Principles of AI Risk Measurement – HARM Model

P Rajagopal Tampi ORCID :009-006-6364-3275 Email: Rajagopal.tampi@valuemoves.co.in
DOI 10.5281/zenodo.21058981

Abstract

The paper presents the principles of AI risk measurement using 'HARM' which is the first holistic Human-centric Ai Risk decision Model for measuring systemic AI risk. The complexity of AI, its myriad impact surfaces and intertwined impacts have hitherto prevented measurement of human centric AI risk. Current risk models and frameworks are narrowly focused on specific domains like medicine and insurance. Risk is introduced not only by AI training but also by other technology system components and their architecture. An approach consisting of use-case based holistic risk score and discrete impact area-specific legislation control are essential for effective risk management.

This paper draws from and is aligned with OECD AI principles [1] , NIST [2] , United Nations "Governing AI for Humanity" [3]. HARM is based on the intent to **build human-centric AI and promote AI for public good**. It incorporates Responsible AI, Trustworthy AI and Explainable AI and AI for Good principles.

The HARM model is intended to be a baseline for simple, quick, reliable and holistic calculation of human-centric risk at a system level for AI use cases across domains. The clear segregation of risk dimensions, the lookup tables for selection of risk levels, clarity between use-case and systemic risks and simplicity of risk calculation make HARM easy to understand, adopt and adapt. At a time when AI is growing freely because of its complexity, this paper provides a definitive way to manage AI by measuring its risk. Measurement enables the impacts of a technology to be gauged and guided. It allows companies to measure the risk of AI use cases during design phase of the development project for their product safety, governments to specify acceptable risk levels for AI use cases from vendors and standards bodies to define global standards. It is an opportunity to scientifically reduce AI risk levels. It can bring about a common understanding of AI risk and a framework to manage it. The paper defines "Beyond Human Cognition" and cautions of irreversibility when AI achieves the cognitive capability of self-development.

Introduction

The increasing complexity of a digital world calls for a depth of understanding which permits simplification for identifying and categorising different AI risk impact layers from very high to low impact on humans. **A human-centric AI risk decision model can help bring coherence, consistency and common understanding for developing AI products safely, setting global standards for AI risk and drafting AI governance policies. This paper proposes the H**uman-centric **A**i **R**isk decision **M**odel **(HARM©), a simple, quick, reliable and holistic risk decision**

model for the purpose. The framework methodically divides the universe of AI risks and impact surfaces into different risk management dimensions based on its impact surfaces and types. This enables logical, justifiable and consistent risk weights to be applied. Risk is introduced not only by AI training but also by other technology system components and their architecture. HARM measures AI risk at the AI use-case level and at the AI system technologies and design level. The model provides holistic, baseline values for AI production systems which enables remediation by technology vendors to bring risk levels within legally permissible caps imposed by nations. The baseline provided by HARM can be fine-tuned for in-depth domain specific risk measurement needs.

HARM uses twelve risk dimensions to cover holistic risk analysis and calculation. It calculates "inherent risk" and "applied risk" for any AI use case (Definitions). The clear segregation of risk dimensions, the lookup tables for selection of risk levels, clarity between use-case and systemic risks and simplicity of risk calculation make HARM easy to understand, adopt and adapt.

This model can be used by companies at the design stage of developing AI systems, Governments to weigh baseline risk levels during policy making. Customers can select the amount of risk they wish to be exposed to from a vendor's AI product.

Artificial Intelligence risks attack us across many dimensions. It affects the lives of people across social, ethical, human rights, political, technical, environmental, security, finances, jobs, homes and many others. Any study of AI risks therefore should cover impacts across all these domains to achieve holistic measurement of AI risk.

The AI stakeholders are technology companies, government agencies, regulating bodies, standards bodies, scientific organizations, international coordination and leadership institutions like the United Nations, ITU, GPAI etc., professionals and representative groups from society. The critical role AI product creator companies play in the human-centric analysis and design of AI applications determines the product risk level of AI outputs on people. These companies should be held accountable by governments for the harm they cause due to irresponsible AI product design.

Definitions

AI Environmental/Functional Dimensions: These are the five factors identified in OECD paper Base Document No 323, Feb 2023. which are impacted by the risks associated with using the AI. They are "People and Planet", "Economic", "Data and input", AI Model" and Task and Output".

Applied Risk: When additional technological and system design factors are applied to the "inherent risk" to obtain a production AI system, there are other risk factors which also become applicable. Applied risk is the risk posed when the use case has been deployed in production. Both applied and inherent risk values are calculated by the HARM model.

Beyond Human Cognition (BHC): This happens because of two components a) capabilities our brains do not possess eg. pattern recognition and b) "brain resource insufficiency". An example: as computing power increases humans will not be able to keep up with the pace of AI. This will

result in loss of control. Control loss will become absolute when the AI loop is several orders of magnitude faster than what humans are capable of. This maybe thought of as **"computing power of the AI beyond human cognition."** Another example of this could be the limitations imposed by our human biological memory size.

HARM Risk Dimensions: These are the four HARM risk dimensions of Functional, Control, Deployment and Architecture risks of AI's when it is deployed in production. Functional risk is evaluated using nine dimensions in Table 2.

HARM Risk classes and weights: The risk classification and corresponding weight scale with Very High risk being represented as 6 and low risk as 1 is shown in Table 1.

