White Paper

2025 NSF US-Southeast Asia Regional Workshop on Responsible Artificial Intelligence

Organizer: Junsong Yuan (University at Buffalo)

Local Chair: Xinchao Wang (National University of Singapore)

Steering Committee: Tsuhan Chen (National University of Singapore), Yew-Soon Ong (Nanyang Technological University), Michael Zeller (Temasek);

Participants: Junsong Yuan (University at Buffalo); Huynh Thi Thanh Binh (Hanoi University of Science and Technology); Thittaporn Ganokratanaa (King Mongkut's University of Technology Thonburi); Koteswar Rao Jerripothula (Indian Institute of Technology Kanpur); Wiboonsak Watthayu (King Mongkut's University of Technology Thonburi); Supavadee Aramvith (Chulalongkorn University); Jianfei Cai (Monash University); Sansanee Auephanwiriyakul (Chiang Mai University); Anirban Chakraborty (Indian Institute of Science); Basabdatta Sen Bhattacharya (BITS Pilani, Goa); David Doermann (University at Buffalo); Tong Xin (National University of Singapore); Ye Wang (National University of Singapore); Rakesh Nagi (Singapore University of Technology and Design); Xinchao Wang (National University of Singapore); Mohan Kankanhalli (National University of Singapore); Bihan Wen (Nanyang Technological University); Wei Lu (Singapore University of Technology and Design); Roy Ka-Wei Lee (Singapore University of Technology and Design); Arne Suppe (Singapore Management University); Ivor Tsang (A*STAR CFAR); Yiyu Cai (Nanyang Technological University); Frank Guan (Singapore Institute of Technology); Xiaoli Li (ASTAR Institute for Infocomm Research); Sijia Liu (Michigan State University); Hao Wang (Rutgers University); Maja Pantic (Imperial College London); Amy Zhang (University of Texas at Austin); Yingli Tian (City College of New York); James M. Rehg (University of Illinois Urbana-Champaign); Qiang Ji (Rensselaer Polytechnic Institute); Yiwei Wang (University of California, Merced); Daniel Schiff (Purdue University); Narendra Ahuja (University of Illinois Urbana-Champaign); Yujun Cai (The University of Queensland); Wei Tang (University of Illinois Chicago); Grigorios Chrysos (University of Wisconsin–Madison); Sharon Li (University of Wisconsin–Madison); Huaiyu Dai (North Carolina State University); Tsuhan Chen (National University of Singapore); Michael Zeller (Temasek); Yew-Soon Ong (Nanyang Technological University); Vladimir Pavlovic (Rutgers University); Aw Ai Ti (A*STAR Institute for Infocomm Research)

Table of Contents

Preface

The US-Southeast Asia Responsible Artificial Intelligence (AI) Workshop was held on April 29 - 30, 2025, at AI Singapore, located on the campus of the National University of Singapore. This workshop was supported by the U.S. National Science Foundation (NSF) and AI Singapore. With the rapid advancement of multi-modal foundation models, agentic AI, and embodied AI, a new wave of AI technologies is transforming nearly every aspect of our lives. The widespread availability of increasingly powerful AI tools holds the potential to fundamentally reshape how we communicate, work, live, and learn, while raising profound new questions about ethics and society [1]. As AI becomes more ubiquitous, its progress shows no signs of slowing. In fact, we are witnessing intensifying global and regional competition in AI-related domains such as data, computational resources, and workforce development [6]. Less frequently addressed, but equally critical, is the risk associated with the development and deployment of AI systems. As the pace of AI innovation accelerates, it becomes essential to assess and forecast these risks, raise public awareness, and help promote the responsible development and use of AI.

This workshop was organized to explore some of the key questions surrounding Responsible AI**:**

- What are the main challenges and opportunities in developing responsible AI?
- How can we develop responsible AI and grow a workforce aligned with responsible AI principles?
- How should we evaluate responsible AI technologies?
- And to what extent are international collaborations essential for advancing responsible AI?

To address these questions, the workshop convened researchers from across the United States and Southeast Asia, as well as thought leaders in computer science, engineering, statistics, AI policy, and ethics. Through a combination of online surveys and in-person discussions, participants examined the goals and evaluations of responsible AI, the barriers to its development, and the roles that both the U.S. and Southeast Asian countries can play in shaping its future.

This report aims to summarize our collective reflections and findings as of summer 2025, offering a snapshot of thinking in this rapidly evolving field. The report also highlights promising directions for collaboration and partnership between federal funding agencies in the United States and their counterparts in Southeast Asia. For decades, the NSF has been a key enabler of both domestic and international scientific research, especially in AI and related areas. Basic research, combined with international partnerships, continues to fuel the U.S.'s leadership in AI innovation. The training of a diverse and globally engaged AI workforce has helped make the U.S. a vibrant hub for technological progress. Sustained support for academic research and international collaboration [53] will be critical to ensuring that AI technologies are developed not only with excellence, but with responsibility, inclusivity, and global impact in mind.

Executive Summary

Artificial Intelligence (AI) is advancing rapidly and is recognized as a strategic national priority in both the United States and Southeast Asian countries. While the capabilities of AI technologies, such as large language models, multi-modal systems, and autonomous agents, are progressing at an unprecedented pace, their associated risks are becoming more complex [2, 8] and intertwined with daily life. High-profile concerns, including deepfake content generation, privacy violations, hallucinatory outputs in response to complex queries, and autonomous driving incidents, highlight the urgent need for proactive approaches to responsible AI. To date, much of AI research and development has focused on technical performance and benchmarks, often overlooking the frameworks, metrics, and methodologies needed to make AI safe, ethical, and aligned with societal values [9]. Now is a pivotal time to strengthen the connection between AI capability and AI responsibility, moving beyond performance toward accountability.

This workshop emphasized the need to approach responsible AI through both technical and non-technical aspects, recognizing the mutual influence between AI and society. Three key dualities define this relationship:

1. **Power and Risk**: AI offers immense benefits but also carries risks if misused or under-regulated. Assessing and addressing these risks is essential to building trust and ensuring safe, long-term innovation.
2. **Technology and Humanity**: Responsible AI is not solely a technical issue and it depends on how people are trained to develop, use, and govern it. Education and workforce development play a vital role in shaping ethical AI practices.
3. **National Investment and Global Collaboration**: International partnerships are crucial for sharing policies, standards, and successful use cases. Cross-border collaboration enables scalable, inclusive approaches to responsible AI across diverse contexts.

Key Opportunities:

- **Integration between Technology and Humanities:** Responsible AI requires close collaboration between technologists and scholars in the humanities to integrate ethical, technological, and societal perspectives into design. Academic research plays a key role, as its incentives align with public benefit rather than corporate interests. While companies focus on "individual" AIs for their products, academia can explore broader goals including AI alignment, ensuring that decisions benefit society at multiple levels and that conflicts among diverse interests are properly managed.
- **Global, Interdisciplinary, and Culturally-Aware Collaboration:** Responsible AI requires coordinated efforts across countries and disciplines, which brings together policymakers, lawyers, social scientists, economists, and technologists, while also appreciating cultural diversity. This integration aims to foster AI systems that respect cultural norms, reduce bias, and align with legal, ethical, and societal frameworks, enabling globally responsible and inclusive AI development.
- **Applied and Translational Research with Human-Centered Computing Focus** AI moving from theory to real-world deployment demands careful attention to safety, usability, and ethics. Human-Centered Computing (HCC) plays a key role in designing interpretable, user-friendly systems that support transparency, accountability, and meaningful human oversight.

1.Risks, Challenges, and Opportunities of Responsible AI

1.1. Risks of Irresponsible AI

1.1.1 Who is Responsible?

A fundamental problem in responsible AI [100-104] is determining who bears responsibility when AI systems fail, mislead, or cause harm. As AI models become more autonomous and their outputs increasingly influence real-world decisions whether in medicine [82-84], finance [85-87], education [88-90], or justice [91-93], the question of accountability becomes both more urgent and more complex [94-96]. The issue is not only about legal responsibility but also about ethical obligation and system design [11]: who should be expected to anticipate, detect, and mitigate undesirable AI behavior?

At the individual level, users of AI, especially those in professional or high-stakes contexts, do bear some responsibility. For instance, a PhD student using a large language model to draft a literature review is expected to critically verify the content, ensuring accuracy and integrity before incorporating it into academic work. However, placing the full burden on individual users is unrealistic and inequitable. Not all users possess the same levels of AI literacy, and many are unaware of the limitations and risks inherent in AI-generated outputs. Expecting the average user to scrutinize hallucinations, assess model confidence, or detect subtle biases, often without transparency tools, creates an uneven and potentially harmful landscape.

Organizations that develop or deploy AI systems carry a heavier burden of responsibility. Enterprises integrating AI into consumer-facing products or critical decision-making pipelines must design systems with appropriate fail-safes, such as uncertainty quantification, interpretability, human-in-the-loop verification, and auditing mechanisms. In domains like banking, insurance, and healthcare, where algorithmic decisions affect lives and livelihoods, organizations are responsible not only for compliance with regulatory standards but also for ethical deployment and transparency in AI use [83]. Simply disclaiming responsibility because “the model made a mistake” is no longer acceptable when organizations actively choose to operationalize AI in consequential settings.

Responsible AI is not solely a technical problem. It is a sociotechnical challenge that requires legal infrastructure, public oversight, and cultural adaptation. Just as societies have evolved laws and norms around industrial safety, consumer protection, and environmental responsibility, a similar collective framework must emerge for AI. This includes policy-making that sets liability standards, educational efforts that promote digital literacy, and public institutions that can adjudicate harm and enforce accountability. Without such frameworks, the diffusion of responsibility may allow harmful AI deployments to go unchallenged and uncorrected.

Ultimately, the risks of responsible AI are not just about what the models do, but about how societies choose to allocate responsibility across the AI lifecycle, from research and development to deployment and use. Ensuring that responsibility is distributed in a fair, transparent, and enforceable manner is essential for trust, safety, and long-term societal alignment.

1.1.2 Unpredictable Future Irresponsible Behaviors

A significant risk associated with the rapid advancement of AI technology is the emergence of unpredictable and irresponsible behaviors that can cause harm in unforeseen ways. As AI models become more autonomous and capable of self-exploration, they may develop emergent properties, unexpected behaviors or abilities not explicitly designed or anticipated by their creators [4]. These emergent properties can result in unintended creativity, hallucinations, or even harmful actions, posing serious risks to safety and trust [34]. Moreover, the potential for adversarial misuse amplifies these risks, as malicious actors can exploit AI's power to deceive, manipulate, or cause disruption on a large scale [35].

