wiley

AI Risk Management: A Discussion with NIST's Elham Tabassi on the NIST AI Risk Management Framework

Wiley Connected
April 6, 2023

In this episode of Wiley Connected, we are joined by Elham Tabassi, Chief of Staff in the Information Technology Laboratory at NIST, who leads NIST's efforts to create an Artificial Intelligence Risk Management Framework (the "AI RMF"). We discuss the overall goals of the AI RMF (1:31), the use of a risk-based approach to AI (6:02), different categories of risks in AI (10:24), approaches to fairness, bias, and explainability in AI (15:09), core risk management functions for organizations (with a nod to the NIST Cybersecurity Framework) (25:18), how broadly the AI RMF applies and how to define "AI" (30:39); and how the AI RMF fits into international efforts on AI (35:20).

Programming note: This interview was recorded prior to NIST's March 30, 2023 official announcement of the Trustworthy and Responsible AI Resource Center, including the first complete version of the companion AI RMF Playbook.

Transcript

You're listening to Wiley Connected, a series of podcasts on tech, law, and policy. In each podcast, technology-focused lawyers at Wiley, a Washington, DC law firm, break down innovation and law with a uniquely D.C. perspective.

Related Professionals

Duane C. Pozza
Partner
202.719.4533
dpozza@wiley.law

Kathleen E. Scott
Partner
202.719.7577
kscott@wiley.law

Practice Areas

Artificial Intelligence (AI)
Compliance
Federal Policy and Regulation
Privacy, Cyber & Data Governance

Duane Pozza

Hello, and welcome to another episode of Wiley Connected. My name is Duane Pozza, and my co-host today is my colleague Kat Scott, and we are both partners at Wiley. As you know, we have been closely following this development of the AI Risk Management Framework. NIST is the National Institute of Standards of Technology, which is part of the Department of Commerce. The AI RMF, as it is called in shorthand, is an important new tool for companies that are developing or deploying AI systems to identify and manage AI risks. We're thrilled to be joined today by Elham Tabassi, who leads NIST efforts on the AI RMF to discuss the framework. Elham is the Chief of Staff of the Information Technology Laboratory at NIST. ITL supports NIST's mission to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve quality of life. She's won multiple rewards for her contributions, including in the area of biometrics. And most relevant for our discussion today, she leads NIST's efforts to create an Artificial Intelligence Risk Management Framework, culminating in the successful launch of version. 1 in January of this year. It's great to have you on with us, and welcome to Wiley Connected.

Elham Tabassi

Duane, Kat, thank you very much for having me. Pleased to be here.

Kat Scott

Great. Thank you for being here. Let's dive right in. As background, we have a wide array of listeners who are interested in AI, both in its promise and potential, but also in addressing the risks that AI may pose. Can you describe at a high level the AI RMF and NIST's overall goal in developing it?

Elham Tabassi

Yeah, happy to, and as you said, AI has enormous potential to improve our lives and every sector of industry, yet it comes with its own certain and specific risks. So, about AI RMF, directed by a congressional mandate, AI Risk Management Framework, or AI RMF, is a voluntary framework for managing risks of AI systems in a flexible, structured, measurable way. Flexible to allow innovation, but also allow organizations with different size and different resources being able to adopt it. Structured in terms of providing terminology and taxonomy and bringing the community a shared understanding of what's risk, what are the trustworthiness characteristics, what are the challenges to risk management. And importantly, in a measurable way, measurable because if you cannot measure it, you cannot improve it. So, if you are serious in improving trustworthiness, we ought to know what it is, what trustworthiness means, what is considered trust, and how to measure them.

AI RMF adopts a rights-preserving approach to AI. It builds on the OECD recommendation on AI, including its principle for responsible stewardship for trustworthy AI, and is aligned to existing standards happening within ISO, including the AI guidance, the standard on guidance on risk management. What I'm trying to say here is that we try to align it with important works that's going on within the community. Of course, we were not the

first one to develop the framework, but we tried to make sure that we leveraged the good work that has happened within the community and use them and reference them and try to address the risks unique to AI and fill the gaps or voids that the other documents do not address. AI RMF outlines a process to address traditional measures of accuracy, robustness, and reliability, but also importantly, acknowledge that sociotechnical characteristics of AI systems, characteristics such as privacy, interpretability, safety, and bias, which are tied to human and social behavior, are equally important when analyzing and evaluating the overall risk of the systems. These characteristics involve human judgment and cannot be reduced to a single threshold or metric or number.