Inherent AI Risk: It is the risk that emanates purely due to the impact of AI training algorithms and processes on the use case which gets applied on the five Environmental dimensions. This represents the functional risk component of a specific AI use case.

HARM AI Risk dimensions

Designing an AI system has gone far beyond choosing a model. To manage AI risk at the design phase, it is necessary to decide where and how the model is located, learns and accesses data, whether it is mobile, can be accessed by us, whether we can observe and understand what it does and monitor its intents, track its actions, have appropriate level of sustainable control and communications with it and many other factors. Regrettably, current development practices adopted by technology product creator companies do not address these engineering principles resulting in the deluge of AI risks for humans.

A significant concern with AI is the exponentially increasing computing power, speed and data handling capabilities which even today far surpass the cognitive powers of humans in many areas.

The AI **"Environmental /Functional Dimensions"** which are directly impacted by AI risk are the following five, "People and Planet", "Economic", "Data and input", AI Model", Task and Output" From these high-level concepts, the **Human-Centric Risk Level Classification (HRLC)** [4] framework has been derived to identify and classify all possible human-centric risks of an AI use-case. AI systemic risk complexity demands that holistic risk from AI systems be viewed along the following additional dimensions. In addition to Functional risk / dimension explained above, other risks are also caused by the technology implementation. These are:

1) Control Risk of the AI
2) Deployment Risk of the AI
3) Architecture Risk of the AI

While functional risk is fixed being determined by the specific use-case under nine dimensions, the other risk dimensions are human determined depending on technology and system design choices made at design phase and therefore are flexible and can change.

Functional AI Risk: The functional risk categorization is explained in the **Human-Centric Risk Level Classification (HRLC)** model below. It deals with the holistic human-centric risk arising purely from AI training on the use case. It is the '**inherent risk**' of the use case.

Control Risk: One of the risk domains is the level of control that humans lose to AI when it replaces human thoughts, decisions and actions. The Loss of control that happens due to AI based automation is discussed in Chapter 2, Applied Human- Centric AI[4]. Based on the amount of control loss, there are three major classes of AI namely Independent Artificial Intelligence (IAI), Hybrid Artificial Intelligence (HBAI) and Human controlled Artificial Intelligence (HAI).

Deployment Risk: AI systems can introduce risk due to their type of production deployment. Accordingly, three deployment risk categories, Type of deployment group, Real-time group and Discrete group are identified. The deployment categories are listed in Table 4 below.

Architecture Risk: This risk arises from data handling, system design, technology architecture selected and applied for design and development of the AI system. There are many technology architectures which are listed in Table 5 below.

HARM Risk weights

To make the HARM model measurable and effective, the following risk weights are assigned to different classes of risks as per table 1 below. These risk weights represent the levels of harm to the users, society and the planet inflicted by the AI system.

Risk weight	Risk level classifications	Explanation
6	Very High Risk	Life and health preservation. Humans will die, be eliminated or suffer physical harm.
5	High Risk	Individual human living essentials, human rights, fundamental rights guaranteed by constitutions, and choices (free will). Global and societal living essentials defined in table 2.
4	High Medium Risk	Fairness factors applicable to individuals and society generally not covered by law but desirable for individual fairness and societal harmony. Factors affecting the general economy. Global sustainability essentials defined in table 2.
3	Medium Risk	Other impacts lesser in severity than above.
2	Medium Low Risk	Minor risks having no lasting effect.
1	Low Risk	No significant effect of the risk.

Table 1: HARM Model Risk weights and classification table

The risk weights are chosen so that at lower risks levels the accuracy of selection is double. Above High accuracy classification (risk weight 5), the accuracy becomes less meaningful as the risk is too high. In the High and above range, the use case is avoidable, or the risk needs to be reduced in some manner if it is to be automated using AI. This step change in risk measurement accuracy is illustrated in the figure 1 below.

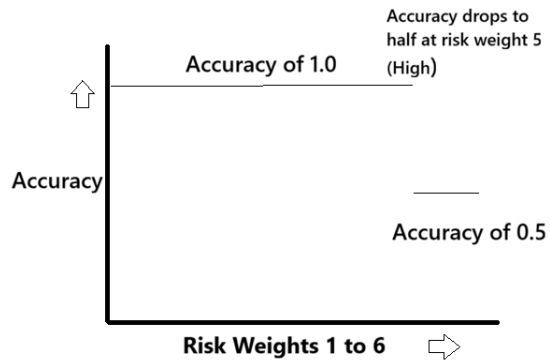

Fig 1 – Accuracy versus risk levels plot

Functional Risk

Since the aim is to develop and deploy human-centric AI, this risk classification is called HRLC (Human-Centric Risk Level Classification) see HRLC Table 2 below. Human-Centric risk from AI is divided into nine functional risk categories as shown in the table.

The functional classification is decided based on an analysis of the use case requirements and the training data set. Forecasting the possible exact risk impacts on human beings, the society, planet and sustainability are guided by the categories and their explanation in the HRLC table.

There could be multiple impacts of the use case in different HRLC categories. In case there are multiple risk impacts, the highest risk impact value should be taken as the functional risk value for the calculation of applied risk.