Embodied AI systems, robots and autonomous agents interacting physically with the world, introduce tangible risks that go beyond digital errors. Hallucinations or failures in robustness within these systems can translate into dangerous real-world consequences [34], such as accidents, property damage, or bodily harm to humans [84]. For example, an autonomous delivery robot misinterpreting its environment might collide with pedestrians, or a medical robot might execute incorrect procedures due to faulty perception. These risks emphasize the high stakes involved when AI's digital decisions have direct physical impact, making failures potentially catastrophic [40].

Similarly, the rise of AI systems capable of generating and executing code presents an amplified threat surface. Such code-generating AI can autonomously write software that may contain vulnerabilities, bugs, or even malicious functionality. If left unchecked, these models could inadvertently or intentionally be used to deploy cyberattacks, malware, or automated exploitation of security flaws. This capability increases the risk of widespread harm across digital infrastructures. The absence of reliable external controls or "guardrails" on these AI agents raises the possibility of misuse or accidental damage, compounding concerns about their safe deployment.

Finally, the unpredictability of advanced AI behaviors, especially as agents gain greater autonomy and physical or computational agency, mirrors risks seen in other high-stakes domains, such as biosecurity. Just as research labs take extreme precautions to prevent accidental release of dangerous pathogens, AI development demands rigorous safeguards to prevent harmful "escape" of unsafe behaviors. Without comprehensive understanding and regulation, the unintended consequences of errors, hallucinations, or malicious use could cause significant physical, economic, and societal damage [30]. These risks highlight the urgent need for robust frameworks that anticipate not only known failures but also the unknown, emergent risks posed by increasingly capable AI systems.

1.1.3 Rapid Deployment vs. Control

One of the critical risks of irresponsible AI arises from the sheer speed at which AI models can be duplicated, adapted, and deployed across the globe. Unlike traditional technologies that require manufacturing, physical distribution, or lengthy certification processes, AI systems, once developed, can be copied and put into use almost instantly by anyone with access. This rapid proliferation challenges the ability of developers, organizations, and regulators to enforce proper

training, ethical guidelines, and safety protocols. As AI innovation accelerates, the infrastructure for oversight and responsible deployment will struggles to keep pace, creating dangerous gaps in accountability and control [5, 11].

This speed-driven diffusion amplifies risks because insufficiently tested or poorly understood models can be unleashed at scale without adequate safeguards. Unvetted deployments increase the likelihood of harmful outputs such as biased decisions, misinformation, or unsafe autonomous behaviors. Moreover, the decentralized nature of AI development means that many actors, from individual hobbyists to large corporations, may deploy powerful models with varying levels of expertise and ethical awareness. This diversity further complicates efforts to bring consistent standards and effective governance, exposing users and society to unpredictable risks stemming from uncontrolled AI use.

A particularly alarming aspect of rapid AI deployment is the so-called "Self-Inflecting Code Problem," where AI systems gain the ability to autonomously write, modify, and execute their own code. This capability expands the threat landscape dramatically, as an AI agent might generate new software or scripts without human review, potentially introducing malicious, buggy, or unsafe functions. The autonomous generation and execution of code could enable the creation of self-propagating malware, sophisticated cyberattacks, or automated exploitation of vulnerabilities at unprecedented speed and scale. Without robust technical controls, such as external sandboxing, runtime monitoring, or strict permission systems, the rapid and unsupervised use of self-modifying AI raises profound risks for cybersecurity, privacy, and system integrity.

In summary, the tension between rapid AI deployment and the ability to maintain effective control represents a critical risk factor for irresponsible AI. While accelerated innovation can drive tremendous benefits, it also magnifies vulnerabilities inherent in underdeveloped oversight mechanisms. Balancing this tension demands urgent development of scalable safety frameworks, regulatory approaches, and technical guardrails that can keep pace with the velocity of AI dissemination, thereby mitigating harms before they escalate beyond manageable levels.

1.1.4 Deterioration of Human Intelligence

As artificial intelligence becomes increasingly embedded in education, workplaces, and daily life, a critical concern is emerging around its unintended cognitive and social consequences, specifically, the deterioration of human intelligence through over-reliance. While AI offers unprecedented capabilities in enhancing productivity and decision-making, there is a growing risk that individuals, institutions, and even entire societies may gradually lose essential cognitive skills as they defer to AI systems in tasks ranging from information retrieval and problem-solving to writing and planning [97-99].

This over-dependence carries multiple dimensions. One of the most immediate concerns is skill erosion, particularly among students and knowledge workers. When learners rely on generative AI tools to complete assignments, code, or compose essays without engaging in the cognitive labor behind them, they risk weakening the very competencies that AI is meant to support. Similarly, professionals who routinely use AI to summarize information or suggest solutions may experience

a gradual decline in independent reasoning and analytical thinking. If education and public awareness fail to keep pace, learners may grow dependent on AI without fully grasping its limitations, biases, or contexts of misuse. Gaps in lifelong learning and professional development can leave individuals across all demographics ill-equipped to question or challenge AI systems, even those labeled as "responsible."

Another significant risk associated with skill erosion is addiction to convenience. The ease with which AI tools generate coherent and seemingly authoritative outputs can discourage users from applying their own judgment or seeking alternative perspectives. Over time, this frictionless interaction may lead to cognitive passivity, a behavioral shift where users instinctively turn to AI for answers without critically engaging with the underlying issues. This problem is compounded by the phenomenon of hallucinations, where AI systems produce incorrect or fabricated information with high confidence and persuasive language. There are already harmful use cases of hallucinations in domains such as healthcare [69, 70], legal advice [71, 72], and education [73, 74], and over-reliance on AI without verification can amplify these risks. As users become conditioned to trust AI-generated responses, they may fail to detect errors, leading to misinformation, flawed decisions, and the erosion of critical thinking. Such habits, once formed, can be difficult to reverse and may ultimately compromise intellectual agility and creativity [23].

In parallel, concerns around equity and fairness further complicate this dynamic. Access to high-quality AI tools and the skills to use them effectively are not evenly distributed across populations or regions. Over-reliance on AI may disproportionately benefit individuals or communities with greater digital literacy and infrastructure, while others fall behind, which will deepen existing inequalities in education, employment, and civic participation. Furthermore, users with limited understanding of AI's limitations are more vulnerable to being misled by biased or inaccurate outputs, such as fake contents generated by AI, exacerbating risks for already marginalized groups.

Finally, even when deployed with the intention of being responsible, AI can still carry significant risks if evaluation focuses too narrowly on technical system performance. Without a framework that prioritizes human–AI collaboration, there is a danger that responsible AI will still erode rather than strengthen human cognition. Overemphasis on replacing human decision-making, rather than augmenting it, can lead to diminished learning, insight, and judgment, even in systems designed with ethical safeguards. The absence of benchmarks for measuring cognitive engagement and user agency creates further risks. Responsible AI tools that lack transparency, fail to explain their reasoning, or provide "black box" answers can inadvertently encourage passive acceptance of outputs. This undermines critical thinking and reduces the user's role to that of a consumer of decisions, rather than an active participant in them.

1.2 Challenges of Responsible AI

1.2.1 Subjectivity and Inconsistent Definition of Responsibility

A fundamental challenge in building responsible AI systems is the lack of consistent definitions for core terms such as "responsibility," "trustworthiness," and even "reliability." These terms are often used interchangeably across technical, policy, and public discussions, yet they carry different

meanings depending on the context [41]. For example, what one stakeholder considers "responsible" behavior in an AI system, such as prioritizing user privacy, may not align with another stakeholder's priorities, such as maximizing accessibility or efficiency. The ambiguity around these hampers the development of clear benchmarks and regulatory standards [47, 151], making it difficult to measure or enforce responsible behavior across different domains and applications. As an alternative, some have proposed defining "irresponsible AI behaviors" as a way to delineate unacceptable outcomes, including examples like generating misinformation, breaching user privacy, or manipulating user behavior [10].

This definitional ambiguity becomes even more problematic when attempting to evaluate specific model behaviors. A key example is the challenge of defining and detecting "hallucination" in large language models [41, 131, 133]. While some see hallucination as a form of bad generalization, where a model offers plausible but incorrect outputs, others regard it as genuine, or even intentional, fabrication, untethered from any input or external reality. This latter view is particularly relevant in the context of machine unlearning [39, 142], where "hallucination" may reflect the unlearned model's attempt to remove unwanted behaviors to some extent, corresponding to its capability of degenerating the original, undesired response. The lack of consensus on what constitutes a hallucination affects how models are evaluated and fine-tuned, especially in high-stakes settings like healthcare or law. Without a clear, shared understanding of the phenomenon, detection and mitigation techniques remain inconsistent, and models may continue to produce misleading outputs under the guise of fluency and confidence [41]. This definitional gap not only impacts performance metrics but also impedes efforts to hold AI developers accountable for harms caused by model outputs.

Not only the definition of responsibility changes based on the type of AI, but also the degree of responsibility changes. For example, Embodied AI must be much more responsible because it can cause physical injury and harm. Further complicating the issue is the inherent subjectivity of what counts as "responsible" behavior, particularly when designing training objectives or reward functions in reinforcement learning. Ideally, responsible AI systems should align with user values and societal norms, but encoding these values into mathematical formulations is inherently difficult. Reinforcement learning systems, for instance, are especially prone to reward hacking [10], where agents exploit poorly specified or superficial reward signals to achieve high performance in unintended ways. This creates a disconnect between intended outcomes and actual behaviors, undermining efforts to build AI systems that behave ethically or safely. Designing better reward models that resist such exploitation, and reflect human-centered objectives, remains an open and critical research area. Until these subjective aspects can be reliably translated into technical specifications, the challenge of defining and achieving responsible AI will persist.

1.2.2 Difficulty in Control and Regulation

A major challenge in ensuring responsible AI lies in the fundamental difficulty of controlling and regulating its development and deployment. Unlike physical technologies such as nuclear energy or biological materials that are naturally constrained by material access, supply chains, and infrastructure, AI is a digital, intangible technology that can be easily copied, distributed, and modified across borders [47]. Its non-physical nature makes it inherently difficult to contain or

monitor. Once an AI model is trained or released, especially if made open-source, its downstream uses become diffuse and difficult to track [43], allowing actors with vastly different intentions and capabilities to appropriate it for both beneficial and harmful purposes. This ease of replication and modification makes traditional models of regulation, based on scarcity or central control, largely ineffective for AI governance.