AI RMF emphasized the need to understand different risk appetites and risk thresholds in different contexts. To that end, it is sector and use case agnostic. It tries to provide a horizontal foundation and an interoperable lexicon for understanding risk management and trustworthiness of AI systems. In terms of intended goal or the larger big intent, AI RMF aims to operationalize values in AI technologies, starting with the values of our society, but also, organizations have the values, so the values of organizations who may design, develop, deploy, or use AI systems. Most importantly, it aims to cultivate a culture of proactively understanding and preventing negative risks and deploy more trustworthy AI. So, in other words, not wait until this technology is built and used to think about risk, but make it from the get go from as early as possible. And what we're trying to do is, going beyond focusing on whether technology works, but also on how it works, where, how, and by whom is technology used, who is left out and why, who is adversely impacted and why that impact happens, and altogether cultivate trust in this technology and managing the risks.

Duane Pozza

Great, and you talked about this a bit, but a key part of the framework is this risk-based approach, and we hear a lot about risk-based assessments, including in AI. So, what does that mean in terms of an approach for organizations that are developing or deploying AI systems, in terms of adopting a risk-based approach or thinking about high or lower risk ways that AI could be used?

Elham Tabassi

Yeah, that's right, AI RMF takes a risk-based approach, and it gets a focus on trustworthy AI systems, which can provide people with confidence in AI-based solutions while at the same time allow innovation to happen and inspire enterprises to develop trustworthy AI technologies. Risk-based approach supports operationalizing common values, protecting the rights and dignity of people, and encourage market innovation. A couple of points here, in the context of AI RMF, risk refers to the composite measure of an event's probability of occurring and the magnitude or degree of the consequence of that event. So, with that definition, because consequence or consequence of the event can be positive, negative, or neutral, this definition of risk can be positive, negative, or neutral. So, AI risk management is as much about minimizing negative consequence as it is about maximizing benefits and potential beneficial use of AI systems.

In terms of the risk management, again, we adopted the definitions from ISO that risk management is coordinated activities to direct and control an organization with regard to risk, and, I want to underscore and say it one more time, that within the scope of the AI risk management, risk management is as much about maximizing positive and beneficial use as it is about minimizing risks. Another thing is that zero risk doesn't exist, so attempts should not be on eliminating risk altogether but understanding the risks and addressing and finding a responsible response to the risks. Altogether, effectively managing the risk or potential harms could lead to more trust for the AI systems and by doing that, build and increase confidence of general public in using AI systems.

I would say one other thing is that AI RMF does not prescribe risk tolerance. While it provides guidance on understanding risk, identifying risk, measuring risk, and managing risks, it doesn't, you know, the risk tolerance how much, for example, private is private, how much safe is safe, how much security is secure, depends on the risk appetite and depends on the context of use, on how AI systems being used and the appetite of the organizations that are designing or developing or using the AI systems. So, the AI RMF does not define high risk or low risk, but talks about risk and risk management because there is no one size fits all and depends on the context of the use, the same technology can pose very high or very low risks. An example they always use is face recognition to unlocking the phone or using our voice for talking to Alexa or Siri or using voice for medical diagnosis. Depend on the context of use, the same technology can have very different level of risks. So, there is no one size fits all, and every time, for each context of use, risk should be understood within that context according to the owner of the application.

Kat Scott

That's really helpful framing, and one thing that you noted was that it's not about necessarily eliminating risk, it's about identifying it and managing risk. So, another big area that we've noticed is this area of defining categories of risks to be addressed and defining different buckets of risks. Can you talk about how those different buckets of risks are identified and defined in the AI RMF and how those were developed?

