

1. Investment into technologies that allow the end users and deployers to understand whether or not the model that they are about to use is suitable for their purposes.
2. This includes, but is not limited to automated verification tools, provided by the developers, that can be made available to AI consumers to assess the technologies that they wish to use.
3. Commitments by the developers to manage and update their products to changing deployment conditions.

### Transparency and Test Evaluate Validate Verify (TEVV): Beyond redteaming

One of the strengths of NIST TEVV framework, and even more so of the AI RMF’s Map Measure Mitigate Manage (M4) framework is their focus on the part of an AI lifecycle that is completely outside of the control of the developers, namely the deployment stage. Therefore, in order to implement the TEVV and M4 frameworks to models, one needs to look beyond the current state of the art in red teaming or conducting bias bounties.

For instance, let us suppose that in the thought experiment above, the developers have conducted their dataset gathering and model design process with the utmost care, and thus, a bias bounty is not able to find biases in the model for the use cases indicated but the model card, in large enough samples with 50% red people and 50% blue people. Even under these circumstances, the above example shows that passing all red teaming efforts is not enough to address the potential harms of AI adoption.

* We note that it is a fundamental arithmetic fact that, outside of a few special cases, [all forms of fairness cannot simultaneously be enforced](https://arxiv.org/pdf/1703.09207.pdf).
  + Specifically, if a bias bounty hunter were to point out that the distribution of red people and blue people was not 50/50 in the US, but rather 30/70, the developers would have to choose between minimizing different measures of fairness, (false positive/ false negative rates vs differential predictive use rates). Furthermore, these could not both be minimized simultaneously.
* It is in the nature of randomness to be lumpy. As seen in the above example, different users may face different, but unobserved, deployment conditions for the AI technology.
  + It is not practical to give a narrow band of deployment characteristics for which the AI technology is safe due to the unobserved nature of the deployment scenario.
  + Even if the end user could conduct a pre-adoption survey to measure and understand its deployment scenario, limiting adoption to a narrow band of conditions is prohibitive to the adoption of this technology. In the example above, this opens the door to further discriminatory practices: why are areas with fewer blue people not able to use this important technology?
* Different users have different needs.
  + Given that all forms of fairness cannot simultaneously be enforced, there is a need for tradeoffs to be made.
  + As we saw in the thought experiment, the model designer’s choice to prioritize minimizing the difference in misclassification error rates over the difference in prediction error was bad for one user but not the other.

For all of these reasons, we believe that the question of transparency needs to go beyond a reporting on design decisions and datasets, and even beyond the reporting of the findings of a red teaming effort. The thought experiment here shows that even under the most idealized of circumstances, the issue of whether or not the adoption of an AI technology will cause harm is, to a significant degree, in the hands of the deployer or end user. The need for this technology is complicated by the fact that the end users and deployers of AI technologies have a very good understanding of the (frequently non-technical) domain, but little understanding of the technical nuances underlying AI. Therefore, in order to bring algorithmic transparency to the end user of deployer, one must also help them translate between potential harms that AI could have on a mission to the specific harms that AI could have to their area of expertise.

### Transparency and Error Analysis: The need for collective assessment

One of the challenges unique to error detection in AI is the fact that AI systems are fundamentally stochastic. In other words, just because a particular individual was not well served (say misclassified) by an algorithm, does not mean that the algorithm itself is biased. In the thought experiment above, the trained algorithm had a classification error rate of 1% (which by itself is unrealistically low, but was chosen to illustrate a principle while keeping the arithmetic easy). There is no way to make an AI algorithm completely free of error. The question of algorithmic fairness, rather, is how the errors are distributed, and whether different populations of interest are affected differently by the errors. For clarification, by population of interest, we do not mean people, but groups in the dataset the algorithm is acting on. This could refer to types of leaves or radio frequency patterns as easily as people or animals.

Therefore, the assessment of whether or not an AI technology is causing harm is fundamentally different than assessing whether, for instance a car break is not functioning as advertised. The stochastic nature of the algorithm means that there are certain expected (and accepted) failures. Specifically, the assessment of AI failure cannot be the responsibility of a single end user.