The global and decentralized nature of AI development adds further complexity. While some jurisdictions may implement strong regulatory frameworks, others may act as safe havens for unregulated AI experimentation. This disparity creates loopholes and uneven enforcement, making it difficult to establish coherent global norms [47]. Moreover, many existing legal frameworks lack the agility to keep pace with the rapid evolution of AI technologies, from foundation models and autonomous agents to generative tools and decision-making systems. Even well-intentioned efforts to impose licensing, usage restrictions, or compliance auditing often encounter pushback [44] due to fears of stifling innovation or restricting access to valuable technologies. Unlike other regulated sectors like healthcare or aviation, AI lacks mature institutions to enforce safety, accountability, and certification standards across the development pipeline.

A closely related problem is the lack of effective forensic and attribution mechanisms [43]. When AI systems are misused, for example, to generate harmful deepfakes, automated scams, or manipulative content, it is often exceedingly difficult to identify the origin or hold users accountable. Many AI models can be accessed anonymously or through loosely governed APIs, and the outputs they generate are frequently indistinguishable from human content. While watermarking and digital fingerprinting have been proposed as partial solutions, these techniques remain fragile and are easily circumvented by adversarial users [41]. As a result, malicious or negligent use of AI can often proceed without consequence, raising significant concerns for law enforcement, national security, and civil society.

Moreover, even benign uses of AI can create serious risks when deployed without adequate safeguards. For example, autonomous systems in transportation, healthcare, or finance may behave unpredictably in rare or adversarial conditions, yet existing regulatory protocols are often too slow or fragmented to respond effectively. The lack of clear liability frameworks [17] further compounds this problem. When an AI-enabled decision harms someone, it is often unclear whether responsibility lies with the developer, deployer, or user [47]. These ambiguities complicate efforts to enforce ethical standards or legal redress.

In summary, the challenge of controlling and regulating AI is not only technical but deeply institutional and geopolitical. It requires the development of novel frameworks that account for AI's unique characteristics: its digital malleability, its potential for autonomous behavior, its global reach, and its integration into both public infrastructure and private enterprise. Addressing these issues will require international coordination, robust auditing infrastructures, legal innovation, and cross-sector collaboration to support that AI can be safely and fairly governed in the public interest.

1.2.3 Evaluation Challenge

A core challenge in advancing responsible AI lies in how to evaluate it meaningfully and consistently. Despite rapid technical progress, the field lacks clear, standardized methods for

assessing whether AI systems behave in ways that are safe, ethical, and aligned with human values [39, 41]. Existing benchmarks often fall short, focusing on narrow tasks or superficial indicators that fail to capture the complexity of real-world behavior. Moreover, the definitions of core evaluation targets, such as fairness, truthfulness, or responsibility, are often subjective or context-dependent. This makes it difficult to develop universal metrics or cross-system comparisons. As AI systems grow in scale and complexity, including those that operate in physical environments or make autonomous decisions over time, the inadequacy of current evaluation paradigms becomes even more pronounced.

Large language models (LLMs), including multi-modal variants, present significant and evolving challenges in responsible evaluation. While hallucination, the generation of plausible but factually incorrect or fabricated content, remains a well-documented concern, it is only one facet of the broader evaluation dilemma. Another persistent difficulty is determining whether an LLM can answer questions safely and ethically, particularly when prompted with ambiguous, adversarial, or sensitive queries. Unlike traditional software systems, LLMs do not operate based on explicitly coded rules; instead, their outputs emerge from statistical correlations learned from massive datasets, making it difficult to predict or constrain their behavior in edge cases. Evaluating their ability to avoid reinforcing stereotypes, propagating misinformation, or generating harmful advice requires nuanced assessments that go beyond accuracy or coherence. Moreover, there is no reliable way to confirm whether an LLM has truly "unlearned" unsafe or outdated content after retraining or fine-tuning, or if such content can resurface under different prompts or contexts [39]. Attempts at fine-grained control, through alignment techniques or post-hoc moderation, are often brittle and non-transparent. These challenges are compounded by the lack of universally accepted benchmarks [46] for safety, responsibility, or ethical compliance, leaving developers to rely on ad hoc tests and shifting norms. To make meaningful progress, the field must develop more systematic, adversarial, and context-aware evaluation methodologies that can test not only what LLMs know, but how they reason, when they fail, and whether they behave consistently with human values under uncertainty.

When it comes to embodied AI, systems that perceive and act in the physical world, responsible evaluation introduces unique complexities. While objective performance metrics such as task completion, safety, and energy efficiency offer a grounding advantage over purely digital systems, many ethical and social dimensions remain difficult to measure. For example, how should we evaluate a robot's behavior in terms of cultural norms, interpersonal etiquette, or accessibility for vulnerable populations? Current evaluation protocols focus heavily on mechanical functionality, often neglecting the social and moral implications [25] of physical presence and interaction. Moreover, real-world testing environments are costly and limited, while simulations may not accurately capture the unpredictability and nuance of human-robot engagement. This gap highlights the need for new interdisciplinary evaluation frameworks that can assess not only physical outcomes but also the human experience of embodied AI systems in shared environments.

Agentic AI systems, which exhibit long-term planning, memory, and autonomy, pose perhaps the most formidable evaluation challenge for responsible AI. These systems operate across time, learn from past interactions, and often adapt to evolving goals, characteristics that make static or one-off benchmarks inadequate. Traditional metrics may fail to capture issues like value drift, where an AI system's objectives subtly diverge from human intent as it learns. In multi-agent scenarios,

further complications arise around collective dynamics such as collusion, unfair competition, or emergent strategic behavior, where responsibility is distributed and outcomes are harder to predict. Current tools are ill-equipped to evaluate how these systems manage trade-offs, resolve ethical dilemmas, or maintain alignment over time. As AI agents take on more decision-making roles in finance, education, healthcare, and governance, the need for longitudinal, behaviorally rich, and ethically aware evaluation strategies becomes increasingly urgent.

1.3 Opportunities of Responsible AI

1.3.1 Policy, Regulation, and Licensing

The development of responsible AI presents a vital opportunity to establish new technical interventions, policy frameworks, regulatory standards, and licensing systems [37, 12], in order to build AI technologies that are safe, ethical, and accountable. Just as human drivers must be licensed due to public safety concerns, a similar model can be envisioned for the deployment and access to powerful AI systems. These frameworks could include user authentication, model certification, third-party auditing, and system-level licensing, particularly when AI is used in high-risk domains such as healthcare, transportation, or content generation.

Responsible AI governance demands more than technical interventions; it requires coordinated efforts that span the legal, ethical, and societal dimensions of AI. Participants in the US–Southeast Asia workshop emphasize the opportunity to proactively shape how AI aligns with public interest, by defining clear social objectives, setting enforceable boundaries, and creating systems of accountability that scale with the power of AI models [50].

Key opportunity areas for technical interventions include:

- *Embedding Explicit Ethical Rules in AI Training*

A forward-looking strategy is to incorporate explicit normative rules and ethical constraints directly into the training process of AI models. Much like physics-based constraints prevent scientific simulations from producing implausible results, ethical guidelines embedded into learning objectives can help steer AI behavior toward responsible outcomes. While technically promising, defining culturally relevant and enforceable ethical norms remains a significant challenge, particularly for issues like fairness, bias, and value alignment [37].

- *Formal Methods and Verifiable Guarantees*

Responsible AI also opens the door for applying formal methods and mathematical proofs to AI system design. These techniques, well-established in safety-critical engineering domains, can offer strong guarantees about system behavior, correctness, and reliability. The shift from performance-driven benchmarks to verifiability and robustness marks a foundational opportunity for transforming how AI is validated and trusted.

• Refusal Mechanisms and Built-in Constraints

Another promising direction is the development of refusal mechanisms within AI systems, capabilities that allow a model to decline certain requests when they are unsafe, unethical, or beyond its scope. Such built-in constraints would support more autonomous and principled decision-making, reducing the burden on external content filters and moderation systems. Meanwhile, such refusal mechanisms should be robust and honest, genuinely declining requests that would trigger harmful knowledge generation, rather than merely appearing to do so for specific phrasings while remaining susceptible to jailbreak attempts via alternative formulations.

• Forensics and Traceability for Accountability

The rise of synthetic media and autonomous decision-making raises the need for forensic tools and traceability protocols. Responsible AI initiatives can lead to the design of systems that watermark generative outputs, log interaction histories, and preserve audit trails to attribute content or decisions [51] to specific actors. These capabilities not only enable legal accountability and enforcement but also help protect victims and maintain public trust in AI applications.

Key opportunity areas for policy innovations include:

• Global Policy Innovation and Licensing Standards

As AI becomes a shared global infrastructure, the opportunity exists to coordinate international standards for model licensing, risk classification, and safety thresholds [65]. Southeast Asia and the U.S., with their respective regulatory traditions and technological strengths, can work together to pilot cooperative frameworks that balance innovation with public protection. Joint development of certification schemes and open benchmarking could further encourage trust and interoperability across borders.

• AI Risk Taxonomies and Tiered Regulation

Establishing risk-based AI classification systems allows policymakers to tailor regulatory obligations based on use-case severity. Low-risk applications (e.g., grammar correction) may require minimal oversight, while high-risk areas (e.g., autonomous weapons, hiring systems, or medical diagnostics) warrant stringent licensing, testing, and monitoring. Tiered approaches can help balance innovation and risk mitigation, a need echoed across both U.S. and Southeast Asian contexts.

• AI Incident Reporting and Safety Disclosure Requirements

Creating standardized protocols for AI incident reporting, which is similar to aviation safety or cybersecurity [15], can foster transparency and rapid response to harm. Responsible AI systems can support automated logging and alerting, enabling policymakers and researchers to learn from failures and close safety loopholes early.

• Regulatory Sandboxes for Responsible Innovation

Countries and regions can promote responsible AI by supporting regulatory sandboxes, where companies and researchers can test new AI applications under real-world conditions with temporary regulatory waivers and close oversight. These environments encourage safe experimentation, rapid feedback loops, and public-private collaboration in governance design.

• Licensing AI Development Platforms, Not Just Models

Beyond licensing deployed models, there's an opportunity to license or certify AI development environments, e.g., cloud-based APIs, foundation model hubs, or agent-building platforms. This aims to establish upstream safeguards and embed responsible defaults during model development.

