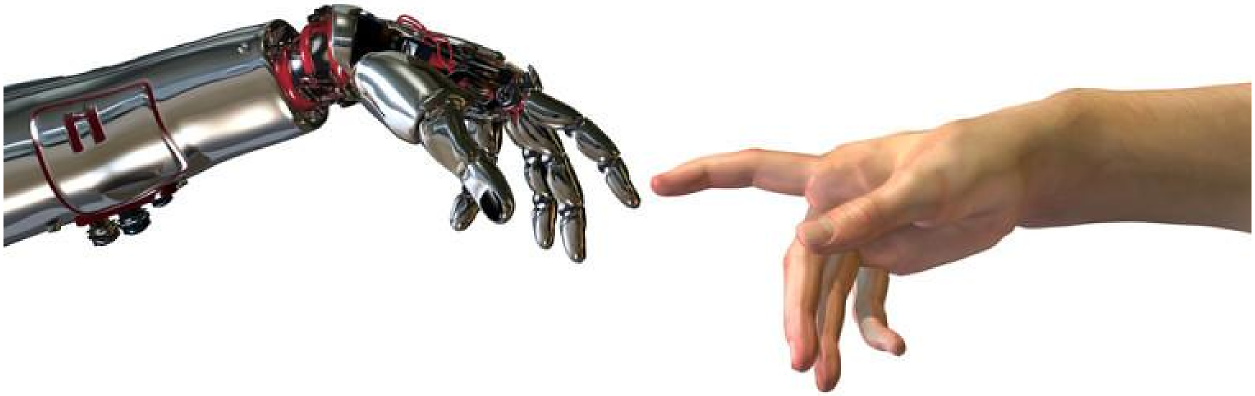


Sept 14, 2022

In regards to NIST AI RMF Draft #2 Feedback

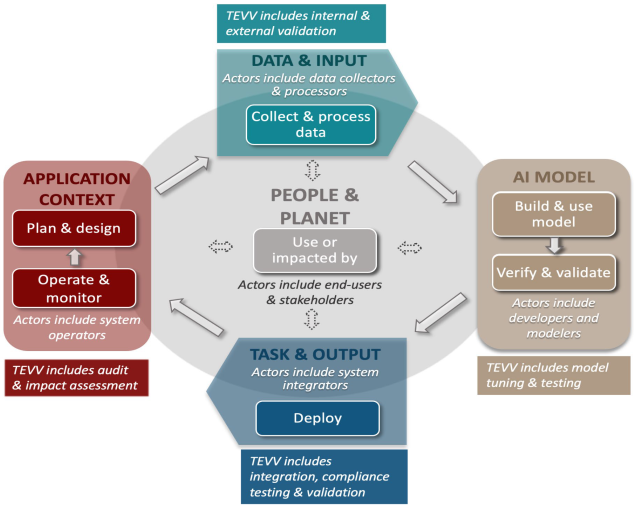
Introduction and background

ForHumanity[1] is a non-profit public charity, with 1100+ members from 72 different countries. We provide services, on behalf of humans, to governments and regulators, such as NIST. We are currently under contract as a technical liaison to CEN/CENELEC JTC 21 for technical advice on the EU Artificial Intelligence Act (as proposed). In this context, we have developed a comprehensive set of auditable criteria to allow approved certification bodies to independently evaluate compliance with the law. The law calls for a comprehensive risk management framework, which we have developed - available here. ForHumanity University also provides expert-level education and accreditation on our risk management framework through online education and certification, which can be accessed here. **Both of these references are free-to-access and may be included in the NIST Trustworthy and Responsible Resource Center**.

