FROM UNDERSTANDING TO OPERATIONALIZATION

DIVERSE INPUTS AND MULTI-STAKEHOLDER FEEDBACK

February 2022
RYAN CARRIER

https://forhumanity.center/

This article will guide the reader through what is meant by "Diverse Inputs and Multi Stakeholder Feedback (DI & MSF) in the context of Independent Audit of AI Systems and ForHumanity's audit criteria on UK GDPR, the EU Artificial Intelligence Act and our own Risk Management Framework. As a critical piece of embedded human agency, DI & MSF increases risk awareness and subsequent mitigations of potential negative impacts to humans and our humanity(rights, freedoms, equality, dignity)

“Diverse inputs & multi-stakeholder feedback is the process of injecting human agency into AI, Algorithmic and autonomous systems to maximize risk mitigation in socio-technical context. “

Bias and unfairness in algorithms are a hot topic in AI Ethics and machine learning circles and rightfully so. There have been numerous examples of bias and unfairness infecting artificial intelligence and machine learning models ranging from racist chatbots to facial recognition systems that work poorly on women, especially those of color. High-profile failures in artificial intelligence (AI) and autonomous systems exposed the risk embedded in the systems we are meant to trust.

In response, the community has responded with some solutions to mitigate bias and improve fairness. A few examples include, Bias mitigation in Data Sets, Why Fairness Cannot Be Automated and Bug Bounties for Algorithmic Harms Implementation of these types of solutions will improve our algorithmic processes and enhance governance and oversight, which are important building blocks towards trustworthy systems. However, even amongst these thoughtful proposals, it feels as if something is missing - an unquantified risk.

Regarding systems that impact an increasing portion of our lives, where are “we” in the equation? Do these systems truly increase human agency? Or are they extensions of corporate or government interests accelerating a trend of making all of our interactions transactional and without human contact. How can we mitigate these risks and ensure that these systems are built for us in a trustworthy manner?

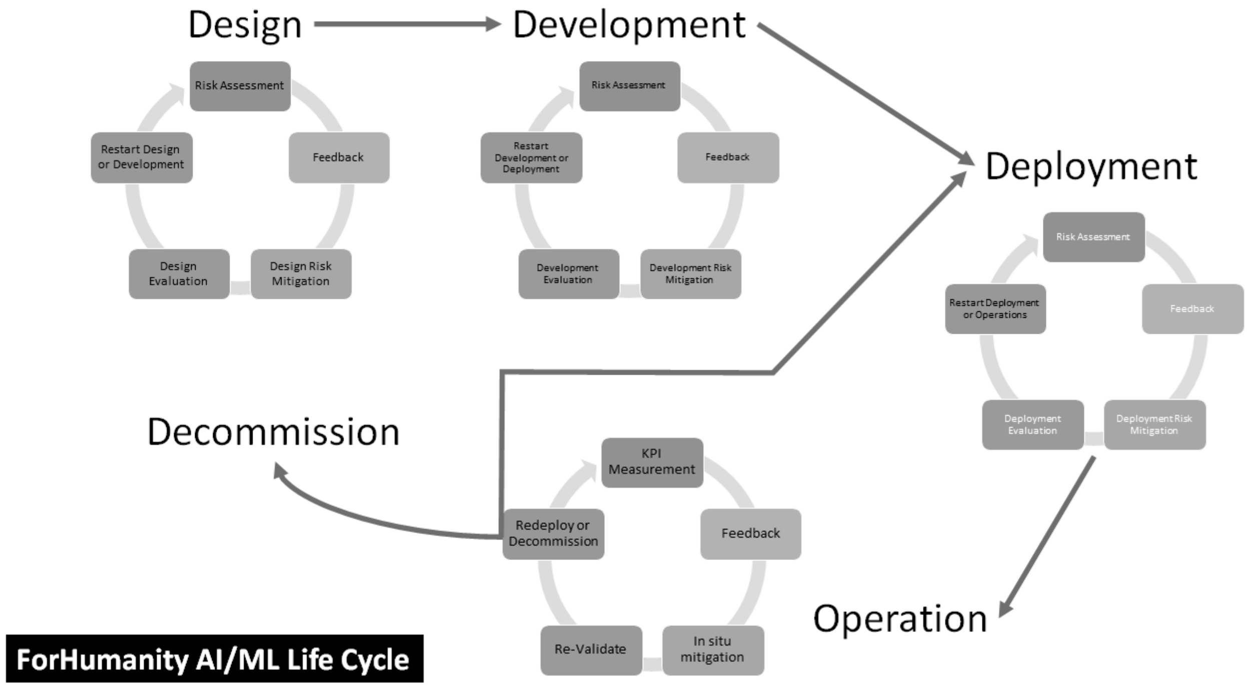
To answer these questions, we must explore the risks and harms associated with AI, algorithmic and autonomous systems (hereafter AAA Systems) to understand the ways in which these systems can manifest risk.