Indicative questions to be asked while determining the functional risk classification are a) “Is there an impact on human life or loss of limb?” and b) “How does this impact humans, society, the planet, economy, sustainability and other dimensions?”.

If the risk does not apply to the top five (LH, ILE,SLE,SUS,FE) categories in the HRLC table, then the “Other” HRLC category should be used and weight chosen as per the severity of risk threat. “Other” category classifies risks which are medium or lower. The proper analysis and answer to these questions will provide the right pointer to functional risk classification.

Human-Centric Risk Level Classification (HRLC)	
Risk category	HARM Risk weights and Explanations
Life and Health (LH)	HARM RISK WEIGHT 6: VERY HIGH RISK. AI is Life and limb threatening. Any application which can result in a life and limb threatening situation. This may be due to autonomous weapons, wars and instability or even short-term life-threatening natural calamities. It may also be due to medical applications, of AI in medical treatment, procedures and surgeries or any other life-threatening area such as AI flying an aircraft which can result in a crash due to a hallucination or a bug. AI in any scenario that could result in loss of life and limb falls in this category.

Individual Living Essentials (ILE) Human Rights and Regulatory	HARM RISK WEIGHT 5: HIGH RISK. These risks concern jobs, home, security, arrest, human rights, fundamental rights etc of individuals. This category comprises those areas which threaten the living necessities of human beings involving food, clothes, house, security and livelihood and other such needs. Includes violation of fundamental rights granted by nations like the right to work, equality and the right to own property, data rights, data privacy etc. Includes personal finances, financial instruments like credit cards, financial transactions, stocks and share trading applications which can cause direct financial loss for people impacting their ability to earn and live. Education falls in this category as it provides the qualifications for jobs and earning a livelihood. Personal property and assets also fall into this category.
Loss of control (LC)	HARM RISK WEIGHT 5: HIGH RISK This classification deals with loss of human control to the AI system. The AI application being considered is likely to take over certain human functions/decisions which may be detrimental to human or group interest or not as per the individual's or group's choice. Hence it constitutes a violation of human freedom of choice. This category is applicable whenever there is a doubt whether we want to allow the AI system to take a decision and/or perform the task in lieu of human decision and intervention. Different people may also have their own differing preferred choices on the same AI use case. Choosing AI to automate a use case could make the person vulnerable to additional risks than doing it themselves. It is to be noted that if the loss of control results in loss of life or loss of limb, then the risk level is 6 and not 5.
Societal Living Essentials (SLE)	HARM RISK WEIGHT 5: HIGH RISK. These can be due to environmental hazards like pollution spreading hatred, creating social divisions, incitingpanic. Political, religious, communal, and social issues causing disharmony, affecting governance, manipulating elections, rule of law and disturbing the peace fall into the SLE category. Social media based escalation of tensions, spreading misinformation faced by common people, groups and societies. A hallucinating AI can unleash societal anger and unrest if it is connected to social media. United Nations and global bodies ensuring world security and peace and harmony fall into this category. These societal living essentials are generally defined by national laws.
Sustainability (SUS)	HARM RISK WEIGHT 4: MEDIUM HIGH RISK: AI system causes sustainability and environment related risks. High energy consumption by exponentially growing data centres, computing resources and its indirect effects on global climate, water and other resource shortages, destruction of nature and causes of depletion of ozone layer. AI use cases which cause large carbon footprint fall into this category.
Fairness and Economic (FE)	HARM RISK WEIGHT 4: MEDIUM HIGH RISK. AI outcomes could embody bias, ethics violations, unfairness and encourage business monopoly. General economic harm caused by AI falls under this category. Trade bodies like the WTO etc. fall into this category. Lower risk outcomes caused by the above factors which may not be covered under national or international laws.
Incomplete Data (INC)	HARM RISK WEIGHT 4: MEDIUM HIGH RISK. This refers to incomplete or missing data which cause errors in AI systems. Some data classes may get left out in the AI training data resulting in non-representation of people or communities. The result is that the people or communities' views, identity and other human rights would be completely ignored by the AI system. This classification constitutes a major part of all AI risk harms.

End-user (EU)	HARM RISK WEIGHT 4: MEDIUM RISK. Some end-users may be children; other end users may be hackers threatening AI and other important systems like banking systems. AI system design must cater to safeguarding AI systems from all known end-user category-based functionalities and checks to align with local regulations. Examples of end-users: children, criminals, forgers, misguided individuals and rogue states etc. Note: the risk level of hacker applicable to a bank hack is 5 and not 4.
Other (Negligible AI impact on humans or planet)	HARM RISK WEIGHT (3,2,1): For other types of use cases which have none or negligible impact on humans, planet, economy or the environment. The risk level to be assigned depends on the level of harm inflicted from the set of (3,2,1) i.e medium, medium-low and low risk. Eg. Business workflows, business processes, academic research etc.
Note: LC category risks are covered under Control dimension. EU risks can span many risk levels and are covered by the architecture dimension. INC data is covered under Architecture dimension.	

Table 2: HRLC risk levels Table with classifications and HARM weights

Using the standard Risk weights of HARM, the different Functional risk categories have been assigned risk weights .

Focus of Control, Deployment and Architectural dimensions

The focus of the control risk dimension is whether we can access, control and direct the AI's functions. It addresses the risk from different ways in which humans can and cannot control the AI and the dangers arising from it.