Elham Tabassi

Yeah, happy to. And maybe this is a point that I will say that everything in AI RMF was developed with close collaborations with the community. We ran and launched an open, transparent, collaborative process starting with a request for information, running three workshops, rounds of drafts for public comments. So, over the 18 months, we heard from more than 240 organizations ranging from technology, civil society, academic and private sector, nationally and internationally standard development organizations. We engaged experts that are computer scientists, mathematicians, statisticians, but also psychologists, sociologists, philosophers, lawyers in development of the AI RMF. All together, we received more than 600 sets of comments. The recording of the workshops, the comments that we received, all of them are posted on our websites, and we ran many, many listening sessions. And so, how those were developed, again, in close collaboration with the community. And the conversations about the risks to the AI systems, let me just also make one other point here that risks to AI systems, risks to any information system and data systems, are also risks to AI systems because at the end, they are information and data systems, but AI systems, as we talked about at the beginning,

impose some unique risks in addition to the risk of information on data systems. In AI RMF, there is an appendix that talk about this, that talk about the reliance of the AI systems on training data, a large amount of training data being used for building models where we don't usually know the granularity that the training data might have, many of the inequalities or viruses in the society be backed into that and several other things.

But how we came up with that taxonomy of risk, I think that's the main point of your question. So, the conversations, particularly in the academic circles, about taxonomy of risks, about trustworthiness characteristics of AI systems, is a lively and energetic conversations and discussions. At the same time, there are high level value-based documents, documents such as the OECD AI recommendation, section 3 of the Executive Order 13960 which talks about trustworthy AI within government, the proposed EU AI Act. They all talk about the values or principles that we want to see in AI systems. So, in our engagement with the community, we try to look at all of them, leverage all of them, and based on those, try to get those values to principles and from principles to practice, come up with categories of risk or, you used the word buckets, and develop the trustworthiness characteristics of AI systems which is valid and reliable; safe; secure and resilient; interpretable and explainable; privacy-enhanced; fair and bias-managed; and accountable and transparent. There are some good work coming out of UC Berkeley that basically talk about 120 or so possible trustworthiness characteristics that has been mentioned in different academic literature and map them into the seven characteristics that I just mentioned. So, the idea was, again, going from values to principles to practice and come down to things that can be meaningful and helpful and usable by designer, developer, deployer, evaluator of AI systems.

Duane Pozza

So, two of those categories or characteristics that you just mentioned are fair with harmful bias managed and explainable and interpretable. Those are both huge topics on their own, but I'm wondering if you could talk a bit about how organizations can start to think about approaching these elements of fairness with bias being managed and also how far organizations should go in making AI explainable or interpretable. And I guess how the RMF can help provide structure and tools to accomplish those goals.

Elham Tabassi

Right. I just want to repeat something that you said that's really important and very, very true, that each of these trustworthiness characteristics are an area of research and topic of research and discovery by themselves, and at NIST we have been building research programs around each of them. True to the nature of NIST and the things that we do, we usually start by terminology and taxonomy because understanding of what it is to measure is an important first step for how to measure building metrics, methodologies, test beds, and benchmarks to do that. So, first you talk about fair and harmful bias managed. First, I want to acknowledge that, as you all know, bias exists in many forms and can become ingrained in the AI systems that assist decision making. And some of those systems can influence our daily lives. The other thing that I want to say is that, as you said in the title, we say harmful biased managed because bias is not always a negative thing. Sometimes bias is intentionally being designed if you think about bias as sort of a demographic disparity. An

example of that is, for example, car insurance payments, that for different ages, there is different level of payment. But while bias is not always a negative phenomenon, AI systems can potentially increase the speed and scale of biases and perpetuate and amplify the harms of some of those biases to individuals, group communities, or society.