Risk exists in every endeavor in our lives, so finding risks embedded in our AIs should not be a surprise. To better understand the presences of risk in AAA Systems, let's examine one version of their system lifecycle.

We can see from our lifecycle image - feedback, risk assessment, and risk mitigation are ever present in the design, development, and deployment phases. This is because the sources of risk can occur at any point in the equation. This diagram only depicts the operational phases of design, development, deployment and decommissioning phase. These systems are also filled with Personal and Non-Personal Data that has consequences to humans. Necessitating unique risk treatment across Data Quality, Informational Quality and Pipeline Quality (technical terms) that allow for precision risk treatment in conjunction with the operational risk mitigations.

It is critical to note that risk can manifest across the entire lifecycle - in the data itself (embedded bias), collecting data, pre-processing, training/testing/validation, architectural input, pipeline data, concept drift, Human-in/on-the-loop, outcome pipeline and post-market monitoring. The sources of risk in AAA Systems are myriad, begging the questions of where and how we mitigate these risks. Before tackling that question, let's examine an industry built exclusively to manage risk in order to see if we can borrow some tricks of the trade.

The insurance industry exists for the management of downside risk on behalf of policy holders, and thus, many people interact with them regularly. However, few people know "how" these risk management experts mitigate their risks.



Beyond internal policies, contractual stipulations, claim requirements, asset/liability matching, premium-setting, economies-of-scale, multi-product offerings, insurance companies still often have sufficient residual risk that they turn to an ultimate backstop - a second industry - Reinsurance. Reinsurers are insurance companies for insurance companies, the world's ultimate risk management firms (e.g. Swiss Re, Munich Re, Berkshire Hathaway). These firms specialize in risks that are difficult to quantify and especially difficult to manage; the sector has a capital value of c.$660bn annually and collects $4+ trillion dollars of premiums per year as a payment to insure the world's most difficult risks.

Beyond intelligent models, astute assessment, increasing safety protocols, education, oversight, governance, the last line of defense for every insurance company and especially their insurers, the reinsurance companies, is diversification. If a portfolio of risk is sufficiently diverse, then should one, or two or even three unlikely negative events occur all-at-once, a diversified insurer will not be bankrupt. As a result, reinsurance companies build vast portfolios of risk, engage in active mitigation, and continually seek hedges to guard against negative outcomes. Make no mistake, true diversification is difficult to achieve, but ultimately, if accomplished, diversification is the ultimate risk management tool. ForHumanity knows that diversification is our best defense against risk attributable to AAA Systems

Consider the following risk scenario, in 2016, Microsoft experimented with the development of an AI-driven chatbot called Tay. Tay was supposed to be the digital equivalent of a female teenager that would learn from interacting with people on Twitter.

> "More diversification reduces the size of each individual negative outcome from each potential risk in the system "

But the account was taken down after less than a day when it learned from interacting to be a Hitler-loving, feminist-bashing troll. There is no chance that the goal of the effort was to create a racist chatbot. Moreover, there is no way that Microsoft's risk management team would want this model to go into the market to fail less than 24 hours later, so what happened?