We argue that, instead, AI safety needs to be considered in the same manner that the safety (or lack thereof) of living next to a coal power plant is evaluated. Namely, it necessitates a large scale survey of the affected populations and an evaluation of how the stochastic errors inherent in the AI system appear across different populations of interest, and how these differences affect the mission at hand.

From the thought experiment, the actual demographics of the communities served by the police department and the mental health service provider are unknown to the agencies. Only by a large-scale sampling effort can one verify that one has a 40% blue population while the other has a 10% blue population. Note that one cannot simply rely on the model for this information, as the model incorrectly predicts a 40.2% blue population and a 10.8% blue population. While these discrepancies seem small in this particular example, these can actually be quite large in reality. After the ground truth proportions of the population characteristics are determined, one must evaluate the negative impact, if any of the adopted algorithm. For instance, in the case where both the police department and the mental health care provider serve the same community with 10% blue population, the model classification error of 1% is acceptable to the police department, while it is not to the community mental health organization.

In short, the validation stage of the TEVV framework for AI technologies, and the analysis of whether or not it causes harm must be a coordinated large-scale effort, and one that the US government is not only in an ideal position to implement, but already has experience in assessing similar harms.

### Transparency and Contracting: Aligning Incentives

Two Six Technologies is a trusted government contractor in the cyber security space, and an emerging partner in the AI space. We deliver high quality tools to partners across the Department of Defense and the Intelligence Communities. Our business and our reputation depends on delivering highly technical and specialized products that exactly meet the sponsor specifications. While as a trusted partner, we strive to deliver above and beyond sponsor needs as and when we can, there are many circumstances that prevent us from doing so in all the ways that we might like. Specifically, it would be very difficult for us, as a contractor, to independently adopt the TEVV or M4 frameworks recommended by NIST, or indeed, work with our sponsors to implement the transparency framework laid out here if it were not specifically folded into the Requests for Proposals.

The US government is a major procurer in AI technologies. As such, it has the capacity to set industry standards across the country, simply by adopting a more rigorous set of standards for the AI technologies that it procures. If the US government sets high standards for the AI systems it procures from the contractors it works with, those same contractors are more likely to give similar, more rigorous services to commercial partners, and thereby raise the bar for commercial AI development as well.

For this reason, we urge the US government to adopt and a requirement that contractors engage with them in the full TEVV and M4 frameworks. At Two Six Technologies, we want to be able to provide validation of our AI products, as well as the necessary tuning and online learning services that come with the manage step of the M4 framework for our customers. But in order to do so, we need our sponsor to build these needs into their Requests for Proposals and work with us to provide safe, responsible AI.

## Conclusion

Two Six Technologies is a proponent of the M4 framework laid out in NIST’s AI RMF, and the earlier TEVV framework proposed by NIST, and would like to see these become standards for the industry at large.

We believe that these frameworks are a good starting point for ensuring safety and fairness in deployed AI technologies, but are not the full picture of what is needed. Specifically, we recognize that AI system safety and assurance is inseparably linked to the context in which it is actually deployed. Furthermore, we recognize the mathematical fact that AI systems and safety is not, and cannot be engineered to be, an absolute goal. Instead, it must be specified to the needs of the end users/ deployer.

Based on this common understanding, we propose that transparency not be a requirement placed solely on the developer, nor be restricted to the model and the underlying data. Rather, in order for an end user/ deployer to have transparency into the potential harms of an AI model, they need to have access to tools that help them reason through their mission’s specific needs. To enable this, investments in such tools for the end users/ deployers must be made.

Furthermore, evaluation of AI safety can only happen at scale (due, to the stochastic nature of the technology). Therefore, transparency for the end user requires large scale analysis of deployment data. In order to enact this, in part, the US government should require contracting partners assess model quality over the lifetime of the model’s deployment (in accordance with the Validation phase of the TEVV framework or the Manage phase of the M4 framework), in the context of the intended implementation.

We thank you for your time in reviewing this comment.