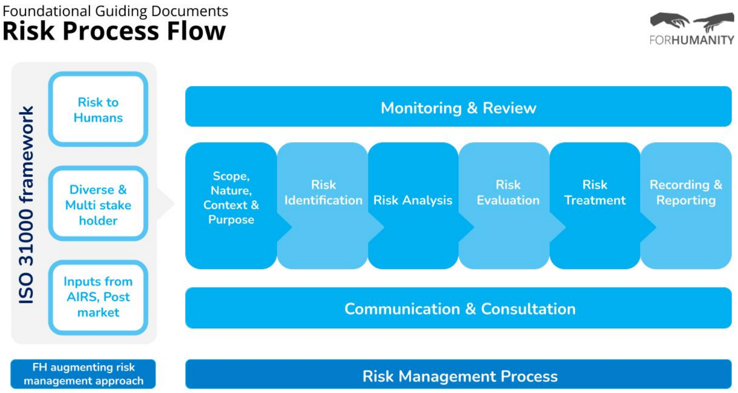
[1] ForHumanity (https://forhumanity.center/) is a 501(c)(3) nonprofit organization dedicated to addressing the Ethics, Bias, Privacy, Trust, and Cybersecurity in artificial intelligence and autonomous systems. ForHumanity uses an open and transparent process that draws from a pool of over 1100+ international contributors to construct audit criteria, certification schemes, and educational programs for legal and compliance professionals, educators, certifying bodies, developers, and legislators to mitigate bias, enhance ethics, protect privacy, build trust, improve cybersecurity, and drive accountability and transparency in AI and autonomous systems. ForHumanity works to make AI safe for all people and makes itself available to support government agencies and instrumentalities to manage risk associated with AI and autonomous systems.

Conformity of NIST RMF with ISO 31000 and COSO ERM/ORMs

NIST is welcome to adopt and adapt this operational risk management framework into its burgeoning AI RMF. The ForHumanity AI Risk Management Framework is operational within Independent Audit of AI Systems, the EU Artificial Intelligence Act (as proposed), UK and EU GDPR, as well as COSO and ISO 31000 risk management frameworks. We note that draft #2 of the NIST AI Risk management framework has incompatibilities with some basic terminology in COSO and ISO 31000 as noted in the process flow diagram below and juxtaposed to the NIST AI RMF process flow diagram



TEVV includes internal & external validation
DATA & INPUT
Actors include data collectors & processors
Collect & process data
APPLICATION CONTEXT
Plan & design
Operate & monitor
Actors include system operators
TEVV includes audit & impact assessment
PEOPLE & PLANET
Use or impacted by
Actors include end-users & stakeholders
AI MODEL
Build & use model
Verify & validate
Actors include developers and modelers
TEVV includes model tuning & testing
TASK & OUTPUT
Actors include system integrators
Deploy
TEVV includes integration, compliance testing & validation



Foundational Guiding Documents
Risk Process Flow
FORHUMANITY
ISO 31000 framework
Risk to Humans
Diverse & Multi stake holder
Inputs from AIRS, Post market
FH augmenting risk management approach
Monitoring & Review
Scope, Nature, Context & Purpose
Risk Identification
Risk Analysis
Risk Evaluation
Risk Treatment
Recording & Reporting
Communication & Consultation
Risk Management Process

As can be seen from a comparison of the process flow graphics, they accomplish similar missions, but use different terminology that may create implementation and integration problems with existing organization's ORMs and ERMs based in COSO and/or ISO 31000.

Consequently, we would recommend a shift in terminology to:

1) Risk Identification
2) Risk Analysis

3) Risk Evaluation
4) Risk Treatment

Word choices to describe the process flow amount to minor tweaks for clarity and conformity with the international standards (a stated goal of the NIST AI RMF framework).

Fundamental Critiques of Principles found in the AI RMF

1) **Unmitigated Risks/Residual Risk:** The NIST AI RMF has no discussion or process for handling unmitigated risk. Residual Risk (defined as the sum of all unmitigated risks) directly impacts End Users when interacting with an AI system. A fair, transparent, trustworthy and responsible AI system should disclose all Residual Risk to end users to inform them about potential negative impacts of interacting with the system. Informed users will then be able to evaluate if the benefits from the system outweigh the risks. This disclosure of Residual Risk is compatible with two valuable risk management models already being implemented.
   a) The EU Artificial Intelligence Act (as proposed) calls for the disclosure of Residual Risk for Annex III High-Risk AI Systems as noted in Article 9.4