Connecting back to our earlier thought, we would argue that the portfolio of risk inputs and risk assessment was simply not diverse enough, or, said differently, at a minimum Microsoft needed more diversification in the risk process to identify, assess and treat this risk. How many more people needed to be in the room before one of them wondered "can this be corrupted through online attack"? Was it one more person, or was it one more person with a unique perspective? Was there a group that wasn't contacted inside Microsoft who might have said "wait, let's test this"? How many more viewpoints and perspectives did they need in order to reign in this model and sufficiently manage the risks before it came to market?

There is no hard and fast measure for those questions, but there is a universal truth to diversification - more is better. It may have decreased incremental value, but more diversification reduces the size of each individual negative outcome from each potential risk in the system. Viewed specifically and solely from a risk perspective, more diversification is better. But more of what?

Diversification - more of what? "Diverse Inputs"

This leads us to the next question - more of what? What diversification do we seek for AAA Systems to achieve the final diversification that Reinsurance achieves for the insurance industry. Returning to the Microsoft example above, it is apparent that there were insufficient considerations for risk. In other words, the risk of "data poisoning" was insufficiently identified and thus insufficiently mitigated. The argument follows for increased diversification of assessors and those providing risk inputs.

Let's start with a two definitions:

Diverse - composed of distinct or unlike elements or qualities
And
Diversification - the act or process of diversifying something or of becoming diversified : an increase in the variety or diversity of something

What sort of diversification helps to identify, assess and mitigate risk? It begins with diversity of thought and lived experience. Artists see the world differently, musicians see the world differently, tech ethicists see the world differently, data scientists, sociologists and on and on and on - far beyond the design and development teams driven by engineers and the trained thought processes they often share in common.

"Diversity is not about gender or color. Organizations need to ask, "how diverse?" to examine the extent of diversity they have"

What about Multi Stakeholder Feedback?

Historically, risk assessment has been largely outcomes focused and done by design, development and, occasionally, risk management teams with an inward focus on organizational risk. This siloed and stilted risk assessment process focusing inward is ill-prepared for risk assessment of socio-technical systems where the human (Personal Data) and our humanity (rights, freedoms, equality, dignity) are integral within the processes/systems.

One of the great challenges facing AI actors in regard to managing risk is a cultural one. The US technology sector often approaches new challenges with a "move fast and break things" strategy that focuses on quick, short-term gains and profit rather than contemplating long-term risk and legal/regulatory compliance. As this sector advances the development and deployment of socio-technical products and solutions, the "things being broken" are increasingly people, interpersonal relationships, communities and societal interactions.

If we restrict our risk assessment and risk input process to designers, developers or even additional assessors within the company building a system (as we have been doing), then we remain clouded and overwhelmed by "corporate" thinking, which is hardly sufficiently diverse and certainly clouded by corporate incentive structures

"ForHumanity advocates for a risk management framework that is omni-directional and multivariate."

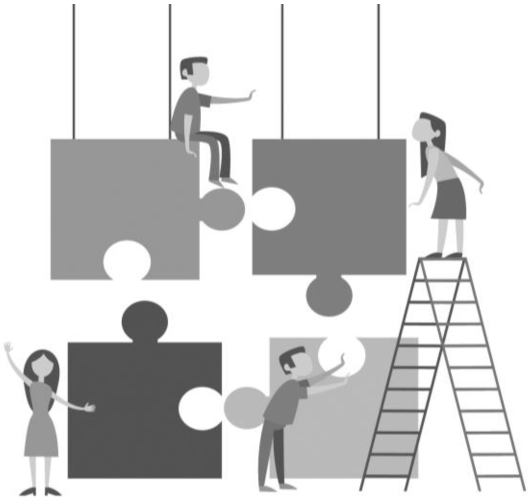
ForHumanity advocates for a risk management framework that is omni-directional and multivariate. Multivariate in that the risk framework considers corporate risk (damage to employees, business reputation and shareholder wealth), risk to humans (damage to users/clients/prospects and unwitting participants), societal risk (damages to systems, groups, communities, markets and collectives) and environmental risks (damages to nature and sustainability considerations).

The word "Multi Stakeholder" is a 360-degree perspective of the system and risks being assessed. Providers of data, providers of inputs, providers of systems (software and hardware), providers of networks, data processors, partners, clients, prospects and most importantly the human users, clients and prospects from many different backgrounds, cultures and user experiences.