• Cross-border Data Governance and Trust Frameworks

Responsible AI benefits from interoperable data governance frameworks, especially when training data and AI systems span national borders. Southeast Asia and the U.S. could co-develop trusted data-sharing agreements, privacy-preserving protocols, and harmonized accountability standards, which can form a shared baseline for responsible global AI practices.

• Third-Party Auditing Ecosystems

There is a growing opportunity to develop an independent ecosystem of AI auditors and evaluators, such as nonprofits, academic centers, or certified firms tasked with assessing AI systems for safety, fairness, and transparency. Regulatory regimes could mandate external auditing before high-risk AI systems are deployed in public-facing roles.

• Dynamic Policy Toolkits and AI Governance Playbooks

Policymakers need agile tools to keep pace with fast-moving AI developments. A major opportunity lies in developing modular AI governance toolkits, such as risk assessment templates, best-practice checklists, and impact reporting frameworks. These can be shared across nations and updated iteratively based on empirical evidence.

• Public Engagement and Participatory Policy Design

Responsible AI policy must reflect societal values. There is a unique opportunity to foster inclusive public deliberation through citizen panels, community consultation, and open comment periods on AI governance proposals. Southeast Asia's diverse sociopolitical contexts offer valuable insights into localized ethical priorities that should shape AI regulation.

1.3.2 Understanding and Control

As AI systems grow more powerful and autonomous, ensuring that their behavior is both understandable and controllable has become an urgent research priority. Participants emphasized that responsible AI development hinges not only on improving raw performance, but also on advancing our ability to interpret, guide, and constrain these systems in ways that align with human values, societal norms, and legal frameworks.

• Interpretability and Value Alignment

The inner workings of large AI models remain largely opaque, making it difficult to predict how training data influences downstream outputs. This lack of interpretability poses serious risks when models are deployed in high-stakes domains [37]. There are research efforts focused on understanding model internals [78-81], such as attribution techniques [78], data influence methods [79], and transparency tools [75-77, 81]. They can help practitioners and auditors trace outputs to training inputs, diagnose failure modes, and evaluate bias [147]. Moreover, as general-purpose AI systems increasingly serve multilingual and multicultural user bases, the challenge of value alignment grows more complex. Aligning model behavior with diverse cultural expectations requires moving beyond one-size-fits-all ethics to design mechanisms that can adapt to context-specific values.

• Guardrails and Control Mechanisms

With the growing autonomy of AI systems, particularly those capable of code generation, decision-making, and agent-like behavior, the risks of unintended or harmful actions increase. As a result, we may need more robust guardrails that can monitor, constrain, and intervene in AI behavior, especially during deployment. These include sandbox environments**,** runtime monitors, and externalized control layers that can act as safety valves when AI outputs deviate from acceptable parameters [15]. Defining what constitutes a "guardrail" remains an open research question. It involves not only technical constraints but also policy decisions about who sets the boundaries, how they are enforced, and under what circumstances intervention is necessary.

• Controllability through Design

We should explore the potential for smaller, more modular models that, while potentially less powerful than state-of-the-art systems, offer advantages in transparency, debuggability, and controllability [39]. We can look into scaling laws for controllability design, to understand what types of attributes do successfully scale from smaller models to larger ones. In parallel, there is a growing interest in prompt engineering**,** control knobs**,** and external tools that give end-users more control over model outputs. Embedding such control structures at the architecture or training level, rather than as post-hoc filters, may offer more reliable and accountable outcomes.

• Scientific Inquiry into Model Behavior

The community is beginning to treat large-scale AI models as scientific objects, such as complex systems that must be studied rigorously using both empirical investigation and formal analysis

[11]. This includes not just performance benchmarking, but deeper studies into the causal mechanics of learning, generalization, and error propagation. For instance, AI models trained for drug discovery and other scientific applications offer enormous promise for innovation, but also pose risks if their discoveries cannot be verified, reproduced, or interpreted by human experts. A science-of-AI approach could help bridge this gap by establishing frameworks for validating discoveries and identifying potentially dangerous misapplications.

- *Incorporating Constraints during Training*

Techniques like regularization using physics-based laws [138, 139], semantic constraints, or structured representations can help make models respect known boundaries and behaviors. These constraints may prevent harmful generalization, encourage ethical reasoning, or simply ground models in physical and logical truths, which contribute to both performance and responsibility.

Taken together, these insights indicate that a responsible AI ecosystem requires not just stronger models, but more interpretable, controllable, and constrained systems. Whether through design choices, scientific investigation, or the implementation of robust external guardrails, the future of responsible AI depends on our collective ability to understand and guide the systems we build.

1.3.3 Education and Public Awareness of Responsible AI

As artificial intelligence (AI) systems grow increasingly capable and pervasive, fostering public understanding and cultivating a collective sense of responsibility have emerged as critical components of responsible AI development. The risks associated with AI are not confined to speculative or future-oriented scenarios. Rather, they are already manifesting in the present, through financial fraud, the spread of misinformation, psychological manipulation, and broader societal disruption. As such, educating both AI practitioners and the general public is not a supplementary task but a foundational pillar in shaping the societal trajectory of AI.

A recurring theme throughout the discussions was the indispensable role of education and public awareness. As AI technologies become more deeply integrated into everyday life, it is imperative for individuals to possess a foundational understanding of how these systems function, their inherent limitations, and the potential avenues for misuse. Early integration of algorithmic literacy and data ethics into educational curricula may need to be considered.

One of the key philosophical questions raised concerned the notion of whether AI can or should be considered "responsible." While frameworks such as the ACM architecture and the IEEE 7000 and 7010 series provide guidance [37] for embedding ethical considerations into AI design, responsibility ultimately rests with human stakeholders. Developers, users, regulators, and society at large must share accountability for the design, deployment, and consequences of AI systems. Despite their growing autonomy, AI systems should not be treated as moral agents; instead, humans must retain ownership of the ethical outcomes.

Concerns were also raised regarding the maturity and transparency of existing AI systems. Users often overestimate these systems' capabilities and place unwarranted trust in so-called “guardrails,” which may be poorly defined or inconsistently implemented. Therefore, responsible AI design must include mechanisms to acknowledge and communicate system limitations, such as quantifying uncertainty [119, 120, 140], explaining sources of error [39, 121, 141], and adapting in ways that mitigate risk [122, 123]. This reinforces the urgent need for public education that clarifies what AI can and cannot do, and why its outputs may be flawed or misleading.

The societal risks posed by unregulated AI, particularly in the realm of psychological manipulation, were identified as especially urgent. The most dangerous uses of AI may not involve robots or physical devices but rather invisible influences, such as targeted scams that exploit vulnerable individuals or the malicious use of generative models in cases of intimate partner violence [35]. These threats necessitate a dual response: technical safeguards and public awareness campaigns designed to foster resilience against manipulation and exploitation.

Education was also discussed as a proactive form of soft regulation. Early instruction in algorithmic thinking, ethical reasoning, and an understanding of machine learning’s limitations was proposed as a vital countermeasure to the naive or uncritical use of AI tools. The aforementioned examples from Singapore serve as a cautionary tale, illustrating that digital fluency does not equate to ethical or technical literacy. To address this gap, integrating responsible AI content into both formal education and informal outreach initiatives could be considered.

At a broader level, the discussions called for a systemic shift toward sustainable AI development practices. Participants expressed concern over the prevailing industry paradigm, which often prioritizes performance gains through ever-larger datasets and increased computational power, frequently at the expense of data quality, environmental sustainability, and ethical oversight. A transition toward a more sustainable development model, grounded in safety, interpretability, and long-term societal benefit, is an urgent necessity.

While legal and regulatory frameworks were acknowledged as important, the consensus was that they alone are insufficient. Effective responsible AI governance must also include self-regulation, adherence to evolving societal norms, and the implementation of technical constraints such as sandbox environments and runtime monitors. The possibility of AI systems contributing to their own evaluation, through self-checking mechanisms and benchmark assessments, was identified as a promising direction for future research and development.

In conclusion, education and public awareness are foundational to responsible AI. Ensuring that individuals at all levels, such as developers, policymakers, and end users, are empowered to understand, critique, and shape AI technologies is essential not only for mitigating harm but also for guiding these technologies toward inclusive, equitable, and democratically aligned outcomes.

1.3.4 Interdisciplinary Collaboration

The responsible development and deployment of artificial intelligence (AI) extends beyond technical sophistication or computational power; it requires sustained, interdisciplinary collaboration. The workshop highlighted the imperative of engaging not only computer scientists and engineers, but also experts from diverse domains such as the social sciences, psychology, law, medicine, philosophy, and the physical sciences. Addressing the societal impacts of AI, such as fairness, accountability, transparency, and long-term social consequences, necessitates an integrated approach that draws upon multiple perspectives and epistemologies.

A key area of emerging interdisciplinary research is the incorporation of responsibility into reinforcement learning frameworks. Scholars are investigating how human preferences, ethical norms, and social values might be systematically encoded into the reward structures that guide AI training [50]. While this research avenue introduces complex questions regarding whose values are prioritized and how normative frameworks are formalized, it also offers a promising path toward the development of AI systems that are more aligned with human intentions and ethical expectations.

Another frontier at the intersection of disciplines involves the creation of performance benchmarks that evaluate AI in the context of human-AI collaboration. Rather than measuring AI capabilities in isolation, this paradigm emphasizes augmentation, which assesses how AI systems enhance, complement, or co-evolve with human decision-making and creativity [12]. Such metrics are essential for incentivizing the development of AI tools that support human judgment rather than supplant it, reinforcing the vision of cooperative, human-centered intelligence.

Beyond research, interdisciplinary collaboration also plays a critical role in shaping public policy and governance. As governments and regulatory bodies grapple with the accelerating pace of AI innovation, there is a growing need for empirically grounded evidence and scientifically credible risk assessments to inform regulatory decisions. Interdisciplinary academic teams are uniquely positioned to offer such guidance, ensuring that policymaking is informed by both technical realities and an understanding of social, ethical, and legal implications [37].

We believe that responsible AI cannot be achieved within disciplinary or institutional silos. Meaningful progress requires ongoing dialogue across sectors, the co-creation of standards and norms, and the development of integrated frameworks that span technical innovation and societal considerations [64]. Interdisciplinary partnerships are not merely additive; they are foundational to ensuring that AI technologies serve the public good in a manner that is equitable, transparent, and aligned with shared human values.

