International Journal of Computer Applications Technology and Research
Volume 14–Issue 212 – 221, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1412.1020

A Risk-Adaptive Trust and Safety Governance Framework for AI-Enabled Online Platforms: An Advanced Methodology for Identifying, Measuring, and Mitigating Digital Harm

Treacy Chiderah Nweze
University of Dallas
USA

Abstract: The increasing integration of artificial intelligence (AI) into online platforms has created significant trust and safety challenges because AI systems can generate, amplify and distribute digital harms across different platform functions and stages of the AI lifecycle. This study develops a Risk-Adaptive Trust and Safety Governance Framework for AI-Enabled Online Platforms as a methodology for identifying, measuring, prioritising and mitigating digital harm. A systematic literature review and conceptual framework-development design was adopted. Relevant literature published between 2017 and 2024 was examined across major academic databases and authoritative sources, with attention given to AI-enabled digital harms, their determinants, risk assessment, risk classification, platform governance and mitigation strategies. The analysis indicates that AI-enabled digital harm is not limited to technical errors or harmful content but can emerge through data practices, model design, deployment conditions, algorithmic ranking, recommendation, moderation, personalisation and organisational decisions. The study further establishes that effective risk assessment requires multidimensional indicators that consider likelihood, severity, exposure, scale, vulnerability and contextual conditions rather than relying exclusively on technical performance measures. On this basis, the proposed framework connects five interrelated processes: identification of digital harm, examination of its determinants, measurement of risk, classification and prioritisation, and proportionate trust and safety mitigation. The framework adopts a continuous feedback approach in which risk assessment informs intervention and the outcomes of intervention generate evidence for subsequent reassessment. The study concludes that effective AI platform governance requires adaptive, evidence-based and accountable trust and safety mechanisms capable of responding to changing technologies, user behaviour and emerging forms of digital harm.
Keywords: Artificial intelligence; Digital harm; Platform governance; Risk assessment; Trust and safety

1. INTRODUCTION

Artificial Intelligence (AI) has become an important part of the operation of modern digital platforms. Online platforms increasingly use AI to recommend content, rank information, detect suspicious activities, personalise user experiences, moderate content and support automated decisions that affect users. These applications have improved the speed and scale of platform operations, but they have also created new forms of digital risk. Gorwa, Binns, and Katzenbach (2020) explain that platforms increasingly use automated moderation systems to deal with large volumes of user-generated content, creating important technical and governance challenges. Similarly, Acemoglu (2021) argues that the growing use of AI can create social, economic and political costs, including threats to privacy, consumer choice, equality and public discourse. As platforms depend more heavily on AI to make or support decisions, errors and unintended outcomes can therefore affect what users see, the information they receive, how their content is treated and how they interact with digital services. This makes effective management of AI-related risks an important part of contemporary digital platform governance.

The risks associated with AI-enabled platforms can occur at different stages of the AI system lifecycle. They may arise from the data used to train a system, the design of the model, the way the system is tested, the decisions made during deployment or the way users interact with the system. Suresh and Guttag (2021) identify multiple sources of harm that can emerge throughout the machine-learning life cycle, from data collection and preparation to model development and deployment. Raji et al. (2020) similarly emphasise the importance of examining AI systems throughout their development and deployment because effective accountability requires more than a single assessment at the end of development. These concerns are particularly important for digital platforms because AI-related harms can change as systems are updated, user behaviour changes and platforms introduce new features. The need for continuous assessment is also supported by Hilbert et al. (2024), whose analysis of algorithmic recommendations across several major platforms identified risks involving bias, discrimination, misinformation, mental health and other forms of user harm. Effective trust and safety governance must therefore identify and assess risks continuously rather than only respond after harm has occurred.

AI-enabled digital platforms can produce different forms of harm, including misinformation, discrimination, privacy risks, unfair treatment, harmful content, manipulation and unsafe automated decisions. Weidinger et al. (2021) identify a wide range of risks associated with language models, including discrimination and exclusion, toxic content, information hazards, misinformation, malicious use and manipulation. At the platform level, Saurwein and Spencer-Smith (2021) show that algorithmic harms can result from algorithmic errors, manipulation, reinforcement of existing inequalities, opaque and discriminatory practices and the increasing power of platforms over users and society. These risks may also affect

users differently depending on the social and technological context in which AI systems operate. Siapera (2022), for example, demonstrates that AI-based content moderation can reproduce racial inequalities when the experiences and voices of affected communities are not adequately represented in moderation processes. These findings suggest that digital harm cannot be managed effectively through a single technical solution. Platforms require governance processes that consider the source and context of risk, the users affected, the likelihood of occurrence, the severity of consequences and the effectiveness of available mitigation measures.

The need for structured AI risk management has also received increasing attention from standards, policy and research institutions. Tabassi (2023), through the National Institute of Standards and Technology (NIST) AI Risk Management Framework, presents AI risk management as a continuous process organised around four major functions: Govern, Map, Measure and Manage. These functions provide an organisational basis for understanding AI risks, measuring their characteristics and managing appropriate responses throughout the AI system lifecycle. The importance of structured assessment is also evident in the work of Green (2022), who questions the assumption that simply placing humans in the oversight role is sufficient to prevent algorithmic harm. His analysis shows that human oversight can itself be ineffective when organisations do not provide appropriate institutional mechanisms for challenging and evaluating algorithmic decisions. AI risk management therefore requires more than technical controls or the presence of a human reviewer. It requires clearly defined governance responsibilities, systematic risk assessment, appropriate monitoring and mechanisms for reviewing whether controls actually reduce harm.