Collectively, DI & MSF describes the pool of risk assessors that ForHumanity requires throughout our comprehensive risk management framework - embedded in Independent Audit of AI Systems. This group satisfies risk assessments such as the Algorithm Risk Assessment, security and human-in/on-the-loop assessments in the Testing & Evaluation Committee At-Risk Report.

Risk Thinking and Education

There is a tension and tradeoff that exists in collecting Diverse Inputs and Multi Stakeholder Feedback. The tension exists at the intersection of training - the understanding and communication of words and ideas regarding AAA systems and the language of risk - juxtaposed to the very nature of their input value - their diverse thinking and lived experiences.



Enabling people to express their concerns, risk ideas and perspectives on negative impacts while allowing them to think differently. Providing risk inputs and expressions of potential negative impacts will often require a certain level of training to understand what they are assessing. Accommodations will have to be made in order to overcome the following challenges in order to provide meaningful input:

- Language and terminology of AI, algorithmic and autonomous system risk inputs
- Barriers to entry associated with accessibility (data inputs and collection)
- Barriers to entry from natural language and technology
- Understanding of the model itself, including scope, nature, context and purpose
- Understanding necessity & proportionality
- Understanding potentially "foreign" Codes of Ethics, Data Ethics and "shared moral frameworks"
- Considering the measurement and metrics of accuracy, validity and reliability
- Understanding outcomes, impacts
- Understanding of AAA systems terminology and basic design.

Stated another way, enabling DI & MSF to be useful and productive has consequences to the very nature of their diversity because the training teaches them "a way of thinking" and alters their "lived experience". The training has to happen, but solution providers and educators need to reinforce the tools of understanding, education and communication combined enabling their manner of diverse thinking and recollection and application to their differing backgrounds.

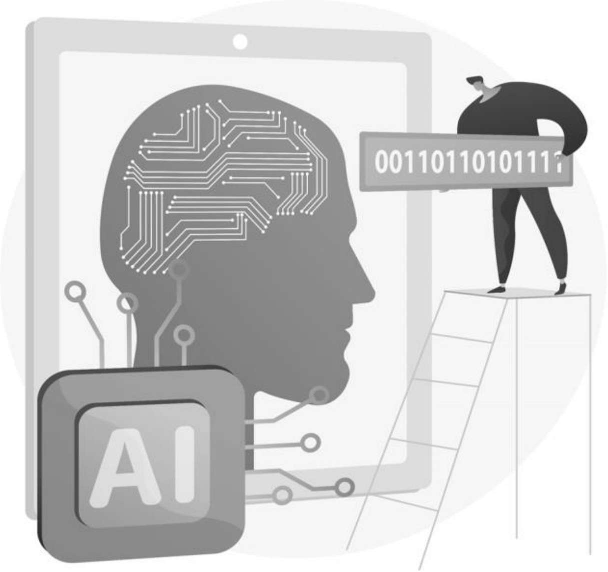
Further training can help assessors avoid some common mistakes when considering concerns, negative outcomes and risk inputs such as:

- Piecemeal and siloed targeting of AI related risks
- Overfocus on "headline grabbing" concerns
- Identification of emerging and new adversarial attack vectors
- Embedded, cognitive and technology barrier bias
- Technical and communications hyperconnectivity, exposing us to rapid proliferation of global risks, requiring creative evolution of risk management approaches
- Overfocus on technical or data quality related risks, versus people-centered outcomes
- Overfocus on success criteria related to immediate intent and functional requirements, versus medium and longer-term risks associated with scope and sharing creep.

Assessors going through training and when providing feedback will be providing a vital service to the organizations that call on them to share their lived experience. Although some may be able to provide their time and knowledge on a voluntary basis, it should be standard practice for people on low or no wage to be remunerated for their time.

Applying Diverse Inputs and Multi Stakeholder Feedback

Now that we have a pool of risk assessors, trained and educated on their roles and duties, where are they put to work? When, where and how become our critical next questions. The following sections will specify the process (ForHumanity's Body of Knowledge provides a template for process guidance) that may be tailored for specific instances of assessment. The Algorithm Risk Assessment (ARA) - a living document - is the repository for Diverse Inputs and Multi Stakeholder Feedback. These inputs join other risk inputs from the internal risk assessment process.