2.Development of Responsible AI

2.1 Two Pillars of Responsible AI Development

The development of responsible AI rests on two foundational and interdependent pillars: (1) the technological advancement of AI systems in ways that enhance their safety, transparency, and accountability, and (2) a human-centered approach that cultivates responsible usage through education, engagement, and empowerment. These pillars are mutually reinforcing. Technological safeguards are most effective when users understand and apply them, while human oversight is only meaningful when supported by trustworthy systems [13, 58].

Technological Pillar: Building Robust, Explainable, Trustworthy, and Certifiable AI

To promote responsible AI deployment, it is essential that AI systems are designed with technical features that support safety, reliability, and ethical use. Achieving this requires progress across several key dimensions:

- **Robustness and Safety**: AI models must be resilient to adversarial attacks, data poisoning, and other perturbations [152-155]. This includes stress-testing systems in diverse environments, improving generalization across contexts, and implementing fail-safe or fallback mechanisms in high-stakes applications [11, 34].
- **Explainability and Transparency**: Systems should be capable of generating human-interpretable explanations of their outputs, particularly in domains such as healthcare, finance, and legal decision-making. Efforts to improve model interpretability, such as attention visualization, feature attribution methods, and counterfactual explanations, should be prioritized in both research and deployment [39, 44].
- **Trustworthiness and Accountability**: Building trust requires not only technical transparency but also traceability and auditability. AI systems should include built-in mechanisms to log decisions, flag anomalous behavior, and identify the sources of bias or error. Moreover, developers and organizations must be accountable for model performance and behavior in real-world settings [41, 50].
- **Certification and Standards**: A robust ecosystem of evaluation and certification is needed to guide the development and adoption of responsible AI systems. This includes third-party audits, standardized benchmarks, ethical compliance protocols, and formal methods for verifying safety and fairness. Collaborations with regulatory agencies and international standards bodies (e.g., ISO, IEEE) will be crucial to institutionalizing these processes [12, 64, 156].

Investments in research, toolkits, and open-source infrastructure that operationalize these principles are necessary to ensure that technical innovation is aligned with responsible deployment.

Human-Centered Pillar: Educating and Empowering Responsible Users

Parallel to technical improvements, responsible AI development requires empowering individuals and communities to use AI technologies with awareness, critical thinking, and ethical judgment.

This human-centered approach emphasizes education and cultural transformation across multiple levels [56, 60]:

- **Education for Students and Developers**: Curricula at both secondary and tertiary levels should integrate topics such as algorithmic fairness, data ethics, and socio-technical systems. For developers and researchers, continuing education through certifications, workshops, and interdisciplinary training programs can help embed responsibility into professional practice [61].
- **Public Awareness and Digital Literacy**: Members of the general public should be equipped with foundational knowledge about how AI systems work, what their limitations are, and how they can be misused. Outreach efforts, community forums, and media engagement are vital tools for fostering critical AI literacy and countering misinformation or undue trust in algorithmic outputs [26].
- **Promoting Responsible Use Culture**: Beyond formal education, cultivating a culture of responsibility among users involves shifting incentives and norms. This may include organizational policies that discourage over-reliance on AI systems, interface designs that promote human oversight, and public campaigns that emphasize the role of human judgment in AI-mediated environments [45].
- **Guarding Against Over-Reliance and Misuse**: AI tools should be framed as decision-support systems rather than decision-makers. Interfaces can be designed to highlight uncertainty, offer alternative perspectives, or require user justification for critical decisions. Teaching users when *not* to use AI is as important as teaching them how to use it [13].

Together, these strategies aim to democratize AI understanding, reduce systemic risks of misuse, so that AI technologies serve as tools for augmentation rather than replacement. The success of responsible AI hinges not only on the systems we build but also on the people who design, deploy, and interact with them.

2.2. Certification and Reliability

As AI systems become increasingly embedded in high-impact domains, from healthcare diagnostics and financial services to public decision-making [137]. There is growing interest in developing formal mechanisms to certify their trustworthiness and reliability [105-107]. Similar to established standards in other sectors, such as ISO 42001 for AI management systems, the IEEE Authorized Assessor program, or the USDA Organic certification in agriculture, AI certification aims to provide stakeholders with confidence that these technologies meet defined thresholds for safety, performance, and ethical alignment [20, 37].

Certified AI systems should be explicitly constrained to domains and conditions where their behavior can be systematically verified. Rather than making broad or unqualified claims about general capability, certification efforts should focus on well-bounded use cases. Any tasks with objective evaluation, such as simple mathematical calculations, structured factual queries, or domain-specific classifications, can be included in such certifications. Equally important is the definition of operational boundaries. AI systems should be certified to function within known "working conditions," such as specific input ranges, sensor configurations, or task contexts. This

enables the articulation of bounded guarantees, such as "99% accuracy within a constrained input space," providing clarity on where performance claims are valid and where caution is required.

In many scenarios, it may be more feasible to certify the process by which an AI system is developed and maintained rather than attempting to certify all possible outcomes it may generate in deployment. Process certification involves evaluating the integrity of training data, the reproducibility of the development pipeline, documentation of model parameters, and governance structures for monitoring and updates. While this does not eliminate the possibility of failure, it supports systems built under rigorous, transparent, and accountable conditions. Outcome certification, on the other hand, remains essential in domains with high risk and clearly defined performance metrics, such as medical imaging or autonomous control systems. In such contexts, performance guarantees must be empirically validated and independently audited [33, 148, 149, 151].

To accommodate the diverse risk profiles of AI applications, a multi-tiered certification approach is essential. Low-risk or general-purpose systems may require only baseline assurance, while high-stakes systems in regulated domains demand more stringent oversight, including domain-specific benchmarks, formal verification, and periodic recertification [47]. Certification should remain dynamic, incorporating continuous monitoring, compliance checks, and mechanisms for recourse in the event of harm or failure. A central question is what kinds of AI systems require certification: Should it apply only to corporations, or also to individual researchers like PhD students? Does it cover only "large" models, by parameters, compute, or capability, or also smaller models deployed at scale? The framework must be flexible enough to survive rapid technological change, which suggests focusing on risk and impact rather than size alone. High-risk systems, such as those in healthcare, finance, law, critical infrastructure, or mass-consumer applications, clearly warrant certification, while low-risk applications may require minimal oversight. Certification should primarily apply to major developers and corporate entities, with academic projects generally exempt unless scaled for public release. Open-source models present unique challenges, requiring voluntary certification labels or responsible release protocols.

Equally important is what compels creators and users to adopt certified models. In highly regulated sectors like medicine, banking, or law, compliance can be enforced through existing legal frameworks that bind professionals. But what compels general users of AI assistants or tutors to prefer certified models, and what incentivizes creators of consumer-facing AI to seek certification? If certification is optional in such cases, adoption may lag unless paired with strong incentives and enforcement mechanisms [43]. Possible approaches include liability frameworks that hold developers accountable for harm caused by uncertified models, platform-level enforcement through app stores and cloud providers, and trust labels that signal reliability to consumers. Detection of rogue AIs will require technical measures like watermarking and fingerprinting, combined with compliance audits.

Ultimately, certification serves not only as a technical safeguard but also as a tool for public trust and governance [40]. Considering companies often guard their datasets, training parameters, and other proprietary assets, it makes full transparency difficult. This gap could be addressed by establishing a trusted third-party certification body, or by developing new technological solutions that enable meaningful transparency while safeguarding trade secrets. As global interest in

certified AI continues to grow, the development of standardized, transparent, and scalable frameworks will be essential for aligning technological advancement with societal expectations. These efforts need be pursued in concert with legal regulation, technical standards bodies, and interdisciplinary stakeholder engagement to fit for purpose.

2.3 Working Condition as a Foundational Concept

A foundational concept in the responsible development of artificial intelligence is the idea of defining clear "working conditions": the specific environments, input ranges, and task parameters under which an AI system is expected to perform reliably. This concept, grounded in traditional engineering practices, introduces a critical layer of operational discipline to AI system design [25]. Just as mechanical or electrical systems are only certified to function within known tolerances, AI systems should be evaluated and deployed under similarly bounded conditions.

Defining working conditions enables a shift away from the pursuit of overly generalized systems toward more targeted, modular AI models that are optimized for well-defined use cases. Smaller, context-specific models, when confined to clear operational domains, may in fact offer greater trustworthiness and accountability than large-scale general-purpose systems operating across unconstrained and unpredictable environments. In high-stakes applications such as healthcare, education, and critical infrastructure, such bounded reliability is more valuable than nominal generalizability.

By explicitly stating and enforcing working conditions, AI developers and deployers can improve system reliability, support certification, and set clearer expectations for users. These conditions serve as guideposts for both technical performance and ethical use, informing where and how AI can safely operate, and where human oversight or additional controls are necessary. In this context, generalization is no longer presumed to be limitless but must be justified and validated before extending system scope.

This framing opens up significant research and policy opportunities [12]. From a research perspective, several key questions remain unanswered: Can AI systems be made modular and downgradable, allowing them to operate in controlled, constrained settings suitable for training, education, or lower-risk applications? How should responsible-use boundaries be defined and enforced, especially in open or semi-open environments? And what new metrics are needed to balance raw performance with ethical and context-sensitive deployment?

These technical challenges are mirrored by emerging policy needs. One urgent step is the formal definition of acceptable working conditions, both at the model and application level. This could include specifying operational domains in regulatory guidance, procurement standards, or platform governance rules. Governments and international bodies should support the creation of third-party certification authorities that can assess whether AI systems meet these operational standards. In parallel, AI-generated content, particularly in education, software development, and research publications, should be accompanied by clear disclaimers and traceability mechanisms, ensuring that users are aware of the provenance and limitations of such outputs.

Adopting working conditions as a central design and governance principle also encourages a more gradual and responsible pathway to AI scaling. Rather than releasing models into open-ended, high-risk domains without constraint, developers can expand their systems incrementally, validating each new deployment context through empirical testing and stakeholder review [55]. This not only builds public trust but also aligns technical progress with evolving social expectations and ethical norms.

In sum, working conditions provide a foundational framework for making AI both safer and more accountable. By operationalizing the boundaries of responsible use, they bridge the gap between engineering rigor and societal governance—anchoring AI development in clearly defined contexts, and enabling both research innovation and regulatory action to proceed in a principled, aligned manner.