You also talked about interpretability and explain-ability, and I think these are all very much tied to each other. In terms of fairness, which is also in the title, fairness in AI includes concerns of equality and equity by addressing issues such as harmful bias and discriminations. Fairness is a complex and difficult to define topics. There are many, many different definitions of fairness and mathematical definitions for fairness that exist, but it also has a dependence on the culture, and perception of the fairness can differ among different societies, different cultures, and may shift depending on the applications. An organization's risk management will be improved by recognizing and understanding these types of differences. In terms of the definition of the bias, NIST put a publication out in March 2022, Special Publication 1270 Towards a Standard for Identifying and Managing Bias in AI, and the report talks about bias being broader than just demographic disparity and bias in data representative. It talks about bias having the three major components that ought to be understood and managed: systemic, computational and statistical, and human cognitive. Computational and statistical biases is, again, the one that everybody's familiar with, that biases can be presented in data sets and algorithmic processes. But systemic biases is another type of biases that's really important and ought to be understood and measured. Systemic biases can be present in AI data sets, in societal or organizational norms, practices, and processes across the AI lifecycle. And the third category is human-cognitive biases, which relate to how individuals or group perceive AI system information to make decisions or to fill in misinformation, or how humans think about purposes and functions of AI systems. A majority of the conversations about bias and bias mitigations focus on computational and statistical biases and ignores the other two biases. I will just say that just because a lot of focus has been on the understanding computational statistical biases – we probably have a better handle over how to measure and understand those – the other two type of biases, systemic bias and human cognitive biases, are equally important to be recognized, identified, and measured, and I think that's where a lot more work needs to...we are not as advanced and as comfortable on understanding these biases as we are with the computational biases.

The other thing that you talked about is explainable and interpretable component. You notice that we distinguish between the terms explainable, interpretable, and transparent. And these three terms in many documents are being used as sort of exchangeable, but they are not. Transparency, the way we are using it in AI RMF, is about transparency of the whole systems and documentations at the different level. Documentation on any decision made across the AI lifecycle, across the risk management for the different part of the AI lifecycle, helps with the transparency. Explain-ability, the way we use it in AI RMF, refers to the representations of mechanisms underlying AI systems operation. So, it's about the parameters, the number of the layers, the weights of AI model. Interpretability, on the other hand, refers to the meaning of AI systems' output in the context of their design functional purpose. So, when you're talking about explaining AI output, within the context, we're really referring to interpretability however it has been defined in the AI RMF. So explain-ability and interpretability, both of them help with those operating or overseeing an AI system, as well as users of the AI system, to get a better, deeper, understanding and insight into the functionality and trustworthiness of the

AI systems. I would just say that interpretability is a difficult thing. And then you asked about how far organizations should go in making AI explainable and interpretable. So, with the distinction in the definition of explain-ability and interpretability, explain-ability can help the developer and maker of AI systems basically provide guidance and input and insight on how the model works, how the system works. Interpretability is more about understanding and explaining the output of the AI systems. I would say interpretability is really important. If AI systems is going through job applications and resumes and denies or rejects some resumes, it's good to have some explanations of why that happened. Or if a loan application is getting denied, or if a AI systems is looking at the medical image, the scan of brain image for a decision if there's any brain tumor there or not, and the algorithms makes a output and basically says that here is a tumor, here is not. Some sort of explanations, some sort of descriptions of how the algorithm make that output would be very helpful. And another difficulty here is that depend on the different audience, depend on who the explanation or interpretation is for, you need a different level or different specificity in the interpretation. If the explanation goes to the patient versus the technician versus the physician in terms of the interpretations of the medical images, different information and different mechanism or communication level is needed to explain the results.

Kat Scott

Absolutely. Switching gears just a little bit. Broadly speaking, the AI RMF is similar to other NIST frameworks, most notably the Cybersecurity Framework. You were mentioning some characteristics of the AI RMF earlier, and there are a lot of shared characteristics between the Cybersecurity Framework and the AI Framework, including that they're both voluntary and flexible. One key characteristic that the AI RMF shares with the Cybersecurity Framework is its structure. Both frameworks consist of core functions which are essentially risk management activities grouped at their highest level. For the AI RMF, the four functions are Govern, Map, Measure, and Manage. Can you describe for listeners how those work, and how they could be implemented by organizations?

Elham Tabassi

Yeah, thank you for that question. I mentioned that we tried to do lessons learned and learn from past experiences and leverage all the good work. And one of them was that CSF Cybersecurity Framework that NIST had put out. It's a very well received document, so we tried to learn as much as we could from the process and one of them was going to be the same structure because again it was very well received. So, AI RMF groups its guidance for risk management at a high level in four functions of Govern, Map, Measure, and Manage. And again, similar to Cybersecurity Framework, each of these guidance, each of these functions, now then are divided into categories and then subcategories to provide a kind of gradual specificity or more detailed guidance.