Despite these developments, a gap remains in connecting AI-related digital-harm identification, risk measurement, risk prioritisation and mitigation within a single trust and safety governance process for online platforms. Existing research has addressed algorithmic auditing, content moderation, algorithmic harms and general AI risk management, but these areas are often examined separately. Imana, Korolova, and Heidemann (2023), for example, demonstrate the importance of platform-supported auditing for examining potentially harmful outcomes from opaque social-media algorithms and improving public oversight. Makhortykh et al. (2022) similarly show that governance of automated agents across online platforms varies considerably and that platforms often focus on particular forms of misuse rather than adopting a broader and generalisable governance approach. These studies indicate the need for a framework that connects risk identification, assessment, prioritisation, mitigation, monitoring and accountability within the specific trust and safety environment of AI-enabled online platforms. The present study therefore proposes a Risk-Adaptive Trust and Safety Governance Framework for AI-Enabled Online Platforms. The framework seeks to provide a structured methodology for identifying digital harms, measuring their likelihood and severity, prioritising risks and selecting proportionate mitigation measures while integrating governance, monitoring, human oversight and continuous improvement.

1.1 Statement of the Problem

The integration of AI into online platforms has created a significant governance problem because the same systems that improve efficiency and user experience can also generate, amplify or facilitate new forms of digital harm. The problem is particularly important because algorithmic systems can influence what information users encounter, how content is ranked and which activities receive greater visibility. DeCook, Cotter, and Kanthawala (2022) show that platforms do not simply respond to harm but also participate in defining what counts as harmful through their policies and governance practices. At the same time, Ozanne et al. (2022) demonstrate that the visibility and perceived accountability of AI moderators can influence how users understand and evaluate automated moderation systems.

The central problem, therefore, is the absence of a sufficiently integrated and adaptive methodology for connecting digital-harm identification, risk measurement, risk prioritisation and mitigation within a continuous trust and safety governance process for AI-enabled online platforms. Existing frameworks provide important foundations, particularly the NIST AI Risk Management Framework, which provides the Govern, Map, Measure and Manage functions for managing AI risks (Tabassi, 2023). However, the general nature of the framework means that it is applicable across many sectors and AI use cases rather than being specifically structured around the operational trust and safety challenges of online platforms. Gorwa et al. (2020) demonstrate the specific challenges involved in automated platform content moderation, but their work does not provide a complete methodology for connecting content moderation risks with broader AI lifecycle risks, measurement and adaptive mitigation. There therefore remains a need for a framework that brings these elements together and enables platforms to identify different forms of digital harm, assess their likelihood and impact, prioritise them according to their level of risk and apply proportionate responses. The study therefore addresses this methodological gap by developing a risk-adaptive trust and safety governance framework that provides a structured approach for identifying different forms of digital harm, establishing measurable risk indicators, assessing likelihood and impact, prioritising risks and applying proportionate mitigation measures.

1.2 Objectives of the Study

The general objective of this study is to develop a Risk-Adaptive Trust and Safety Governance Framework for AI-Enabled Online Platforms as an advanced methodology for identifying, measuring and mitigating digital harm. The study seeks to:

1. identify the major forms and sources of digital harm associated with AI-enabled online platforms;

International Journal of Computer Applications Technology and Research
Volume 14–Issue 212 – 221, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1412.1020

2. examine the key factors that influence the emergence and severity of AI-enabled digital harms across different platform functions and stages of the AI life cycle;

3. develop measurable indicators and criteria for assessing the likelihood, severity and potential impact of digital harm;

4. design a risk-adaptive methodology for prioritising digital harms according to their likelihood, impact, affected users and contextual characteristics; and

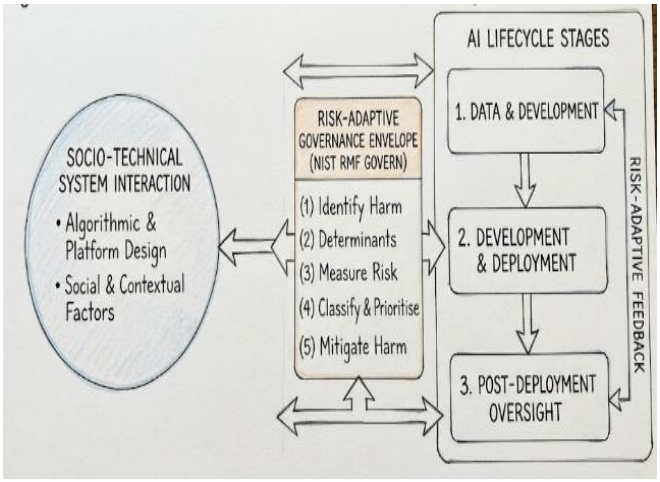
5. develop appropriate trust and safety mitigation strategies that correspond to different levels and types of AI-enabled digital risk