      The risk management measures referred to in paragraph 2, point (d) shall be such that any residual risk associated with each hazard as well as the overall residual risk of the high-risk AI systems is judged acceptable, provided that the high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse. Those residual risks shall be communicated to the user. In identifying the most appropriate risk management measures, the following shall be ensured: (a) elimination or reduction of risks as far as possible through adequate design and development; (b) where appropriate, implementation of adequate mitigation and control measures in relation to risks that cannot be eliminated; (c) provision of adequate information pursuant to Article 13, in particular as regards the risks referred to in paragraph 2, point (b) of this Article, and, where appropriate, training to users. In eliminating or reducing risks related to the use of the high-risk AI system, due consideration shall be given to the technical knowledge, experience, education, training to be expected by the user and the environment in which the system is intended to be used

   b) The well-tested FDA drug and clinical trial model that allows the marketing and selling of FDA approved drugs to share both the benefits of the drug and the known side effects - in other words, side effects are the Residual Risks associated with the drug.

ForHumanity recommends that the inclusion of required disclosure of Residual Risk is the most critical adjustment NIST could make in the AI RMF. Disclosure of

Residual Risk will create informed users, that include both downstream acquirers of AI systems and end users, including natural persons. Disclosure of Residual Risk also demands that the provider of the system has fully considered the risks of deploying the product.

---

2) **Ethical Choice:** The risks associated with Ethical Choices made in the design, development, deployment, and decommissioning of AI systems remain neglected in the NIST AI RMF. The language has significantly improved by introducing the wording "socio-technical" systems. However, the identification of risk associated with the interface between natural persons and machines has not fully been captured. Humans operate with moral frameworks, and oftentimes these moral frameworks are shared by many, but under the strain of specific decisions regarding instances of Ethical Choice, shared moral frameworks may diverge. Frequently today, these decisions regarding instances of Ethical Choice are handled by untrained and sometimes even unaware designers, developers, and data science teams, instead of by trained AI Ethics experts.

   In addition, both the UK government and now California have passed bills commonly known as the Children's Code. These laws are progressive and govern the actual interface between Children and Artificial Intelligence, especially in the area of Personal Data collection and use. These laws introduce numerous instances of Ethical Choices, especially in the physical design of the interfaces and therefore require unique and specialized expertise to examine the pros/cons or tensions/trade-offs to document and reach conclusions about these instances of Ethical Choice.

   Instances of Ethical Choice create risk to end users and are not accounted for in the NIST AI RMF. Examples of these risks include this small subset listed below of the discipline called Algorithm Ethics:

   1. Necessity assessments
   2. Proportionality Studies
   3. Concept Drift monitoring
   4. Data Representativeness
   5. Sufficient accuracy,
   6. Construct validity and ground truth
   7. Innovative AI applications and comparable industry-standards
   8. Diverse Inputs and Multi-Stakeholder Feedback

ForHumanity would recommend identifying Ethical Risk as a critical consideration in socio-technical systems and further recommend objective, expert evaluators to manage these risks.

Further Abbreviated Comments on Draft #2 NIST AI RMF

- **People impact:** The governance process is defined to focus on business impact, and business context and not the context of people impacted.
- **Risk categories considered:** NIST considers Trustworthy AI valid, reliable, safe, fair, and one where model bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced. However, it does not consider *Diverse, Governed, Ethical and Human Controlled, rights & freedom are upheld and Sustainable*.
- **Privacy enhancement vs Privacy assurance:** While NIST AI RMF considers privacy, it is focused on Privacy enhancement, not Privacy assured and Data Protection by design. It also tries to represent that privacy-enhanced includes human rights and freedom. This is not true. Human agency, rights and freedom, and preserving human values are distinct from privacy expectations.
- **OECD classification:** Should NIST AI RMF continue to use OECD classification of AI systems to determine AI risk and impact, then the risk assessment process will be hindered in achieving sufficient precision. The human assessors need the freedom to identify risk inputs and risk indicators specifically to properly evaluate and analyze the risk to develop sufficient risk treatments
- **Ground truth validation as part of Testing and evaluation:** Testing, Evaluation, verification, and validation must include ground truth validation, just as the DoD AI Risk Management framework has called for.
- **Emergent Risks:** NIST AI RMF mentions risk measurement as a challenge. It also mentions the need for considering emergent risk but does not provide details on how that needs to be collected or gathered.
- **Risk Appetite and Risk Tolerance**: NIST AI RMF speaks about defining risk tolerance as a challenge. They do not mention anything about risk appetite, which should come first. Establishing risk appetite acceptance and defining risk tolerance is not currently in NIST AI RMF. Risk Appetite and Risk Tolerance are conflated. While they should be two unique considerations, a process flow starts with the organization's acceptance of risk appetite and then a metrics-based definition of Risk Tolerance.
- **Mapping with EU AI Act:** The mapping of the EU AI ACT is wrong in several areas - we would be glad (based on our technical status with CEN/CENELEC) to directly assist the drafting team with the management of their chart.
- **Accountability:** Accountability is a fixed responsibility within an organization (across all AI actors). However, the broader perspective of accountability where it takes into