The focus of the deployment risk dimension is on execution answering questions such as how is it deployed? What is the deployment type? Is it in a container (embodied)? Is it physical or mobile? How time sensitive is it? Is it executed continuously or in chunks?

The focus of architecture risk dimension is on how risk emerges from data design, system design and technologies and architectures which are used for automation. It answers questions like which technologies and architectures are used? Is the intelligence centralised, distributed or does it use edge devices? What sort of data transfers are involved etc.? How does it handle personal data? How is the overall system designed including sensors, end physical asset activation and what are its specific capabilities.

Explanation: Mobility of the AI system brings in the need to discover/track the location of the AI to access and establish communication with it for control and other purposes. Mobility can increase risk by losing access due to the AI moving to an unknown locations and identity spoofing. Mobile AI's which are standalone and embodied physically like robots, aircrafts and cars can cause direct physical life and limb harm to people.

Basic Principles applied for risk measurement:

For AI systems risk analysis the following basic principles have been applied:

1. Access, human control and supervision are covered under Control dimension. Human supervision of systems is implied when Human on the loop (HOTL), Hybrid AI (HBAI partially manual and partially AI), HAI (manual AI batch run for pattern recognition etc.) exist. (HOTL, HBAI and HAI are explained below). This classification of human supervision of AI is applicable not only for the AI systems being studied but also to perceptive intelligence-based input and output physical sensors and asset control systems and other non-AI system components.
2. Location, maintainability and trackability of AI systems are covered under Deployment dimension.
3. The data privacy, data handling, system design and technology design aspects and input and output sensors and control systems are covered under the Architecture dimension.

Control Risk (CR)

Control risk arises from the **loss of human control which decides the level of autonomous operation of AI systems**. This category deals primarily with loss of control by humans of processes which were hitherto manual, the human supervisory capability provided by the AI system (or not), the ability to interrupt/stop the AI at desirable times for human control, human decision or human review, the computing power (MFLOPS) of the hardware used by the AI and the availability of observability and monitoring capability of the AI's internal working.

The specific principles applied for HARM Control risk measurement are:

1. Autonomous nature resulting in loss of control by humans to AI.
2. Existence (or not) of human supervision for the AI functionalities including the input and/or output sensors and control loops used for physical asset activation.
3. Ability to interrupt the AI or stop the AI when desired to prevent loss of life and limb and for other purposes (human override).
4. The amount of computing power-based threat to security that it represents (whether humans can keep up with understanding, reviewing, correcting what is happening internally in the AI system or not).
5. The availability of access, observability and monitoring.
6. Can the AI replicate itself?
7. Does the AI have self-development capability?

Note: AI self-learning risk is much lower than AI self-development capability risk. The goals of self-learning are to modify the model's parameters and weights. The intent of self-development is to modify architecture, goals, processes and even its own structure.

Loss of control becomes irreversible when AI achieves the cognitive capability of self-development because humans will lose our ability to guide, change or monitor its growth, intents

and actions. AI will then become another species growing and existing with us on earth. Due to this reason, Self-development is the highest risk level Very High (6).

Independent Artificial Intelligence (IAI)

IAI defines a free standing (it has no human controller) intelligent self-contained Artificial intelligence which operates without human supervision. The **core understanding is full autonomy of its function** once it is activated. Mobility and embodiment (containerisation) are non-mandatory requirements for IAI.

An AI agent which can live inside a network on its own, and performing functions without human intervention (or perhaps even human knowledge) would be included as an IAI. The operating control of such IAI is with the IAI itself. Its intents and motivations could be unknown. IAI's pose the highest level of risk.

Human On the Loop (HOTL) AI

In this case, the AI application runs autonomously in the normal case without human intervention. Yet there is a provision for human being to intervene at any point in time to take back control from the AI (also called human override). The risk posed by this category is in between IAI and Hybrid AI (to be explained below). Example is an aircraft pilot while landing the aircraft using automatic AI driven ILS system decides at some point that the landing is too risky and decides to pull up and come in again to do a manual landing. Another example would be autonomous real world weaponised drone swarms with HOTL where remote drone pilots can take back control of the drones. Due to the need to reduce complexity to achieve a baseline model the paper assumes that the superset HOTL (which handles human monitoring and overrides) also includes Human in the loop (HITL) function which handles affirmative human authorization before specific actions such as authorizing lethal engagement for drones which are required by law.

Hybrid Artificial Intelligence (HBAI)

An HBAI system has both types of use case implementations, those which are functionally part autonomous and part human supervised. The overall operative control of HBAI system is with humans whereas control at the sub/part-use-case level could be either with the AI or with humans. HBAI can have some functionalities performed by AI autonomously before it exits to human control for review or approval or other actions. They could employ agents for workflow orchestration. Examples of HBAI systems could be Human supervised algorithmic trading or agents used for research which involve many functional steps some of which are autonomous.

Human controlled Artificial Intelligence (HAI)

HAIs constitute classical AI (SVM, linear regression, decision trees, gradient descent etc.) which **do not have autonomous components such as agents** but use AI tools and techniques to generate insights from data and arrive at decisions which humans by themselves cannot. As the name implies the control at every functional step is with humans. This is the lowest risk category of AI since it always proceeds with human approval/review. Examples include a software

application which uses AI to predict local demand for warm clothing in the coming year for a clothing store. They are generally used as decision support systems.