2. LITERATURE REVIEW

2.1 Forms and Sources of AI-Enabled Digital Harm

AI-enabled digital harm cannot always be adequately characterised as technical failure. As Saurwein and Spencer-Smith (2021) note, algorithmic harms on social-media platforms can take the form of "incorrect selection, manipulation of the recipient, extension of existing inequalities, facilitation of harmful activities, and concentration of power in the platform". According to Milano et al. (2020), recommender systems give rise to a set of ethical issues that extend beyond the immediate recipient of an recommendation to other relevant users and stakeholders in a given ecosystem. Other scholars have argued that digital harm needs to be viewed more broadly than as related only to illegal or harmful content, and includes impact from processes such as the selection, ranking, amplification or suppression of information. As a result, so too can the source of harm be more socio-technical than simply technical, arising instead from algorithmic design and platform-specific factors interacting with social factors.

Figure 1: Multidimensional Framework of AI-Enabled Digital Harm



Source: Author's compilation from literatures

Literature reveals harm can also stem from features introduced during the data and development stages that come before deployment. Suresh and Guttag (2021) propose numerous points in the machine-learning life cycle when harmful effects can be introduced including data collection, labelling, development and deployment. Sambasivan et al. (2021) show that poor data practices can create "cascades" of harm in high-stakes AI systems, with seemingly trivial problems becoming worse as they move through the development and usage life cycle. These existing results undermine the presumption that the final algorithmic decision is the main source of harm. Any system can cause harm even where its internal computational workings operate as intended. This is because its design itself may be based on intractably invalid assumptions, incomplete data and unrepresentative portrayals of the population to be served. The upshot is that properly recognising digital harm may call for analysing the full process by which a digital system is thought up, built and deployed, rather than merely focusing on its proximal outputs.

A further problem for existing classifications of harm is that AI systems – even more so than traditional information technology – can present types of harm that are hard to foresee in advance of being released. Bender et al. (2021) describe risks associated with the training data, scale and environmental impact of large language models that allow them to mimic harmful or harmful-sounding forms of language. Bommasani et al. (2021) enumerate broad 'capabilities' of foundation models and the second-order risks associated with their use, given that a general purpose model can be refined or adapted for many downstream uses. This presents a challenge to a traditional taxonomy of harm, where categories are clearly linked to a specific application, as general purpose models are embedded in multiple products and systems. Because the core technology is the same in each case, yet the harm produced varies depending on who makes use of it, how it is used, which data it processes and other contextual factors, the categories of harm we use need to be broad enough to cover technical failures, social impacts, misuse, systemic effects and context-specific harms.

2.2 Determinants of AI-Enabled Digital Harm Across the AI Lifecycle

The emergence of AI-enabled harm is not reducible to the quality of the model, but rather the interaction of technical, organisational and social factors. To this end, related work by Mehrabi et al. (2021) has found that bias can stem from many sources of machine-learning models, including data, modelling decisions and socio-technical interactions between technical systems and the social environment. This is extended by work by Selbst et al. (2019), which demonstrates that concerns about fairness cannot be addressed by framing an AI system as a technical object in isolation, for the algorithmic system is embedded in the sociotechnical context of its development and deployment. This contradicts attempts to mitigate social harms through improved data and mathematical notions of fairness. Technical refinements are

imperfect, but how relevant is the improvement in terms of having first identified the true nature of the problem and social factors in the system? AI risk stems from the interplay between technical properties and socio-technical conditions. A further factor is the quality and origin of the data. Gebru et al. (2021) identify the problem of poorly documented datasets, which make it difficult to understand why data was chosen and how it was collected, and to establish how it should be used. Mitchell et al. (2019) emphasise this at the model level through model cards, which prompt developers to specify use cases, performance metrics, intended audiences, limitations and other contextual information. Both approaches reveal an existing weakness in standard AI development processes: technical performance can be presented without sufficient information about the assumptions and boundaries of that performance. An AI system may measure well on aggregate evaluation data while failing on certain population groups or in specific circumstances. Lifecycle risk assessment must therefore establish the scope of system performance as well as whether it performs well, in particular populations and for specific purposes.

The organisational and governance environment matters for the probability and impact of harm as well. Wirtz, Weyerer and Kehl (2022) observe that the AI governance framework must adopt a risk-based, multi-stakeholder approach because the risks from AI are so numerous that they cannot be addressed by a single governance instrument. Floridi et al. (2018) similarly advocate an ethical framework that directly links opportunities and risks from AI with the principles of beneficence and non-maleficence, autonomy, justice and explicability. While such principles provide normative anchors, they cannot fully specify the actions needed to identify, measure and prioritise these risks within operational settings. There remains a practical disconnect between principles of good governance and the management of operational risk – effective prevention of harms from AI must be accompanied by governance systems capable of translating general principles into specific demands, triggers and mechanisms of control.