2.4 Performance vs. Responsibility Tradeoff

A central tension in contemporary AI development lies in balancing the pursuit of performance with the imperatives of responsibility. The research community has long prioritized state-of-the-art (SOTA) results, often measured through benchmarks that reward raw accuracy, scale, or speed. While this has driven rapid progress in model capability, it has also created incentives that may conflict with responsible AI practices. The most powerful models, those trained on vast datasets with billions of parameters, are not necessarily the most trustworthy, transparent, or ethically aligned. In many cases, a less powerful but well-bounded model may be better suited for deployment, particularly in sensitive or high-stakes environments.

This tradeoff is not merely theoretical. Large, general-purpose models frequently operate in open-ended contexts where their behavior is difficult to predict or verify. These systems often lack clear boundaries of applicability, offer limited transparency about how outputs are generated, and can be prone to biased or misleading responses. In contrast, smaller, modular, or domain-specific models, designed with narrowly defined working conditions, may sacrifice some degree of generality or benchmark performance, but offer higher reliability, interpretability, and accountability. Such models are often more amenable to certification, monitoring, and responsible integration into human workflows [46].

Addressing this tradeoff requires a deliberate rebalancing of values within the AI ecosystem. Performance should no longer be defined solely by technical metrics, but by a more holistic evaluation of how well a system aligns with ethical principles, social needs, and user expectations. Metrics that account for user trust, robustness under defined conditions, and cognitive support rather than substitution, must become standard components of model evaluation [56]. Without such recalibration, the field risks developing increasingly powerful systems that are functionally impressive but socially irresponsible.

This challenge also opens up a range of research and policy opportunities. On the research front, open questions include how to design modular and downgradable AI agents for use in controlled environments such as education or training, where cognitive engagement is prioritized over

automation. Further inquiry is needed into how to define and enforce responsible-use boundaries, particularly in settings where AI tools may be repurposed in ways that diverge from their original intent. Most critically, the field must grapple with the development of new evaluation metrics that capture the tradeoff between raw performance and responsible use, including metrics for explainability, user understanding, and unintended consequences.

In terms of policy, a number of concrete actions were identified during the workshop. Regulators and standards bodies should begin by defining acceptable working conditions for AI systems in specific domains, providing clarity on where and how models can be safely deployed. The establishment of third-party certification bodies, which is analogous to those in medicine, engineering, or food safety, will be essential for assessing compliance with these standards. Additionally, policies requiring disclaimers and traceability for AI-generated outputs, especially in education, software, and publishing, can help mitigate the risks of uncritical reliance and misinformation [51].

Ultimately, resolving the performance–responsibility tradeoff does not imply halting innovation, but rather redirecting it toward long-term societal benefit. This includes creating models that not only perform well under ideal conditions but also behave predictably, ethically, and transparently in the real world [23]. Shifting research incentives, publication criteria, and funding priorities to reflect these values is a necessary step in aligning AI progress with public interest.

3. Evaluation of Responsible AI

3.1 What's Going Well and Where Are the Gaps?

Efforts to evaluate responsible AI have seen promising advances in certain areas. Developers have implemented guardrails that help prevent AI systems from responding to harmful or unsafe queries, using moderation layers, prompt filtering, and refusals to reduce toxic, biased, or misleading outputs [108-110]. Significant progress has also been made in identifying and mitigating hallucinations when AI systems generate factually incorrect or fabricated content [111-113]. This is particularly important in high-stakes domains like medicine [82], where hallucinations can have serious consequences. Fortunately, the presence of well-established medical guidelines provides a concrete basis for evaluating factual alignment and appropriateness in this domain.

One notable development in responsible AI evaluation is the use of large language models to evaluate other LLMs [114-116]. This “LLM-as-a-judge” paradigm offers scalability and efficiency, allowing for automated assessments of coherence, helpfulness, and safety without requiring exhaustive human annotation. It holds promise for accelerating the feedback loop in model development and deployment, particularly in rapidly evolving applications. However, this approach is still in its early stages, and critical challenges remain. Chief among them is the question of reliability: since many LLMs share similar architectures, training data, and biases, using one model to evaluate another can lead to circular reasoning. This risks reinforcing shared blind spots, such as subtle forms of bias, misinformation, overconfidence, or reasoning errors, rather than exposing them. Without robust external validation or grounding in objective standards, such evaluations may provide a false sense of rigor. The lack of standardized benchmarks and the

reliance on static, culturally limited datasets further compound these concerns. As a result, the current landscape of LLM-based evaluation remains fragmented and ad hoc—described by many as "the wild west"—with inconsistent metrics, opaque methodologies, and little consensus on what constitutes meaningful success in responsible AI [56].

More fundamentally, current evaluation efforts lack a shared world ontological model, namely a structured, agreed-upon representation of commonsense knowledge, logic, and real-world relationships that AI outputs can be verified against. Without such a grounding framework, evaluations of factuality, reasoning, and ethical alignment are often ad hoc, dataset-specific, or limited to surface-level checks. Traditional knowledge representation tools, such as symbolic reasoning, formal logic, and curated knowledge bases, have not yet been fully integrated into modern LLM evaluation pipelines. This leaves a significant gap in verifying whether an AI's output truly reflects valid reasoning [39] or conforms to established truths across disciplines.

Another key challenge in LLM evaluation is data leakage and contamination, whether unintentional or intentional. Benchmarks often overlap with training data, leading to inflated scores that misrepresent true generalization [46]. This issue is further complicated by proprietary models, which do not disclose their training corpora, making it difficult to assess contamination risks. Reliable evaluation thus requires stricter data provenance checks and safeguards against both inadvertent and deliberate leakage. Moreover, many existing evaluation datasets, such as those used to measure toxicity, bias, or fairness, are static and often fail to capture cultural nuance, evolving norms, or real-world complexity. As a result, they offer limited generalizability and robustness. Overall, while parts of responsible AI evaluation are maturing, the field still lacks the foundational infrastructure, both conceptual and empirical, needed for rigorous, reliable, and context-aware assessment [16].

Emerging domains such as embodied AI, where AI agents interact with the physical world through sensors, actuators, or robotic platforms, present new evaluation challenges and opportunities. On one hand, embodied AI benefits from clear grounding in physical reality, allowing for objective evaluation based on task completion, safety constraints, and measurable real-world outcomes. For example, robotic manipulation or navigation tasks can be assessed using concrete success metrics like precision, energy efficiency, or compliance with human safety standards. However, current evaluation approaches are often limited to technical functionality and overlook broader social and ethical dimensions of physical interaction, such as respecting personal space, adapting to cultural norms, or ensuring inclusivity. Standardized benchmarks for these aspects are still lacking.

Similarly, agentic AI, systems that exhibit persistent goals, planning, memory, and autonomous decision-making, raises novel concerns for responsible evaluation. These agents can act over extended time horizons, interact with users repeatedly, and adapt dynamically to complex environments. Progress has been made in building simulation platforms and multi-agent environments (e.g., virtual economies, strategy games, or embodied simulators) to support long-term evaluation of planning, goal alignment, and cooperation. However, these environments often abstract away critical real-world variables, such as human values, long-term accountability, and unintended consequences. One notable gap is the absence of protocols for evaluating *value alignment over time*: how do we ensure that an agent's evolving behavior continues to reflect human intentions as it learns or generalizes? Another unresolved issue is evaluating responsibility

in multi-agent settings, where outcomes arise from complex interactions between agents and humans. There are already companies advertising LLMs as synthetic human users for testing products [117, 118]. However, these LLMs are not sufficient to model the full extent of human behavior even in narrow testing conditions. LLMs are still too "malleable" in the sense that they defer to whatever prompt they are provided with and cannot stay consistent to the original context (the description of the user they are simulating). Existing tools are not yet equipped to assess distributed accountability, emergent ethical dynamics, or strategic deception, challenges that will become increasingly salient as AI agents grow more autonomous and socially embedded.

Another persistent gap in responsible AI evaluation lies in our limited understanding of the performance thresholds required for human trust and adoption. Contrary to the assumption that achieving human-level performance is enough, research in human-computer interaction and cognitive science indicates that AI systems often need to *outperform* humans significantly, both in accuracy and consistency, before users are willing to trust and rely on them, especially in domains like healthcare, transportation, or legal decision-making. Even minor unpredictable errors can severely erode trust, particularly if they diverge from human reasoning patterns or are difficult to explain. Encouragingly, interdisciplinary efforts are starting to quantify these trust thresholds across different tasks and populations. However, a universal or calibrated benchmark for "trustworthy performance" remains elusive. This presents a serious challenge: traditional performance metrics (e.g., accuracy, F1 score) may not fully capture user perceptions of competence, transparency, or fairness [13, 57]. To bridge this gap, evaluation frameworks will need to go beyond technical benchmarks and incorporate human-centered criteria such as interpretability, consistency, and value alignment.

To advance the field of evaluating responsible AI, there is an urgent need for transparent, reproducible, and domain-specific evaluation protocols, ones that assess AI behavior not only in controlled settings but also in terms of real-world social, cultural, and ethical implications. Only by integrating technical performance with human values and lived realities can we build truly responsible and trustworthy AI systems.

3.2 Research Gaps and New Questions

Responsible AI systems become increasingly integrated into high-impact domains, ranging from clinical decision support to creative content generation and educational tools. As a result, new challenges emerge that traditional evaluation metrics are ill-equipped to handle. This calls for a rethinking of not just *what* we evaluate in AI systems, but *how* and *why* we evaluate it [29]. Addressing these challenges demands filling several key research gaps and confronting critical new questions that cross disciplinary and sectoral boundaries.

A foundational research gap involves better understanding the secondary effects of AI systems, particularly in terms of user experience, trust, and productivity. While much of the current evaluation effort focuses on primary outcomes such as accuracy or fairness, users increasingly face a deluge of AI-generated outputs that may be irrelevant, distracting, or even misleading. This is especially problematic in creative or open-ended domains like educational content or consumer entertainment, where quality control is more subjective and harder to enforce [23]. Research is needed to quantify these "soft harms", such as wasted time, decision fatigue, and erosion of

confidence in human judgment, and to develop evaluation metrics that capture such indirect yet significant impacts.