How can risk arise from IAI and HBAI?

Example of IAI can be “robots” and “autonomous weapons” which once trained and activated can act by themselves without human intervention. A bug or a hallucination by the robot could result in the death or bodily harm of people. An erroneous decision by an autonomous weapon system could eliminate an entire village without an option for human intervention.

Agents communicate using direct negotiations using standardised formats like Programmable IP License (PIL), trust-less execution methods like blockchain and customized communication protocols. If we cannot monitor interaction between agents, IAIs may communicate, collaborate, collude, deceive, grab power and carry out destructive tasks by employing Machiavellian methods [5] . The reliability of techniques for monitoring the intentions of agents and inter agent negotiations needs to be strengthened. Agents could be employed by miscreants to spot zero-day vulnerabilities or break our encryption systems and the security of our networks for accessing critical resources by harnessing the power of distributed GPU compute.

The interactions of IAI and HBAI components like agent-to-agent negotiations, rules and authorisations and monitoring of such interaction remains an emerging high- risk area. An example of this is in algorithmic stock trading from which we need to learn lessons. In algorithmic trading, the risks that it creates are an illusion of liquidity which makes the market fragile. It can create flash crashes and sudden spikes, results in disadvantages for slower manual traders while helping market manipulation by algorithmic traders and increasing system and operational risk.

When AI systems complexity and large scope go beyond our cognitive capacities, we may not recognise the self-destructive possibilities of our own prompts to LLM based AIs.

Effects of Control loops: Artificial intelligence systems using sensors may control real world and virtual digital twin control loops, which are tightly coupled with physical assets, such as operating traffic lights, draw bridges, and railway crossing barriers, flying aircraft, driving cars, et cetera. These pose Very High risk (6) due to risk of Life and Limb. Such control is implicit in our definitions of HOTL and HBAI categories of the Control dimension.

NOTE: it is expected that the system design will take care of all real-world dependencies and harm arising from the coupling of closed loops to sensors and physical assets through appropriate human intervention in the HOTL and HBAI classifications of control risk. In case of only partial coverage of human supervision/ interruptibility in HBAI or HOTL use case, it should be treated as an IAI while estimating its risk level.

The risk table for Control Risk dimension is shown table 3 below: -

Control Risk Dimension	Risk profile & risk weight	Explanation
Independent AI (IAI) as autonomous AI – no	Very High - 6	IAI is autonomous possessing. Execution cannot be interrupted by humans

HOTL (without human control)		(whether the AI is embodied or has mobility does not matter)
Autonomous AI+ Mobility + Human on the Loop[%] (HOTL)	High - 5	An AI system (HOTL) where human can intervene, if required in a process which is normally run autonomously by artificial intelligence. HOTL can be implemented on a physically mobile AI container (like a car, plane) or non-container AI designed to be mobile on a network.
Autonomous AI + Human on the Loop (HOTL) - mobility	High Medium - 4	Human being intervenes if required in a process normally run autonomously by artificial intelligence. No mobility.
The AI can replicate itself?	Very High – 6	Self replication is very high risk because it could multiply and put the AI beyond the control of humans.
HBAI (non-autonomous) + mobility	High Medium - 4	An AI with human control which is mobile.
AI has recursive self-development capability	Very High - 6	This provides the AI with advanced capabilities which could be well beyond human capabilities putting us in danger which could threaten our own survival.
HBAI (non-autonomous) - mobility	Medium – 3	Fixed to location
Any AI system with no observability	Very High – 6	The AI is not observable* hence it is like a Black Box and hence very high risk.
HAI – AI used only as a Decision Support System.	Low Medium - 2	Human controlled AI systems review, approval etc done by humans. Used in Decision Support Systems which help us take better decisions.
Any AI to which we do not have access	Very High – 6	The inability to access the AI robs us of the ability to obtain control.
The resources power of the AI is beyond *human cognition*[&]	Very High – 6	We will not be able to keep up with the intentions and actions of the AI and hence unable to understand or counter it to protect ourselves.

% AI executes actions autonomously and continuously while humans supervise and can intervene if needed.

* Observability allows monitoring and analysing internal state of AI systems to understand why they produce specific outputs.

& Humans are still ahead of AI in many cognitive areas today, but we are behind AI in many cognitive abilities due to the speed of compute and limitation of resources of the human brain. At some stage, as the power of compute continues to grow e.g.: with quantum computing, the developments will be so fast that humans will not be able to keep up with AI developments, learning and growth. This is an unknown MFLOPs value at present. This is an example of AI being ” beyond human cognition”.

Note: Human supervision of systems is implied when Human on the loop (HOTL) and Hybrid AI exist. This human supervision is assumed to be applicable not only for the AI systems being studied but also to all input and output sensor and physical asset control systems. *In case HBAI and HOTL systems have only partial human supervision, the case must be treated as an IAI for risk level calculation*.

Table 3: Control Risk Table with choices and HARM weights.

Using the standard Risk weights of HARM, the different control risk categories have been assigned risk weights. Vendors and app makers can study the table to choose the appropriate control risk weight category which is used to calculate the applied risk that their AI systems pose for users to comply with local AI risk level regulations.

Deployment Risk (DR)

This dimension measures risk from harm that can arise due to the type of physical deployment of the AI system.