As discussed above, DI & MSF should be considered throughout the lifecycle of the systems (Design, Development and Deployment)

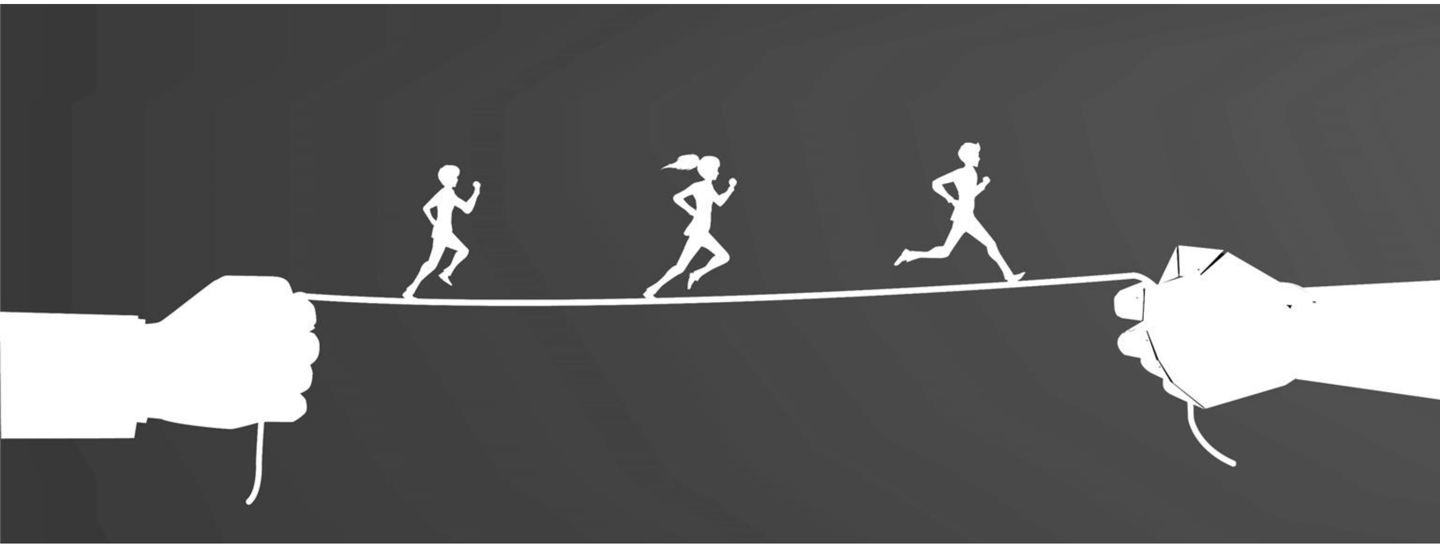
After deployment, continuous monitoring, post market monitoring and Adverse Event Tracking Systems take over and encompass most of the risk assessment, only in the occasion of a refresh of the whole systems would we reconvene Diverse Inputs and Multi Stakeholder Feedback.

Questions about timing, costs and overkill might arise from regular inclusion of Diverse Inputs and Multi Stakeholder Feedback, however, this work is designed to counteract massive problems in the existing process from sunk cost bias, confirmation bias, insufficient risk assessment, lack of accessibility, and countless other exposures to risk that could be mitigate with a more robust process.

Getting Started

The design and development team must establish the scope, nature, context, purpose including a lawful basis for the system. Necessity Assessments and a Proportionality Study present the first opportunity for risk input.

The next step belongs to the Ethics Committee and their application of the Code of Ethics and diversity policy to determine if/when the pool of Diverse Inputs and Multi Stakeholder Feedback assessors are sufficient. Taking account of availability of assessors, the Ethics Committee will determine the pool allowing the assessment process to begin. Once the pool of assessors has been established, the design and development team need to share, explain and describe the system to the assessors.