So, sort of a high-level description of these four functions. Govern defines the structures, systems, processes, procedures, roles, and responsibilities of the teams that needed to be in place for an effective, efficient risk management that helped cultivate a culture of understanding and identifying proactively and purposefully and continually understanding risk, risk assessment, and responses to the risks. So that's the guidance in the

Govern function. The Map function provides guidance about establishing context that's needed to frame the risk related to AI system within that context of use. From the trustworthiness characteristics that we just talked about, which one of them relate to the context of use, what are the interplay of these trustworthy characteristics within the context of use, and who is being impacted, the degree of impact. So, all of those guidance is within the Map function. The Measure function uses quantitative, qualitative or mixed methods, tools, techniques and methodology to analyze, assess, benchmark, and monitor AI risks and their related impacts. It basically use the knowledge and information developed during the Map function to come up with some sort of a assessments and measurement of the risks identified in the Map functions or identified as part of the process of the measure. Measure functions use the informations from the map and measure and gives guidance on allocations of risk resources to the risks mapped and measured in the previous stages and give guidance on how to develop plans to respond, to recover from, and communicate about incidents and risks or events. There's no set order of which function to use first and next. We generally suggest that going from map, measure, and manage, and govern, is we say that that the function of map, measure, manage is in a sea of governance. Everything starts with governance and with governance, governance really gives the right infrastructures right foundation for good risk management. Another important thing is that none of these functions is something that is just being done once. We talk a lot about continual risk management, continuing and frequently go and do the cycle of map, measure, management, it depends on the law of the AI act during the life cycle, keep iterating over the activities and the guidance. It's not, again, a one-time done deal but a regularly and continually understanding the risks, assessing, measuring, and managing them.

Kat Scott

So, how broadly applicable is the AI RMF? I mean, we maybe should have started here, but how would you define AI generally, and does the AI RMF's applicability in terms of applying these functions really depend on a strict definition of Ai?

Elham Tabassi

Yeah, that's a very good question. So, you may notice that in AI RMF, we don't define AI but AI systems, and the scope of the AI RMF is risk management of AI systems. And this gives us a better landscape boundary to work on that because AI is a scientific discipline, and its definition has been changed in the past 65 years or so.

While we were working on providing a definition of AI system and AI RMF, again, we consulted with the whole community, and we were very cognizant to adopt a definition, like anything else in the AI RMF, to be able to stand the test of time, at least for a short time, given the rapid pace of changes in this environment. So, generative AI, ChatGPT, is something that a lot of people are thinking about this in the past month or so, but when we were developing AI RMF, we were looking at those and making sure that the guidance of AI RMF is applicable to them. So, again, going back to the definition, we don't define AI, we define AI system. We adopted the definition of the OECD on AI systems, and the definition that we have is the AI system as an engineered or machine-based system that can, for a given set of objectives, generate outputs such as prediction recommendations or decision influencing real or visual environment. And AI systems are designed

to operate with varying level of autonomy, so this definition that Internet or machine-based system that for a given set of objective generate outputs such as predictions recommendations is applicable to generative AI. And the risk-based approach as we talk at the beginning of the call is a very powerful approach to understanding the impact and consequences of the output of the AI systems and is applicable to generative AI. Granted, that risk profile for different categories, for example, risk profile for generative AI, would be more complex and understanding of that would be much more challenging than risk profiles for, for example, deep neural networks, which in turn, is more challenging and complex than risk profiles for logistic regression. Logistic regression inherently is interpretable and explainable but deep neural are not and generative AI makes it even harder. So, in terms of a broad applicability of AI RMF, I think we will see what happens next. I think whatever we know now in six months would be very old information, with all the things happening too quickly. But in the developing of AI RMF, we try to avoid any sort of strict definition or prescribing anything. We had these attributes of the AI RMF definitely early on become an impact to the community and one of them was flexible. That's all I have to say, that we don't define AI, we define AI systems, and the definition is, at least for now, general enough to cover AI systems being discussed right now.