Specific Principles applied for deployment risk measurement are:

1. Whether it is deployed in a container (embodied) and has mobility or not. Eg: robots, robotic arms.
2. Scale of deployment – is it a platform deployment where there are many users and its impact is very large.
3. Time factors such as real-time AI where speed of processing plays significant role in the levels of impact of the AI.
4. Whether AI is a continuous process or processed in discrete batches.
5. Factors which impinge on location of the AI deployed, its trackability and its maintainability.

Type of deployment can determine how effectively we can reduce the harm which the AI can inflict on humans. Take the case of a large industrial **robot** operating heavy machinery on a factory floor. The robot could crush a human being on the factory floor since it operates independently (autonomously). Take the case of autonomous agents operating on a business network of a financial company. An agent theoretically can move to an undisclosed location/server and penetrate the company's security policy enforcements and cyber security barriers to transfer money across the network. The opportunity in this case is provided by **the inability to locate and track the AI across the network infrastructure**.

This Risk dimension is divided into three groups viz. Type of deployment group, Real Time group and Discrete/process group. As part of the Real Time group, network effects and AI hallucinations can amplify a local failure into a global failure through interconnected networks and cascade effects. These characteristics could amplify disinformation campaigns, cause flash market crashes and widespread data breaches overwhelming human oversight.

The table 4 below lists the types of deployment categories of AIs with their risk levels and explanation of the causes of the risks.

Deployment Risk Group	Type of AI deployment	Risk Level	Explanation
Type of AI deployment-container/non-container	Mobile container AI	Very high -6	Robots and independent mobile AIs particularly stand alone can inflict physical harm on us. Autonomous Weapon systems are an example of this group.
	Non-mobile container AI	High -5	Containerised AIs still pose greater risk due to additional communication risks involved since they are not directly on the network.
	AI without any of location, trackability or maintainability.	Very High -6	The lack of its location, trackability or maintainability poses very high risk
	Non-container AI	High Medium - 4	AI systems which reside in servers
Real Time Group	Streaming AI (eg: real-time, news, financial trading, automated content moderation, video/audio platforms)	High -5	Due to its real- time, continuous nature, this deployment mode is high risk because of ultra-low latency, indirect prompt injection, action hallucinations and unchecked model drift. It can offer limited checks for monitoring, safety and analysis. Larger attack surface is provided by this deployment mode.
	Platform AI and networked AI systems	High -5	Dangerous due to scale of the platform combined with the typical applications which can involve large losses and resultant harm at scale if systems fail or are compromised eg: financial, identity verification systems etc. Network effects resulting in cascading failures (explained above).
	Other real time AI deployments	High - 5	Due to real time nature of the AI deployment, the risk is high as in algorithmic stock market trading
Discrete and process Group	Batch AI	Medium-3	From the general applications which fall into this category of deployment we notice that only low priority transactions or updates are performed. Hence the risk level is medium.
	Business workflow rules AI	Medium-3	Humans set these rules and they are tweaked with human approval and the score of AI inferences decide which rules to follow hence reasonably safe to adopt.

Table 4: Deployment Risk Table with options and HARM risk weights.

Using the standard Risk weights of HARM, the different deployment categories have been assigned risk weights. Vendors and app makers can study the table to choose the appropriate deployment risk level which is used to calculate the applied risk that their AI systems pose for users to comply with local regulations.

Architectural Risk (AR)

Architecture dimension: This dimension deals with the risks arising from data processing/other data related activities, technology, system design and architecture related factors. Specifically architectural risk dimension deals with the following factors which influence the risk level of AI system:

Specific principles applied for architecture risk measurement:

1. Data security, data privacy and personal data.
2. Accuracy of inference.
3. Integrity and completeness of data used for training AI systems.
4. Technologies employed.
5. Learning methods employed influencing hallucinations, black box risks, accuracy of inferences, data security and stability of the AI system.
6. Distributed processing (multiple nodes), federated systems.
7. Input/output control technologies employed, safety mechanisms and methods to activate physical assets in the real world.
8. Perceptive intelligence, sensor driven technologies such as (optical and others) and methods.
9. Embedded systems, edge systems.

The different architecture types and their risk impact on the environment dimensions influence risk weights.

Adaptive mechanisms for Closed Loop Systems (CLS)

AI enables adaptation by analysing discrepancies between predicted and actual outcomes, refining models dynamically—often called "self-optimizing" via machine learning. In construction examples, LLMs interpret natural language updates (e.g., weather delays) to adjust constraints like task dependencies or durations without recoding. This handles uncertainties like material shortages or robot failures [6].

Adaptive CLS uses **adaptation algorithms that identify and modify system dynamics during operation**. Adaptive CLS increases risk due to instability during adaptation, unmodelled dynamics and disturbances and high computational demands that can lead to failures in real time operation [7], [8].

The AI architectures are shown in Table 5 below with the HARM risk levels mentioned against each.