2.3 Measurement and Assessment of AI-Driven Digital Risks

Measurement of AI-enabled risk presents an even more significant challenge since different types of harm can't always be expressed on a common scale. Fellnder et al. (2022) promote risk-based methods of data-driven assessment that understand legal, societal and ethical considerations in the assessment of AI-enabled systems. Similarly, Giudici et al. (2024) show how to operationalise AI risk in more tangible measures such as accuracy, fairness, explainability, and sustainability. Such approaches are an important step away from principle-based forms of governance toward effective measurement that recognises systemic complexity. However, measurement must face the difficulty that these indicators capture different aspects of system performance and social impact. For example, high accuracy alone does not guarantee fairness, and 'being fair' on one measure will not necessarily mean that it is safe in another circumstance. Thus, risk assessment should not be based on a single, generic, universal measure; what is necessary is an array of indicators that considers the nuance of the risk.

Another illustration of the problem is the example of measuring fairness. Binns (2018) demonstrates that conflicting definitions of fairness rest on different philosophical assumptions, so that choosing a particular fairness measure is a value judgement as well as a statistical task. Similar points are made by Mehrabi et al (2021), for whom algorithmic bias has multiple causes and is expressed in different ways, in a way that cannot be reduced to a single statistic. The point, then, is that measurement is not in itself a normative choice: a system can be fair according to one measure, and unfair according to another (focusing on equality of treatment or outcomes, equality of outcome or opportunity, global balance or local attention to disadvantaged groups). Risk assessment, then, must distinguish measurement from assessment. Metrics give us data, but to determine whether the system makes unacceptable risks, we must consider the context, the community affected, the nature of possible damage and the relative importance of competing values.

These are complemented by additional approaches to assess model risk and thereby improve confidence in system safety. Mitchell et al. (2019) introduce model cards, a way to communicate model performance, intended use and limitations, while Raji et al. (2020) build a systematic process of internal algorithmic auditing, which links auditing to each phase of system development. Both methods acknowledge that risk assessment cannot be a one-off test conducted at the end of system creation. But documentation depends on the accuracy of the information given by developers and auditing might be constrained by the harms that only appear after system release and interactions with end users or other systems. Identification of harm and confidence in safety require ongoing collection of evidence, from both a comprehensive pre-deployment risk assessment as well as a thorough post-deployment review.

2.4 Risk Classification and Prioritisation of AI-Enabled Digital Harm

Why organise? Having identified many different risks, a second challenge is deciding which ones to organise around. Wirtz et al. (2022) show that there is not one, but many types of AI risk, each necessitating a different governance treatment, rather than a single controls solution. Giudici et al. (2024) view AI risk management through scalar measures of risks and propose a systematic way of assessing them. These perspectives suggest the need for classification because organisations do not have the capacity to organise and treat every identified risk identically. Problems emerge, though, when overly simply categorising various forms of harm ignores its context. A risk that is categorised as having a low frequency, for example, may have disastrous impacts if it is borne by a particularly disadvantaged population and a frequent, relatively low-impact problem may need a different

approach altogether. Classification therefore should be sensitive to not just the probability of an event, but the extent, magnitude, reach and impact of its consequences.

This challenge becomes even more clear when algorithmic outcomes impact groups differently. Binns (2018) shows the many competing definitions of fairness mean different solutions to the question of which outcomes are acceptable. And Selbst et al. (2019) warns the very logic of technical definitions of fairness is flawed because they abstract away from the social conditions under which algorithmic systems are designed and employed. In this way, risk priority cannot always be derived from a technical measure alone. While a difference of, say, 5% in social risk may be insignificant within a particular context, in another it may mean 5% more for the (say, 0.05%) of the population that will experience it. Risk classification is thus outside the boundaries of technical precision and must include contextual interpretation. Potential avenues for development should be evaluated along the criteria of severity, likelihood, scale, vulnerability, reversibility, and amplification capacity, not only margin of error.

Prioritisation also needs some way of translating results into organisational action. Mkander and Floridi (2021) state that ethics based auditing can help to provide the link between a set of principles and the concrete governance of AI, and Rahwan (2018) presents a 'society-in-the-loop' approach that asserts governance needs to take account of social interests instead of one group of operators or a technical system (or set of systems) deciding. Both of these lead to the conclusion that internalist, or technically dictated risk ranking is inadequate. Risk prioritisation involves value judgements over which interests to consider, where, and when to have a risk threshold. Robust risk classification would therefore need to clarify its criteria and have procedures for review, escalation and revision. Priorities should also be kept flexible to the extent technology, use patterns, deployment scale or social circumstances can change the relevance of the identified risks.

2.5 Trust and Safety Strategies for Mitigating AI-Enabled Digital Harm

AI risk mitigation requires a combination of a technical, organisational and human interventions and is not a single adjustable 'knob' to be turned to prevent safety and reliability failures. Shneiderman (2020) advocates for human-centred AI where automation is high but human control is maintained so that AI applications can foster reliability, safety and trustworthiness. This approach is complemented by a focus on the interaction between humans and computers, such as by Amershi et al. (2019) who highlight how human-AI interaction should be designed to convey the AI system's capabilities, limitations and behaviour. These strategies challenge the notion that automation improves safety; in many scenarios, it is better to have human intervention in decision making because the trade-offs are more complex or include multiple, contrasting values and significant hazards. How to mitigate risk should thus be proportionate to the risk, while considering the capabilities and limitations of the AI system.