Duane Pozza

Yeah, that's certainly a hot topic of generative AI models right now. We have a last couple questions about looking forward, obviously much more to talk about with this. You'd mentioned at the beginning the work that NIST is doing is very much grounded in consensus activity including at the international level, the OECD, for example, on risk management and approaches to AI. Obviously, a big thing going on right now is the EU's proposed AI Act, which is going through some drafts and changes right now. But one question is, how do you see, if at all, the AI RMF fitting in with the work that Europe is doing, with the proposed AI Act or standards development generally? I know there are coordination activities going on with the US-EU Trade and Technology Council, for example. Is NIST involved in those, and how do you see these efforts fitting in?

Elham Tabassi

Yeah, so we released that roadmap when we released the AI RMF, and three things top on the list of the roadmap is alignment with standards, is advancing test evaluation verification validations, and third is development of the AI RMF profiles. So, your first questions about standards and measurement, both of them high up on our list. We had a cross work from AI RMF to one of the ISO standards on risk management. When we put the cross work out, the ISO standard was in...that's why we just did the cross work at the section title, but we intend to be more active in participation in standard development activities, particularly within ISO, the text of AI RMF. Everything in AI RMF, as we just talked, reflects the wisdom of a really large sector of the community, so we should be able to contribute all of those knowledge, all of those wisdom to international standard development. In terms of measurement, we started more than a year ago on trying to set up evaluations with human in the loop. We already have some really good evaluations I think well received by the community on measurement of the biometrics, information retrieval, so we will be working on those.

In terms of the Trade and Technology Council, yes NIST is involved, and we have been very much involved in development of the joint AI roadmap that was delivered as part of the second TTC deliverables on December 5. And we also did a workshop on February 16 that basically talked about the joint AI roadmap which, what it tries to do, is to kind of figure out areas that we are in alignment and by doing that also shed light on the possible gaps, such as establishing three working groups and starting work in three areas on shared terminology and taxonomy, starting with definition of terms such as risk, bias, and interpretability; another one on importance of commitment to work towards international standards; and the third one around measurement and working together on advancing measurements of risks and trustworthy AI.

Kat Scott

So, building on that work and I guess our final question looking forward is how can organizations get involved here? We know that the version that NIST released is version 1.0, so not the final word. What are you looking for in terms of external feedback and what's the best way for organizations to engage moving forward?

Elham Tabassi

Yeah, thank you for that. We definitely need to keep the engagement. I always say that the only reason we are able to do our work and put good quality work out is because of the support of the community and external engagements. What they can help, first of all, any comments on the playbook. So along with AI RMF we also released AI RMF playbook that provides more guidance, more descriptions around how to use, implement AI RMF, provides suggested actions, informative references, and ways to improve transparency for each of the guidance in each of these subcategories. The playbook will be updated every six months because a lot is happening and to keep up with the pace of developments. So please look at the playbook, give us input if you're using AI RMF, and let us know how you're using it. Another item on the roadmap is measuring the effectiveness of the AI RMF. We won't be able to measure the effectiveness of AI RMF if we don't hear from the community or know how they are using it. Another area that we like to get involvement and engagement is development of the use cases and profiles. Profiles are a tailored set of guidance in AI RMF to certain or specific use case AI RMF profile for hiring would basically slant or tailor or make every subcategories very specific to the use of AI in hiring. So, there's a call for contributions on development of the AI RMF. That's another way that they can be involved. Standards and measurement, again, any ideas that they had. We're trying to build more test beds and benchmarks around measurement of the AI systems from the sociotechnical lens. All of the information is on our websites. How to contact us is also on our website. It is a version 1.0. We put a revision cycle of every 3 to 5 years built into the AI RMF, for the playbook every six months, so just please keep the engagement. Let us know of the tools, methods, and metrics that you are encountering, you're using, for measuring any of the trustworthiness characteristics, suggestions on how to measure the effectiveness, suggestions for advancing standards and measurements, and how you're using the AI RMF, all of those will be extremely valuable to us.

Duane Pozza

Thanks so much again for joining us today and thanks for all your efforts on the framework. It's such an important and exciting development. This has been a really interesting and informative discussion. Thanks again.

Elham Tabassi

Thanks so very much for having me. Really enjoyed our discussion. And to everybody listening, thank you very much and please keep in touch.