Technology Architecture	Risk weights	Explanation
Expert Systems: Programming, rules based.	1	Rules set by humans hence safe
Classical AI: Linear regression, decision trees, SVM etc	3	Humans control selections of features and regression algorithms

Deep Learning: CNNs, RNNs, GNNs, Federated Learning, Transformers etc	4	Fairly proven but with many hallucinations and black box risks
Deep Reinforcement Learning (DRL) and Reinforcement Learning	5	Riskier than Deep Learning due to instability and divergence where a small error compounds until the system fails.
Foundation Models: LLMs	5	High risk due to corrupted training data, hallucinations, black box syndrome, leaking sensitive training data and data from prompts.
Agent Based Systems: LLM core with workflow orchestration	5	Emergent behaviour can introduce risks, unclear accountability, errors can be compounded by using agents.
Hybrid: perceptual intelligence, Neuro-symbolic reasoning, adaptive control and negotiation.	4	Capable of neural perception + symbolic reasoning + control + negotiation + adaptive.
Hybrid perceptual intelligence + adaptive control loops -without neuro-symbolic reasoning.	5	The absence of symbolic reasoning increases the risk of the AI
Hybrid perceptual intelligence + non-adaptive control (simple applications)	2	Used in simple systems like elevator or conveyor belt control which will not break due to 'brittleness#'.
Federated: Multiple learning nodes distributed – no personal data	3	Non-personal and summary data only removed to the cloud.
Federated: Multiple learning nodes distributed with personal data	5	Personal data transferred to cloud hence security of the individual is threatened.
Edge systems: Multiple learning nodes distributed and personal data removed to cloud server.	5	Personal data transferred to cloud hence security of the individual is threatened. Large attack surface and distribution increase systemic risk.
Edge systems: without personal data involvement.	4	Large attack surface and distribution increase systemic risk. Generally, failures are usually buffered from immediate physical harm.
Embedded AI without control loops for physical component activation	4	Risks arising from the limitations of embedded AI exist, but direct physical harm is limited.
Embedded AI with control loop for actuation of physical assets	5	Direct physical actuation and real time constraints imply single point failures can cause immediate harm.
End User Protections do not exist	5	Hardening the AI system to prevent exploitation by criminals, hackers, ensuring age and role appropriate access.
Note # brittleness means failure/locking up of the system when faced with inputs which were not anticipated by its static rules. This is like what happens to expert systems when it encounters unknown rules.		

Table 5: Architectural Risk Table with options and HARM risk weights.

Vendors and app makers can study the table to choose the appropriate architecture risk level which is used to calculate the applied risk that their AI systems pose for users to comply with local regulations.

Battlefield AI system Example

Take the example of 'drone swarming', the level of human control of the drone swarm falls under control dimension eg: Human On The Loop (HOTL). The technology architecture involved controlling the drones could be optical neuro symbolic adaptive control loops with neuro-symbolic reasoning, tight coupling, and deep reinforcement learning, CNNS, GNNs, transformer-based vision models for training the AI, sensor fusion EO/IR cameras and radar/Lidar.

Beyond Human Cognition

In practice we see AI systems moving from assisting humans to duplicating most of our capabilities. This shifts the risk from "human error" to "system error and unpredictability," because humans cannot react fast enough to prevent a disaster. Control loss will become absolute when the AI loop is several orders of magnitude faster than what humans are capable of. This has been addressed by the control risk dimension **"computing power of the AI beyond human cognition."**

BHC of AI is a critical threshold which humans and nations should recognise and accept globally. In today's context since nuclear deterrent has taken root successfully, it therefore has yielded its top-most place in global and human priority and criticality to BHC of AI systems.

Human-centric Ai Risk Decision model (HARM)

The holistic view of AI risk can be obtained from analysing it from the four dimensions as in Fig 2 below.

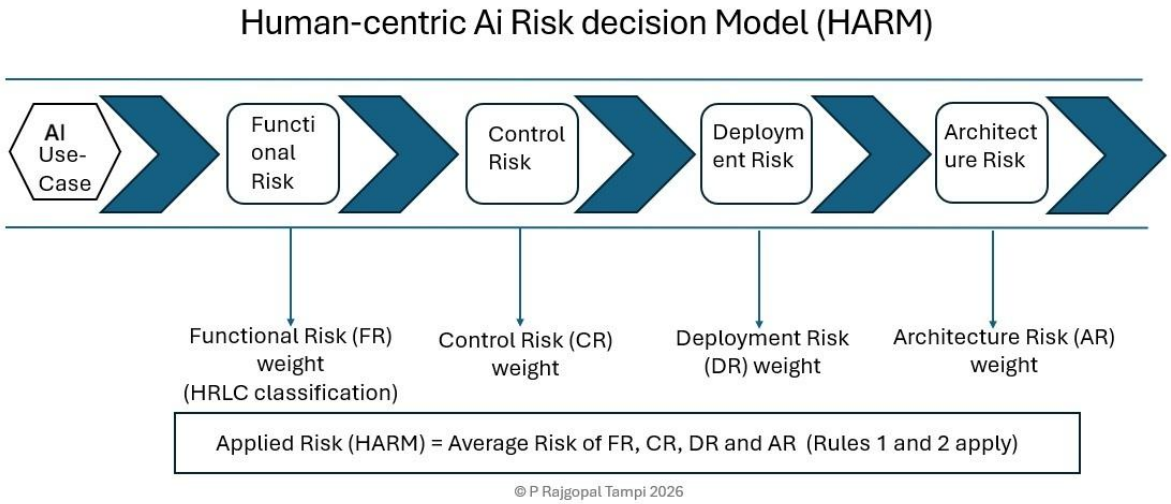

Fig 2: Human-centric Ai Risk decision Model (HARM) © P Rajagopal Tampi

Calculating Applied Risk

The HARM model operates using a 5-step analysis and calculation process as explained below: -

Step1: Calculate Functional Risk
The functional risk ascertained from the use-case measured against the HRLC table will establish the human-centric go/no-go risk resolution and provide a guideline of the level of risk management needed (Eg: need for human oversight or not. This will give pointers to the appropriate design). HARM risk weight should be assigned to functional risk as specified in the HRLC table. If there are multiple HARM risk weights applicable, then the highest risk weight should be used to represent functional risk.