Mechanisms of transparency and accountability may also act as mitigations. Gebru et al. (2021) propose comprehensive documentation of datasets as a technical approach for transparency and accountability; in a different vein, Mkander, Morley, Taddeo and Floridi (2021) look at an ethics-based auditing framework to assess automated decision-making systems and identify intervention points. These aspects are attractive because the shift of mitigation from modifying the algorithm to also encompass processes within the organisation is a positive step towards recognising risks. Nonetheless, transparency or auditing does not necessarily lead to safety. Transparency does not necessarily mean that information is communicated in an intelligible way, or that it is sufficiently relevant to be used effectively; auditing does not necessarily mean that a problem is addressed. There needs to be an appropriate acknowledgement of risk, responsibility, mitigation, and verification.

In addition, platform governance extends to interventions that change the information that is visible and how it spreads rather than only removing content. Gillespie (2022) illustrates how platforms leverage algorithmic reduction to minimise the prominence of 'misleading, risky, or offensive content', leading to a broadening of platform moderation beyond removal and suspension of accounts. Floridi et al. (2018), however, contend that responsible AI balances various principles, including beneficence, non-maleficence, autonomy, justice and explicability, and this presents a fundamental dilemma in AI-enabled safety: a single intervention to alleviate a particular harm may raise questions about another concern, such as autonomy, justice, expression, or accountability. Therefore, mitigation cannot simply be appraised by whether it prevents a particular harm. Instead, it should be appraised according to its effectiveness, proportionality, side effects, fairness and how well it aligns with legitimate user needs. Trust and safety thus demand flexible mitigation, which is able to change in response to evidence about its risks and benefits.

3. METHODOLOGY

3.1 Research Design

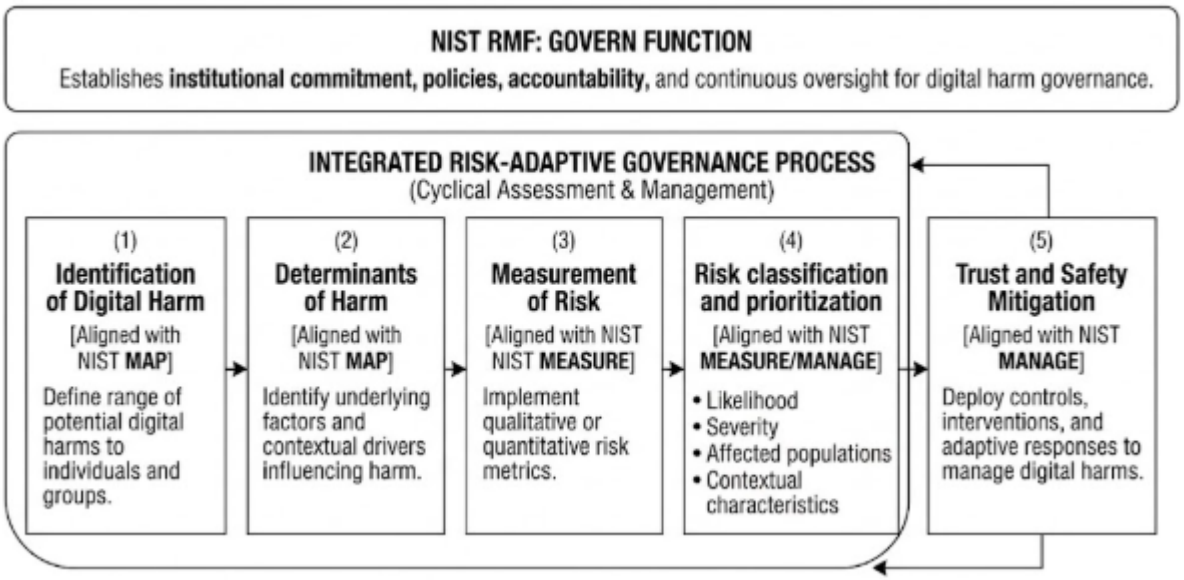
This study adopted a systematic literature review and conceptual framework-development design to examine digital harms associated with AI-enabled online platforms and develop a risk-adaptive trust and safety governance framework. The study systematically reviewed relevant scholarly and authoritative literature to identify the major forms and sources of AI-enabled digital harm, the factors influencing their occurrence and severity, approaches for measuring and assessing risk, and strategies for mitigating identified harms. A conceptual synthesis was then used to integrate these findings into a unified methodology for platform-level trust and safety governance. This approach is appropriate because the study seeks not only to describe existing knowledge but also to develop an integrated

International Journal of Computer Applications Technology and Research
Volume 14–Issue 212 – 221, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1412.1020

framework for identifying, measuring, prioritising and mitigating digital harm.

harm, algorithmic risks, platform governance, trust and safety, risk assessment, vulnerable users, or strategies for preventing and mitigating AI-enabled harm. Studies that mentioned AI

Figure 2: Integrated Risk-Adaptive Governance Framework for Digital Harm



Source: Author's compilation from literatures.