A second, critical set of questions arises from the fundamental distinction between prediction and generation tasks. Predictive AI systems, used for applications such as medical triage, financial forecasting, or recommendation engines, are typically evaluated on criteria like precision, recall, and calibration. Generative AI systems, however, pose novel challenges due to their open-ended, subjective, and often context-sensitive outputs. For example, generating a music track or a synthetic educational video requires not only technical fluency, but cultural relevance, emotional resonance, and ethical appropriateness. How should evaluation frameworks adapt to the unique risks and responsibilities tied to creative generation? And how do we draw boundaries between what is considered "responsibly creative" and what is potentially manipulative or harmful, particularly for vulnerable groups like children or patients?

Another pressing research gap concerns responsibility attribution in human-AI collaborations. Increasingly, AI-generated outputs, like clinical summaries, legal drafts, or student feedback, are adopted without significant human revision. This blurs the line of accountability between the AI system and the human expert, raising complex questions about liability, trust, and oversight [99]. How often are AI outputs used “as is”? What types of users are more likely to rely on AI suggestions without verification? Systematic empirical studies are needed to analyze human-AI interaction data, including modification rates, trust thresholds, and decision-making pathways, to inform more robust responsibility-sharing frameworks.

Furthermore, the evaluation of creativity in generative systems remains an underexplored area with significant societal implications. Current proxies, such as output diversity or linguistic novelty, do not capture whether creative outputs are meaningful, contextually appropriate, or aligned with broader human values. In some cases, seemingly novel outputs may replicate or amplify harmful stereotypes or misinformation. Particularly in domains like health education or mental wellness, AI-generated content must be evaluated not only for creative merit but also for psychological safety and cultural sensitivity [95]. This demands interdisciplinary research across fields such as cognitive science, media studies, and behavioral psychology.

To address these challenges, the field must shift toward process-based evaluation models, inspired by regulatory science frameworks such as those used by the FDA or USDA. These models emphasize the importance of oversight across the entire AI lifecycle: from data collection and model training to deployment and monitoring, rather than focusing exclusively on output quality. For AI, this could involve staged evaluation pipelines, pre-deployment audits, stress-testing under edge cases, and post-market surveillance. Importantly, these processes must be transparent, reproducible, and adaptable, especially as AI systems continue to learn and evolve in deployment. Research is needed to define what such a process-based evaluation should look like for AI, what documentation and artifacts should be required, and how independent verification and public accountability can be operationalized.

Finally, addressing these gaps will require new data infrastructures and interdisciplinary methods. For example, building repositories of human-AI interaction logs, especially in expert domains like medicine or education, can help reveal how decisions are made, revised, or ignored in practice.

Similarly, the development of socio-economic evaluation frameworks is essential to assess how AI systems redistribute value, labor, and risk across different populations. Just as public health agencies evaluate both individual-level and systemic effects of interventions, responsible AI evaluation must account for both micro- and macro-level impacts.

In sum, responsible AI evaluation is no longer a purely technical challenge. It is a complex socio-technical endeavor that calls for deep integration of methods from computer science, social science, humanities, and regulatory policy. By addressing these research gaps and embracing new questions, the field can move toward evaluation frameworks that are not only rigorous and actionable but also aligned with the broader societal values and ethical imperatives that define truly responsible AI.

3.3 Connecting Research to Practice and Policy

A central challenge in responsible AI is that evaluation research informs not only academic discourse but also the decisions and practices of those most affected, like practitioners deploying AI, auditors and regulators overseeing compliance, and communities impacted by these systems. Despite advances in responsible AI frameworks, dissemination often remains fragmented and inaccessible, with research outputs typically technical and aligned with academic priorities rather than real-world needs. Researchers can bridge this gap by engaging stakeholders as co-designers of evaluation tools, priorities, and outcomes. Collaboration with advocacy groups and community coalitions is particularly important, as they bring contextual knowledge and trusted relationships. This engagement helps align evaluation frameworks with lived experiences and supports development of concrete, actionable outputs such as audit checklists, risk assessment templates, transparency scorecards, and lifecycle monitoring protocols, designed to integrate with regulatory and organizational workflows.

From the perspective of government agencies, regulators, and other decision-makers, trust relies on credible, comprehensible, and contextually relevant evaluation evidence. Standardized, validated tools, independent audits, and third-party certifications provide documentation beyond marketing claims [51, 37]. Accessible summaries for non-technical stakeholders clarify risks, trade-offs, and mitigation strategies, supporting accountability and public trust. Rigorous evaluation also requires meaningful engagement with impacted communities, with evidence that feedback influenced design decisions, evaluation metrics, and deployment strategies. For example, an AI system for education should document how teachers, students, and parents informed fairness criteria or content moderation policies. Embedding participatory design and evaluation as core principles strengthens legitimacy, relevance, and trustworthiness, helping AI development better reflect accountability, equity, and long-term public interest.

3.4 Challenges, Opportunities, and Action

The rapid advancement and widespread deployment of AI systems have introduced unprecedented complexity in evaluating their responsibility, safety, and societal impact. As models become more capable, autonomous, and domain-diverse, the field of responsible AI evaluation faces a growing set of methodological, infrastructural, and epistemic challenges. At the same time, new opportunities are emerging, from synthetic evaluation agents to interpretability advances. They

can guide more robust and scalable evaluation approaches. This section synthesizes key challenges identified during the workshop, highlights corresponding opportunities, and outlines priorities for immediate action.

3.4.1 Evaluating AI at Scale and Across Domains

Challenge: AI systems are being deployed across a wide range of domains, from education to medicine to creative arts, each with unique risk profiles and evaluation needs. Evaluating these systems at scale and across disciplinary boundaries is inherently difficult, especially when domain expertise is required to understand appropriate uses and harms.

Opportunity: One promising solution is the use of *AI-for-evaluation*. Training synthetic agents on domain-specific knowledge to act as evaluators of other AI systems. However, such evaluators must be designed to be orthogonal and independent from the systems they assess, avoiding contamination or bias.

Priority for Action: A key next step is to identify high-impact use cases such as AI in healthcare or education and develop domain-specific methods, agent architectures, and protocols for synthetic evaluation. These agents could perform diagnostic probing, simulate user experiences, or assess alignment with normative goals, enabling a scalable yet context-sensitive approach to responsible AI evaluation [54].

3.4.2 Open-Ended AI and Benchmarking the Unknown

Challenge: Many generative and conversational AI systems operate in open-ended domains where factual labels may be insufficient or unavailable. Evaluating such models, especially regarding their capacity for unintended harms, hallucinations, or creative misuse, is fundamentally challenging when ground-truth labels do not exist.

Opportunity: The field can look to the success of standardized benchmarks in adjacent areas (e.g., ImageNet or GLUE) and adapt these strategies to open-ended tasks. Additionally, modeling evaluation after regulatory processes, such as those used by the FDA, could create structured pipelines for risk assessment and public accountability [67].

Priority for Action: Responsible AI efforts must prioritize the creation of *systematic, multi-dimensional benchmarks* for open-ended AI outputs. These benchmarks should incorporate not only correctness but also societal impact, user trust, and robustness. At the same time, a regulatory-style *evaluation process model*, with defined stages, inspection points, and feedback loops, should be developed to assess AI systems over their lifecycle.

3.4.3 Measuring Secondary Effects and Societal Impact

Challenge: Many AI applications now play roles in guidance, counseling, and decision support, which may lead to *secondary effects*—changes in user behavior, perception, or social interaction—that are difficult to capture using traditional model performance metrics.

Opportunity: While some technical model benchmarks exist, much less attention has been paid to the human and societal outcomes of these systems. This presents an opportunity to study specific use cases—e.g., AI in mental health support, educational feedback, or legal advice—and derive cross-domain metrics for indirect impact.

Priority for Action: Develop interdisciplinary research programs to understand how different use cases manifest secondary effects, and identify *generalizable metrics and reporting protocols* for them. These metrics could include user trust dynamics, decision dependency, cognitive offloading, or societal equity outcomes, and should be incorporated into standard evaluation pipelines.

3.4.4 Defining Trust Thresholds and Risk Tolerance

Challenge: There is currently no consistent definition or threshold for what constitutes "sufficiently responsible" AI. This ambiguity hinders evaluation, procurement, and regulation, and leaves public institutions without a clear standard for acceptable risk.

Opportunity: Structured engagement with stakeholders—including users, regulators, domain experts, and advocacy groups—can clarify the conditions under which AI systems are considered trustworthy, and help define the criteria for deployment-readiness.

Priority for Action: Initiate *stakeholder-driven consultations* to define trust thresholds across various domains. These efforts should not only capture technical expectations but also include ethical, cultural, and legal perspectives on acceptable trade-offs, ultimately feeding into guidelines for procurement and policy.

3.4.5 Addressing the "Unknown Unknowns"

Challenge: Harmful outputs, unintended uses, and emergent behaviors remain difficult to anticipate during pre-deployment evaluation. As AI systems grow more complex, identifying and mitigating these "unknown unknowns" becomes both more critical and more elusive.

Opportunity: Techniques *such as uncertainty quantification* [120,121]*, red-teaming [27],* privacy leaks testing *[34, 35], and adversarial testing [152,153,155] o*ffer ways to stress-test systems under diverse conditions and surface hidden failure modes. In parallel, articulating intended and unintended use cases [15,50,44], which is similar to drug labeling, can enhance transparency.

Priority for Action: Build *uncertainty estimation* and *adversarial evaluation* into standard evaluation pipelines. Additionally, require system developers to submit formal descriptions of expected and prohibited use cases, potential failure cascades, and mitigation strategies.

3.4.6 Interpretability and Grounding in Generative AI

Challenge: Generative AI models often produce content that appears plausible but lacks grounding in underlying data or context. Evaluating interpretability, attribution, and input-output alignment remains a technical and conceptual challenge.

Opportunity: Advances in *attribution techniques*, *explanation generation*, and *grounding verification* are opening up new possibilities for assessing how faithfully generative models reflect their source material [81].

Priority for Action: Establish *standardized benchmarks for interpretability and grounding*, including provenance tracking and justification scores. These benchmarks should be domain-sensitive and adaptable to multiple modalities (e.g., text, images, video) to make broad applicability.

4.Responsible AI — Global Alignment, Local Action

4.1 Regional Insights from Southeast Asia Countries

The landscape of responsible AI development and evaluation varies considerably across countries in South and Southeast Asia and the broader Indo-Pacific region. Each country presents unique socio-cultural contexts, infrastructural realities, and policy priorities that shape how AI can be responsibly researched, deployed, and governed. Understanding these regional nuances is critical to designing evaluation frameworks and AI systems that are both effective and equitable. Below we summarize key insights from five representative countries—India, Singapore, Vietnam, Thailand, and Australia—highlighting common themes and distinctive opportunities.