account individual responsibility, function responsibility, entity responsibility, ecosystem responsibility etc. In addition, it does not speak about authority and resourcing for Risk management. We also see a general deficiency of accountability and oversight with the C-suite and Board of Directors.

- **Risks associated with tools and API's:** Validation is focused from the perspective of training data; no consideration is provided for risks arising from API, GitHub, open source, auto code, co-authoring code, etc. It also does not consider applications that use reinforcement learning or federated learning where the risk and control context evolves differently. It, however, states the risks associated with the supply chain and third parties in general. Testing, evaluation, validation, and verification are part of risk management, requiring monitoring or conducting metrics tests instead of overseeing associated risks.
- **Risk Disclosure**: NIST AI RMF does not have any reflection on disclosure. It only deals with risk control and internal reporting. Traceability of controls, treatments, and mitigation should be at least documented and in critical areas, disclosed for sufficient public oversight (not including trade secrets or IP). Furthermore, this is incompatible with the concepts of Trustworthy and Responsible AI founded on transparency and governance.
- **Adverse Incident Reporting System:** NIST AI RMF classifies "user impacted by" as one of the AI risk actors. This category covers insights or risk inputs from users, the general public, etc. However, the need for a structured mechanism like an Adverse Incident Reporting System (AIRS) or handling a whistle-blowing tool for complaints is not mentioned. While these may appear as an extension as things evolve, this must be clearly articulated.
- **Diverse Inputs & Multi-stakeholder Feedback**: Diverse Inputs & Multi-stakeholder Feedback (DI&MSF) is offered as an expectation (as mentioned above). However, the approach and risk management associated with DI&MSF, including DI&MSF acting as an attack vector, are not dealt with in the framework. Furthermore, sufficient diversity is a matter of Ethical Choice, best handled by trained ethics officers acting according to the Code of Ethics.
- **Post-market monitoring:** NIST AI RMF does not provide any reference regarding post-market monitoring mechanisms, including AIRS, ground truth validation, black box (for enabling future review of events) for all autonomous systems, etc.
- **Human-in-the-loop risks:** NIST AI RMF mentions the need to appropriately define and measure Human-in-the-loop or Human-on-the-loop (HTL). However, the monitoring of risks and effectiveness of controls and treatments associated with HTL are not covered in NIST AI RMF.
- **Risk, Reward, and Pro-innovation:** NIST AI RMF speaks about being pro-innovation. In the basic context of Reward/Risk, a risk management framework should only impact the denominator. Each reward unit is more valuable (pro-innovation) when a unit of risk is reduced.
- **Systemic Societal Impact:** NIST AI RMF does not consider systemic societal impact. They are looking at the direct impact and not the long-term systemic impact on society, including AI systems impacting democratic values (e.g., voting). ForHumanity has built detailed guidance regarding the Systematic Societal Impact Assessment for examining risks that impact individuals, communities, and society.

- **Integration with COSO:** NIST AI RMF mentions that organizations need to establish and maintain appropriate accountability mechanisms for the organizational integration of risk. However, it is non-explicit about how the integration work (ForHumanity recommends abiding by the COSO approach to integration).

Link:
https://www.nist.gov/news-events/news/2022/08/nist-seeks-comments-ai-risk-management-framework-guidance-workshop-date-set

ForHumanity
Executive Director - Ryan Carrier
https://forhumanity.center