Figure 2 illustrates an integrated risk-adaptive governance framework for digital harm. The figure shows how an AI-enabled platform would assess risk across the four functions of the NIST AI RMF-govern, map, measure, and manage-under directed by the overarching function of govern. After mapping harm identification and its root causes, measuring risk, and determining what threat should be ranked as most imminent by severity, engagement level, specific population, and contextual conditions, trust and safety controls are applied by managed, with an integrated feedback loop to ensure data-driven refinements can be undertaken.

3.2 Literature Search Strategy

A structured literature search was conducted across Scopus, Web of Science, ScienceDirect, IEEE Xplore, ACM Digital Library and Google Scholar. The search combined keywords relating to artificial intelligence, AI-enabled platforms, online platforms, digital harm, algorithmic harm, AI risk, trust and safety, platform governance, risk assessment, risk mitigation and AI governance. Searches also considered combinations relating to specific platform functions, including recommendation, content moderation, ranking, advertising, generative AI and automated decision-making. Reference lists of relevant publications were additionally examined to identify important studies not retrieved through the database searches.

3.3 Study Selection and Eligibility Criteria

Studies were selected according to predefined inclusion and exclusion criteria. Priority was given to peer-reviewed journal articles, conference papers and authoritative institutional publications published between 2017 and 2024, reflecting the contemporary development of AI-enabled digital platforms. Studies were included where they addressed AI-related digital

only incidentally or focused exclusively on technical AI performance without relevance to digital harm, platform risk or governance were excluded. Titles, abstracts and full texts were screened for relevance before final inclusion.

3.4 Data Extraction and Evidence Coding

Relevant information was extracted using a structured coding framework. The extracted variables included AI technology or platform function, source and type of harm, affected users, AI lifecycle stage, determinants of harm, risk indicators, potential consequences, assessment methods and mitigation measures. Particular attention was given to factors relevant to risk adaptation, including the likelihood of harm, severity of consequences, scale of exposure, vulnerability of affected users and contextual conditions. The coding process provided the evidence base for connecting different forms of digital harm with appropriate risk assessment and trust and safety responses.

3.5 Data Synthesis and Framework Development

The evidence was synthesised thematically around five interconnected dimensions: (1) identification of digital harm,

(2) determinants of harm, (3) measurement of risk, (4) risk classification and prioritisation, and (5) trust and safety mitigation. These dimensions were subsequently integrated into a risk-adaptive governance process in which identified harms are assessed according to their likelihood, severity, affected populations and contextual characteristics. The structure is consistent with the NIST AI RMF's Govern, Map, Measure and Manage functions, while extending the logic specifically to AI-enabled online platforms and digital harm. The visual imagery can be seen below:

3.6 Risk Assessment and Prioritisation

The proposed methodology assesses digital risk by considering the likelihood and potential severity of harm, together with exposure, affected-user vulnerability, scale and contextual conditions. These factors are used to classify identified harms into different risk levels and determine which risks require immediate or enhanced intervention. The risk-adaptive approach recognises that not all AI-enabled harms require the same level of governance response; higher-risk situations require stronger safeguards, closer monitoring and greater human oversight. This is consistent with the principle that AI risk management should account for both the probability and magnitude of potential harm.

3.7 Trust and Safety Mitigation

This model maps each of the risk categories to a proportional set of trust and safety interventions. Possible mitigation measures include human oversight, content moderation, recommendation ranking, transparency tools, technical measures, access controls, user controls, incident response and active monitoring. The choice of severity of measures will be adapted to the severity and nature of the risk to ensure that trust and safety governance is not static but instead a risk-adaptive activity that varies with the level of reported harm, platform context and evidence of the system's effectiveness. NIST has an emphasis on responsible corporate governance of AI through continuous monitoring, clear delineation of responsibilities and risk-based governance of the lifecycle of AI.

3.8 Ethical Considerations

This paper was based entirely on secondary literature, using only publicly available leading sources and it did not employ human participants or collect personal information. Ethically the study was concerned with the selection of literature, the fair representation of other authors' work, attribution of sources and the avoidance of personal speculation. As the framework proposed involves users of platforms who may be regarded as 'vulnerable', the essay also considered the necessity for proportionate safeguards and protections.

4. DISCUSSION OF FINDINGS

The research found that the construct of AI-enabled digital harm was better conceptualised as a risk continuum than a discrete set of harm incidents. Harm could emerge over time via everyday platform functionality and remain insignificant until exposed repeatedly or interacted with other system processes. For example, a recommendation system alone may not produce harmful content but could promote the frequency, visibility and audience of such content, thereby altering the impact. Equally, ranking, categorisation, automated moderation and personalisation may produce cumulative effects via incremental decisions repeated millions of times by users. This indicates that traditional harm models, built around discrete incidents, could miss risks that accrue during less obviously harmful system functioning. Similarly, Peckham (2024) has argued that trustworthy AI should make clearer links between different types of specific AI deployments, harms and governance responses. This implies that trust and safety strategies need to identify emerging or time-dependent risks before they become serious, rather than primarily reacting to them after the fact.
More fundamentally, the cause of a harm can often be far upstream from the point of it becoming apparent. A harmful result observed on a platform may have stemmed from, say, a prior design assumption, a decision on data, a limitation of the model, a failure in the evaluation, or even the wrong policy or regulatory choice. Addressing just the symptom, therefore, may provide a temporary cure, but not a fundamental solution. This makes a distinction between symptom management and risk avoidance particularly significant: the platform might, for example, successfully tackle the problematic content but not the recommendation engine that keeps surfacing it to users. The approach set out here, by contrast, can offer analytical robustness by tracking risks back along the AI supply chain and pinpointing intervention points before they translate into user-facing problems.