<u>ForHumanity audit rule</u>

To ensure representativeness, the Ethics Committee shall be responsible for determining the makeup and completeness of the diverse inputs and multi stakeholder assessors as well as the feedback.

The assessor’s task will be to listen, understand and apply the information to their assessment process. Risk Taxonomy provides the guideposts for the assessors, where to look? The ForHumanity **Risk Taxonomy** starts with Risk Categories & AI principles

Traditional Risk Categories	EU High Level Expert Group AI principles	FH AI Ethics Principles
Strategic Financial Reputational Operational ESG Business	Human agency and oversight Technical robustness and safety Privacy and data governance Transparency Diversity, Non-discrimination, and fairness Societal and environmental well being Accountability	Human Centric Ethical Fair Actionable Operational Accountable Auditable Certain Transparent
These are typical organizational risk categories, which are likely static across specialist operational risk areas e.g. Strategic / Financial IT risk, Strategic / Financial Marketing risk. Requires evolution to incorporate risk categories associated with impacts to individuals, society, and the environment in addition to the organization.	EU High Level Expert Group AI Principles is one generally well accepted list of ethics-focused principles is included below. It can be mapped to AI control capabilities / domains and impact types. Then to criteria required to effectively manage socio-technical systems risk.	Principles that are applied within ForHumanity for minimizing downside risks to people, society and environment. **Risk universe in AAA systems shall consider risks that has a potential to impact people, people groups, society, and environment.**

Trained on the risk taxonomy, the assessors can apply their expertise and perspective to the system and identify concerns, negative impacts and risk inputs in the system. Concerns, negative impacts and risk inputs must be measured in two primary vectors - severity and likelihood. Assessors can use the scale below to subjectively rank each identified risk relatively and consistently. It is the relativity that is critical, not the exact level as concern, negative impact and risk is perceived by people differently.

Guidance on impact understanding

1. Life/Death decisions
2. Physical or mental harm
3. Loss of Rights/Freedoms
4. Restrictions of Rights/Freedoms
5. Restriction to the Access of Goods and Services
6. Discriminatory outcomes
7. Unfair outcomes
8. Identify theft/loss of identity
9. Disclosure of personal information
10. Damage to reputation
11. Monetary loss
12. Harassment and increased undesirable exposure
13. Annoyance and hassle
14. Minor repairs
15. Petty disturbance

The exact hierarchy is not important, as different individuals/stakeholders will value some of those risk elements slightly differently. This hierarchy is meant to provide guidelines and guardrails for the analysis and aggregation of risk impact. The most important analysis required at this phase is the identification of these risks, who may be impacted, and if those impacts are different to one stakeholder over another, especially if those differences occur across Protected Category Variables, like gender or race.

Analyzing likelihood and severity of consequences

Some assessors will want to assess the concern, negative impact and risk input in steps, others will choose to evaluate severity at the same time as likelihood. This process can be adapted to the assessor, as long as in the end a fair and unfettered adjudication of severity and likelihood occurs. Both internal and external risk assessors must consider the severity of risk, as well as the likelihood of those risks..

Scale of Likelihood:

1. Persistent (greater than 95% chance)
2. Very likely (70%-95%)
3. Likely (50%-69%)
4. Possible (30%-49%)
5. Unlikely (1%-290%)
   - Mostly unlikely (16%-29%)
   - Very unlikely (5%-15%)
   - Remote (1% to 4% chance)
   - Rare or Unusual (less than 1% chance)

*Note, this analysis should pay more attention to less likely risks with enormous impacts, because these risks, often known as Black Swan events, are often undervalued in their likelihood and subsequent severity that can have devastating effects. Therefore, when considering the risk to humans, greater attention to the estimation of unlikely events with significant impact will enhance risk mitigation. In these moments, the organization has a critical opportunity for diverse inputs, from people who are not weighed down by sunk-cost bias and therefore are freer to evaluate catastrophic risk. Catastrophic risk that if originally considered by designers and developers from the outset, might have hindered development in the first place.

The system may have many possible risks to people, but if the chance of those risks occurring is remote, then the risk management team must weigh that composite risk accordingly. For example, if there is a risk of death from a system, but the likelihood of that happening is 1 in a trillion chances, then the composite of that risk is lower than if the likelihood of death is 1 in 1000. The point is that there are no systems that exist without risk.