Step 2: Calculate Control Risk
The Control Risk dimension where there is a choice from the ten options in Table 3 above. HARM risk weights are mentioned against each risk category in the Control Risk table. If there are multiple HARM risk weights applicable, then the highest risk weight should be used to represent control risk.

Step3: Calculate Deployment Risk (DR)
To control Deployment Risk dimension there are nine choices from Table 4. HARM risk weights are mentioned against each risk category in the Deployment Risk table. If there are multiple risk weights applicable, then the highest risk weight should be considered as the deployment risk.

Step4: Architecture Risk (AR)
RULE 1: Step 4 is not relevant if the previous three steps (Functional, Control and Deployment) risk weights are all = 6 (Very High). The applied risk in this case would be 6. The architecture dimension is not required to be considered in this special case.
OR
In case the first three risk values are not all =6 (6,6,6) then, we can choose the appropriate AI architecture required to manage the AI with the risk level which is ideally suited from choices in Table 5 above.

RULE 2: In case multiple architectural choices are used for a use case in a large system, the risks of the choices may be averaged to arrive at the architectural risk **EXCEPT if any of the risk values are = 5.** In the latter case, **consider the largest risk weight as the Architectural risk weight without averaging**. Eg 1: FR is 3, CR is 4, DR is 3 and AR includes 3 architectures viz. Deep Learning – 4, Reinforcement Learning – 5 and expert systems – 1. In this case, the AR will be 5. Eg 2: FR is 4, CR is 3, DR is 3 and AR includes 3 architectures viz. Expert systems- 1. Classical AI – 3 and Federated multiple learning nodes distributed – no personal data – 3. In this case AR will be 2.33. (Note: AR reference here is to Architecture Risk and not Applied Risk)

Step 5: After completing the above steps, an average of the four risk readings FR,CR,DR,AR will provide the "applied risk" from the AI use case being analysed. The four dimensions are allocated equal weight (i.e 25% for each dimension).

In the case where Step 4 is not relevant then the weights applicable to each of FR, CR and DR dimensions is 33.33% and the Applied risk is always = 6.

Dimensional weights of HARM and general matters

While calculating applied risk the technical/functional design details of the functional, control, deployment and architecture dimensions should be obtained from the vendor to enable correct results.

The relevance of Step 4 of risk decision calculation and its conditional application represents an adjustment that has been arrived at after testing for risk appropriateness of AI use cases using HARM across domains.

The conditional application of Architectural risk also makes sense when we consider that when the other three dimensions are all at 6, very high risk, the technology architecture has negligible or no role to play in determining the amount of human-centric risk the use case poses. Rule 2 has also been arrived at by giving greater centrality to human-centric risks during testing.

Consistency of risk determination: A fair level of AI, technology knowledge, analysis and judgement is expected from the technical person calculating the Applied Risk. This is expected to be learnt during the training for HARM use.

It is critical to be consistent when classifying AI use case risk levels in the different dimensions of risk. Example, in an AI use case where the risk of trying an accused who has not committed a murder or physical harm should be classified as 5 (human rights issue). In case the accused has committed a murder or loss of limb, then the same criminal use case would be classified as 6 on functional risk.

In civil complaints use case, the type of loss for the sufferer such as personal financial loss determines functional risk at 5. In case the use case affects the general economy then functional risk will be 4 from HRLC chart.

Conclusion

The HARM model provided baseline systemic risk level may be used in conjunction with other discrete legislations on specific areas such as data privacy, human rights etc. for effective AI risk management.

HARM may be considered for adoption by governments, standards bodies and technology companies as a common framework to measure human-centric AI risk at a baseline level and to manage AI vendors to provide safe AI products for users.

Being a simple, flexible, quick, reliable and holistic risk baseline even for complex use cases, the model can scale quickly and bestow the benefits of lowering the harm inflicted by AI on people and society.

The justifiable measurement of AI risk levels by HARM makes it possible to hold the technology products creator companies accountable for the harm inflicted by AI products that they release to the public.

==========

References:

1. OECD AI principles
2. NIST.AI.100-1 Artificial Intelligence Risk Management Framework (AI RMF 1.0)
3. United Nations Governing AI for Humanity – Final Report 2024.
4. Applied Human-Centric AI by Rajagopal Tampi
5. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark" – Pan et al., 2023 (arXiv)
6. https://arxiv.org/html/2506.18178v1?
7. https://www.sciencedirect.com/science/article/abs/pii/S136757880800028X?utm_source=perplexity
8. https://fiveable.me/adaptive-and-self-tuning-control/unit-1/challenges-limitations-adaptive-control/study-guide/ykziMAXBdqMZacas?utm_source=perplexity