Moreover, consequences of a specific risk cannot be separated from a system's deployment context. The consequences of a single technical failure can vary significantly based on platform scale, level of automation, users' cultural diversity and level of human oversight (as per who monitors the system). An error that affects only a limited set of users may have relatively insignificant consequences, while the same error in a widely used platform can turn into a large scale, well-repeated problem. Hence, technical risk and social impact are not interchangeable concepts. Mntymki et al. (2022) argue that responsible AI governance should go beyond the technical system and consider organisational and institutional actors in the deployment context as well. This type of risk analysis is not adequately captured by an assessment of model properties alone.
The other aspect to consider is organisational decision-making which shapes the risk-taking behaviour of the system. Decisions about tolerable levels of errors, staffing, escalation protocols, extent of monitoring and delegated authority to automated systems can contain or propagate a problem across the whole platform. AI harm in other words, results partly from decisions about management, not simply unintended side effects. This undermines the design-centric explanation of who is to blame (the developer) and highlights the need for coordination across technical teams, platform managers,

policy groups, safety teams and others. Jussupow, Spohrer and Heinzl (2024) additionally show that this form of AI governance depends on, among other things, the actors, object of governance, timing and mechanisms of implementation. This suggests organisational responsibility should be institutionalised across the responsible bodies that design, deploy and monitor AI systems.

Measurement is also important because simply identifying a potential harm does not necessarily tell us how important the harm might be. The platform will need to assess the probability of the harm occurring, the severity of the outcome, how many users are exposed and whether certain demographic groups are disproportionately affected. Measurement therefore turns broad concerns about AI safety into information that can be used to make governance decisions. But a platform's sole level of measurement of AI safety cannot be precision. A model may be accurate but it can still cause large social or distributional harm. Polemi et al. (2024) pointed out the lack of appropriate measurement for socially-related threats and human threats as an ongoing limitation to AI trustworthiness assessment. This indicates that the identified methods should treat metrics of technical performance as just one element of a comprehensive framework for risk assessment which considers social impacts and factors.

Equally, however, the process of measurement should not give the impression of certainty. Digital-platform risks may be hard to detect in the early roll-out phase, or evidence may evolve as AI systems engage with larger and more diverse populations. Thus, lack of any clear reports of harm will not necessarily equal lack of risk. And this is especially true for new AI capabilities, for which historical evidence may be sparse. So risk assessment should allow for uncertainty and for re-evaluation as new evidence emerges. This approach would allow for risk assessment to shift from a one-off certification activity to an ongoing learning cycle. In relation to the framework put forward, this means that risk scores and labels should be subject to review rather than being 'hard-coded' as labels once they are provided.

All of this means that another governance dilemma is emerging: which risks merit our attention? Given that platform resources are finite in terms of money, technical capacity and human expertise, it is unreasonable to impose the same high level of intervention on every identified risk. Risk assessment is ultimately aimed at directing attention to those risks that seem likely and/or severe, but this can be achieved by risk prioritisation. High priorities warrant intervention while risks of lower order may not be worth sustained concern. However, such lower-priority risks can still be significant if it is difficult for platform operators to prevent or reversibly mitigate their impacts or if their effects can somehow intensify very quickly. Peckham (2024) again underlines the potential for differentiated governance responses to AI harms depending on the type of risk involved. This is why risk classification should be used to aid rather than replace judgement. Numerical categorisations can aid in consistency and comparison, but can never in themselves determine the relative social impact of a harm associated with the AI system. Two risks with the same numerical categorisation may require different responses because the population affected is different, the scale of the impact is different, or because the scope of action is different. Therefore, context should not be sacrificed to the numerical categorisation. This is especially critical in trust and safety contexts where a platform action may have effects that go beyond the immediate incident or individual. A truly adaptive approach to classification should therefore enable challenge, review and revision of categories as circumstances and evidence evolve.

Interrelated with these are the other three concerns mentioned above; this 'can only be managed if we think of governance as a circular process rather than a linear one', thus the concepts of continuous assessment and intervention. Hence the fact that identification makes measurement possible; measurement makes prioritisation possible; prioritisation makes mitigation possible; and the results of mitigation creates the data that enables further assessment. This cyclical process further reflects an adaptive governance mode as opposed to a static or static–compliance model where you analyse once, put controls in place, and treat controls as static thereafter. Mntymki, Fortuin, Stam, and Brandt (2022) highlight the lifecycle and organization features of responsible AI governance, while Lahusen, Maggetti, and Slavkovik (2024) similarly focus on the interaction of technology, human actors, and institutions in trustworthy AI governance. The main advantage of this framework, therefore, rests on conceptualising governance as an ongoing feedback loop.