Before Diverse Inputs and Multi Stakeholder assessors are finished, allow them the chance to provide you with potential mitigations. As the risk management process moves forward and risk inputs are turned over to the Algorithmic Risk Committee for risk treatment, allow your DI & MSF assessor to offer their suggestions for risk treatment. The endgame is risk management. If the system is useful and beneficial that value will be enhanced with maximized risk mitigation regardless of the source. It is rare to eliminate all risks, therefore any opportunity to identify risk inputs and risk treatments should be taken.

This process is not a silver bullet or complete panacea. Diverse inputs and multi stakeholder feedback will lead to surprises and initial reactions of incredulity. It is imperative that the persons conducting the ARA report record fairly and impassively the collection of perceived risks. This phase of collecting risk inputs is not about solving or mitigation, instead this step in the assessment is about the identification of risks to humans, communities, nature, and society-at-large. This process is highly subjective, but the pre-elimination of risk inputs is a disservice to the process and a risk assessment itself.

Another consequence of the process is the subjectivity of severity and likelihood. Most of us can agree on the difference of importance between a life/death decision and a petty disturbance, however, it is equally reasonable for two people to disagree about the relative importance of the difference between monetary loss and identity loss or the difference between the damage to a reputation and discriminatory outcomes. There is no right/wrong to be imposed upon the severity of a risk. Likelihood is an estimate, not a fact. The process is designed to meet assessors where they are and enable them to express themselves comfortably and in a manner that makes sense to them. The resultant subjectivity will require thoughtful consideration when aggregated and conclusions are reached. Accumulation of inputs from assessors will smooth out outliers and personal bias.

In the analysis of risk and impact to stakeholders, ForHumanity calls to produce a specific report, an Algorithmic Risk Assessment (ARA) designed to resolve 2 key issues of risk:

1. The riskiness of the processing (Systemic/High/Go Ahead/Low)
2. The collection of risk inputs and risk treatment (not accounted for in other aspects of the ForHumanity AAA Systems Risk Management Framework)

Special considerations for external Diverse Inputs and Multi Stakeholder Feedback

The identification of risk is a personal thing. It can be a feeling, an impression, a guess or a well-considered calculation to which right and wrong may not always be assigned. Furthermore, perspective is crucial. One person's risk is another person's opportunity and vice versa.

Perfection is unattainable. Regardless of the number of skilled risk assessors, with diverse backgrounds, there is no certainty that the assessment will identify, correctly prioritize and treat all risk. This fact is not an excuse for limited action, but instead a cautionary limitation to those operating, conducting or relying upon the ARA to manage expectations. The Algorithmic Risk Committee in conjunction with the Ethics Committee will have to determine when an ARA process has achieved enough risk assessors.

The Ethics Committee shall also comprehensively consider the risk to individuals, communities, nature, and society-at-large before proceeding. Some jurisdictions and some systems may be required by law to seek approval or review from the local authority prior to proceeding.

Lastly, Diverse Inputs and Multi Stakeholder Feedback cannot identify every risk perfectly correctly, therefore, this process is just one of the important components and risk managers should be ever vigilant for emerging risks and "unknown unknowns". It is a difficult process to imagine the range of outcomes that may proceed from a system, and it would be a rare, and likely simple system where all possible negative outcomes can be imagined in advance and forever.

Diverse Inputs and Multi Stakeholder Feedback is necessitated by the ubiquitous advancement of socio-technical systems - driven by the very nature of these systems and their multidisciplinary applications and risk. AI, algorithmic and autonomous system's comprehensive inclusion of the human through Personal Data coupled with their increasing impacts to our humanity (rights, freedoms, equality, dignity) demand a greater vigilance of risk and negative impacts - DI & MSF delivers one vital measure of the needed vigilance.

FORHUMANITY

Credits: Many thanks are needed on this paper starting with the whole of the ForHumanity community. Special thanks go to Sundar Narayanan, Sue Turner, Hema Lakkaraju and Ashley Coffey.

Images: Freepik