5. CONCLUSION

The study concludes that AI-enabled online platforms require a form of trust and safety governance that is continuous, evidence-based and responsive to changing levels of digital risk. The findings demonstrate that digital harm cannot be effectively managed through isolated technical safeguards, general ethical principles or reactive content-moderation practices. Instead, effective governance requires a connected process through which platforms identify emerging harms, determine the conditions that produce them, assess their significance, establish priorities and apply proportionate interventions while continuously reviewing the results of those interventions. The proposed Risk-Adaptive Trust and Safety Governance Framework provides this integrated structure by positioning risk as the basis for determining the intensity and type of governance response. Its contribution lies in connecting digital-harm assessment with practical trust and safety decision-making and in recognising that risk levels can change as technologies, users, platform functions and social conditions evolve. The framework therefore provides a methodological foundation for moving AI-enabled platform governance from predominantly reactive safety management towards a more anticipatory, adaptive and accountable approach.

6. RECOMMENDATIONS

1. AI-enabled services should incorporate formal governance processes so that trust and safety measures vary with the extent and type of the risk identified.

2. Operators of AI platforms should construct common metrics to assess the probability, gravity, magnitude and repercussions of AI-enabled harm.

3. Platforms need to put in place feedback loops that evaluate the impacts and any unforeseen effects of measures once implemented. A mitigation action cannot just be judged a success because it has been shown to address the issue it was designed to solve, but it's impacts on the users, fairness and transparency and overall behaviour of the platform need to be considered as well.

7. REFERENCES

Acemoglu, D. (2021). *Harms of AI* (NBER Working Paper No. 29247). National Bureau of Economic Research. https://doi.org/10.3386/w29247

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S. T., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 81*, 149–159. https://proceedings.mlr.press/v81/binns18a.html

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). *On the opportunities and risks of foundation models* (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

DeCook, J. R., Cotter, K., Kanthawala, S., & Foyle, K. (2022). Safe from "harm": The governance of violence by platforms. *Policy & Internet, 14*(1), 63–78. https://doi.org/10.1002/poi3.290

Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., Heintz, F., & others. (2022). Achieving a data-driven risk assessment methodology for ethical AI. *Digital Society, 1*, Article 13. https://doi.org/10.1007/s44206-022-00016-0

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689–707.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society, 8*(3), 1–10.

Giudici, P., Centurelli, M., & Turchetta, S. (2024). Artificial intelligence risk measurement. *Expert Systems with Applications, 235*, 121220.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 1–15.

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review, 45*, 105681. https://doi.org/10.1016/j.clsr.2022.105681

Hilbert, M., Thakur, A., Flores, P. M., Zhang, X., Bhan, J. Y., Bernhard, P., & Ji, F. (2024). 8–10% of algorithmic recommendations are 'bad', but… an exploratory risk-utility meta-analysis and its regulatory implications. *International Journal of Information Management, 75*, 102743. https://doi.org/10.1016/j.ijinfomgt.2023.102743

Imana, B., Korolova, A., & Heidemann, J. (2023). Having your privacy cake and eating it too: Platform-supported auditing of social media algorithms for public interest. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), Article 134, 1–33. https://doi.org/10.1145/3579610

Lahusen, C., Maggetti, M., & Slavkovik, M. (2024). Trust, trustworthiness and AI governance. *Scientific Reports, 14*, 20752. https://doi.org/10.1038/s41598-024-71761-0

Makhortykh, M., Urman, A., Münch, F. V., Heldt, A., Dreyer, S., & Kettemann, M. C. (2022). Not all who are bots are evil: A cross-platform analysis of automated agent governance. *New Media & Society, 24*(4), 964–981. https://doi.org/10.1177/14614448221079035

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics, 2*, 603–609. https://doi.org/10.1007/s43681-022-00143-x

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & Society, 35*(4), 957–967. https://doi.org/10.1007/s00146-020-00950-y

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics, 27*, Article 44. https://doi.org/10.1007/s11948-021-00319-4

Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication, 27*(4), zmac010. https://doi.org/10.1093/jcmc/zmac010

Peckham, J. B. (2024). An AI harms and governance framework for trustworthy AI. *Computer, 57*(3), 59–68. https://doi.org/10.1109/MC.2024.3354040

Polemi, N., Praça, I., Kioskli, K., & Bécue, A. (2024). Challenges and efforts in managing AI trustworthiness risks: A state of knowledge. *Frontiers in Big Data, 7*, 1381163. https://doi.org/10.3389/fdata.2024.1381163

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, 20*, 5–14. https://doi.org/10.1007/s10676-017-9430-8

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 39, 1–15. https://doi.org/10.1145/3411764.3445518

Saurwein, F., & Spencer-Smith, C. (2021). Automated trouble? The role of algorithmic decision-making in media and communication. *The Computing Journal*. 26(3), 45-51.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Siapera, E. (2022). AI content moderation, racism and (de)coloniality. *International Journal of Bullying Prevention, 4*, 55–65. https://doi.org/10.1007/s42380-021-00105-7

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv*. https://doi.org/10.48550/arXiv.2112.04359

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly, 39*(4), 101685. https://doi.org/10.1016/j.giq.2022.101685