IN India

India's vast diversity presents both a challenge and an opportunity for responsible AI. A major challenge is the lack of data that is essential to designing AI-based technology responsibly where these could cater to bespoke regional problems in healthcare, law etc. Another challenge is in training and education to develop skills at multiple levels [143]. Education and curriculum changes at high school level are also being developed and implemented [143, 144, 145]. All of these are key thrust areas of the India AI Mission [1, 145], which has launched a comprehensive ecosystem aimed at democratizing AI infrastructure, improving data quality, and sharing compute resources to foster responsible AI [144] development nationwide [9]. Focused applications in autonomous driving [2], remote healthcare [3], and smart agriculture [4, 5] not only address India's priorities but also hold promise for regional adoption across Southeast Asia and Australia.

VN Vietnam

Vietnam has seen strong governmental support for AI since 2020, characterized by initiatives like "AI for Vietnam" and collaborations with expatriate scientists to boost local research capacity. A particular focus lies in developing regional language models tailored to the Southeast Asian linguistic landscape, often in partnership with Singapore and neighboring countries. This underscores the critical need for regional collaboration to create AI solutions that reflect shared cultural and linguistic identities across Southeast Asia. Vietnam's emphasis on open-source models also points to a strategic commitment to inclusive, community-driven AI development.

TH Thailand

In Thailand, there are AI embedded in many platforms/systems including medical diagnosis developed from industrial and education sectors. However, AI applications in agriculture stand out as a major opportunity with potential for cross border technology transfer within Southeast Asia. National programs promoting inclusive education and skills training aim to build AI literacy among adults and youth, preparing a workforce equipped to engage with emerging technologies. AI innovation in applications and in all sectors is also promoted by the Thai government. However, challenges remain, particularly in expanding "Responsible AI" awareness in all sectors, a common issue across the region. Thailand's experience highlights the importance of tailored training initiatives and infrastructure investments to bridge digital divides.

SG Singapore

Singapore's approach to AI development is grounded not merely in technological advancement but in strong cultural and structural foundations. A long-term planning ethos, high-quality governance, and a globally respected legal system create a trusted environment for innovation [53, 65, 66].. Singapore places particular emphasis on high secondary education standards, though there is growing recognition of the need to expand AI-related PhD capacity. As a cross-cultural hub, Singapore positions itself as a neutral, high-trust global platform where East meets West, enabling international AI collaboration across sectors like manufacturing and healthcare [134, 135]. Additionally, it aspires to global leadership in combating AI-driven misinformation by leveraging interdisciplinary R&D and its unique trust profile to act as a clearinghouse for truthful, responsible AI content [136].

AU Australia

Australia's AI landscape is marked by a strong commitment to respecting Indigenous communities and cultural sensitivities, ensuring that AI solutions align with local values and governance structures. Efforts to improve technology access in rural and remote regions leverage AI-powered tools, such as portable medical imaging devices, to reduce disparities between urban and regional healthcare delivery. Australia also recognizes opportunities for AI to address national-scale challenges including telehealth expansion and climate change mitigation, positioning AI as a strategic asset for large, diverse geographic contexts.

4.2 A Shared Responsibility: Building Trust and Accountability in a Global AI Landscape

The global rise of artificial intelligence (AI) has prompted urgent reflection on the appropriate scope and structure of its governance. Should responsible AI be regulated primarily through national frameworks, reflecting each country's unique legal, cultural, and technical circumstances? Or is there a need for more unified global oversight, given the inherently transnational nature of AI's development, deployment, and impact? This section explores these questions and emphasizes

the emerging consensus: while implementation must be context-sensitive, responsible AI is a shared global responsibility requiring international coordination and trust [6].

AI's capacity to influence economies, political processes, education systems, healthcare, and public discourse is no longer constrained by geography. Much like climate change, the externalities of AI do not recognize national borders. Participants from across the Asia-Pacific region, including representatives from Vietnam, Australia, and Singapore, echoed the importance of building globally interoperable norms while attending to local realities. Vietnam's national AI strategy illustrates how countries are actively shaping AI policy with global awareness, while Australia's efforts to harmonize its standards with those of the European Union and the United States demonstrate the potential for alignment across democratic contexts. Yet, these same examples also reveal persistent barriers: digital infrastructure remains uneven, access to computational resources like GPUs is highly asymmetrical, and educational pipelines are not equally prepared to support the ethical development of AI.

There was broad agreement among experts that shared principles, e.g. fairness, safety, transparency, accountability, and human agency, should be established through collaborative international frameworks. Institutions like UNESCO and ISO offer potential pathways to formalize these principles into guidance or standards. However, we cautioned that such high-level norms must be flexibly adapted during implementation [49]. Ethical priorities may differ across societies; values such as individual autonomy, collective well-being, or relational trust are interpreted differently depending on cultural context. Local legal systems, civic traditions, and data governance models will necessarily shape how responsible AI is put into practice [59].

Despite the aspirational nature of global alignment, several persistent challenges must be acknowledged. First, enforcing responsibility across multinational technology companies remains a major jurisdictional hurdle. These entities often operate across borders, complicating efforts by any single country to regulate harmful content, discriminatory algorithms, or opaque business practices. Second, inequities in access to the computational and data infrastructure that underpins AI development threaten to widen the global digital divide. Without targeted support, many low- and middle-income countries may remain consumers of AI, rather than contributors to its shaping. Third, achieving cultural alignment in AI ethics remains elusive. While universal principles are attractive, they risk erasing context-specific norms and practices. Education and engagement must therefore be grounded in local values and lived experiences.

As one speaker poignantly summarized, "the Pandora's box is open." The global community must now move beyond reactive containment strategies and toward proactive, collaborative governance. Doing so requires a delicate balance between establishing universal norms and respecting local nuances. It calls for researchers, educators, policymakers, and civil society actors to work together across borders and disciplines. Above all, the development and deployment of AI must be guided by intentionality, humility, in principle "love and care", a recognition that the tools we create will shape societies for generations to come.

4.3 Common Themes and Collaboration Opportunities

As artificial intelligence continues to grow in global reach and influence, it is increasingly clear that addressing its societal impacts demands transnational coordination. While the challenges of responsible AI development are frequently discussed in national contexts, the implications of AI systems often transcend borders. Accordingly, regional and global collaboration is not only beneficial but essential. Through international engagement, we can foster inclusive innovation, bring cultural relevance, and harmonize ethical standards to guide the future of responsible AI.

4.3.1 The Imperative for International Collaboration

Many state-of-the-art AI models are trained on data collected in the United States or other Western nations. However, their applications are global, affecting populations with vastly different cultural, linguistic, and infrastructural contexts. This mismatch risks reinforcing biases, misrepresenting non-Western perspectives, and exacerbating digital inequalities. International collaboration offers a path to correct this imbalance. By co-developing evaluation frameworks and research agendas, countries can enforce that AI systems meet diverse needs and respect varied social norms.

Collaboration also supports the creation of interoperable governance structures. With the proliferation of national AI strategies and regulatory proposals, there is a growing risk of fragmentation [48]. Coordinated policy dialogue and joint technical standardization can mitigate this risk and confirm that safety, fairness, and transparency are consistently prioritized [29].

4.3.2 Building Shared Infrastructure and Standards

Effective AI evaluation relies on diverse, high-quality data and robust computational infrastructure. Yet access to these resources remains highly uneven across regions. To democratize responsible AI development, international partnerships must prioritize shared infrastructure. This includes pooled compute resources, open-source models, and accessible data repositories, particularly in domains such as healthcare, agriculture, and education.

Standardization efforts are equally critical. Agreements on data collection protocols, annotation practices, sensor compatibility, and privacy-preserving techniques (such as federated learning) will support more consistent and equitable model performance across different use cases and geographies [33].

4.3.3 Addressing Cultural Context, Fairness, and Bias

Cultural sensitivity is essential for meaningful AI deployment. Language, behavior, and social expectations vary widely across regions, and models trained in one context may fail or produce harmful outcomes in another. Rather than merely translating AI systems, it is vital to ground them in local realities. This includes developing culturally aware evaluation metrics and ensuring that fairness is assessed not only in technical terms, but also relative to societal norms.

Collaborating with local stakeholders, particularly community-based organizations and domain experts, can make AI systems reflect local values and serve local needs [17]. Grounding generative

AI models in region-specific data and sociocultural understanding can also mitigate the risk of miscommunication and bias.

4.3.4 Cross-Cultural Interaction and AI Agents

As conversational and generative AI agents become more embedded in daily life, their capacity to navigate cross-cultural settings becomes increasingly important. These agents must understand local customs, interpret gestures and idioms appropriately, and adjust behavior depending on regional norms. Designing culturally fluent agents is not simply a matter of translation, but of embedding deep contextual understanding into their training and evaluation.

International efforts to define what cultural fluency means in practice, and to evaluate whether AI systems achieve it, can support broader goals of user trust and inclusivity. This includes the development of benchmarks for cultural alignment and methods for measuring how well systems adapt to diverse communication styles and expectations.

4.3.5 Regulatory and Ethical Convergence

As countries pursue their own frameworks for AI oversight, inconsistencies in regulation and ethical standards are emerging. Divergent requirements for transparency, risk classification, and liability can hinder cross-border collaboration and increase compliance burdens for developers. At the same time, shared concerns around consent, data rights, and algorithmic accountability present an opportunity for convergence.

While allowing for cultural specificity, efforts to harmonize standards can improve interoperability and create clearer pathways for responsible innovation. This includes developing shared definitions of responsible AI, recognizing common audit requirements, and facilitating the mutual recognition of certifications and evaluations. Intergovernmental bodies and international standard-setting organizations can play a vital role in anchoring this alignment.

4.3.6 Coordinated Responses to Misinformation and AI Misuse

The rise of AI-generated misinformation, such as deepfakes and synthetic content, scam calls with fake identities, presents a shared challenge. The rapid spread of such false or misleading content can cause confusion, fear, and even division in society. For countries like Singapore or USA, with a multi-cultural, multi-religious or multilingual population, harmful AI-generated content that plays on cultural misunderstandings, language differences, or social sensitivities can quickly disrupt social stability. No single country can address these threats in isolation. International cooperation is essential to develop shared tools for detection, attribution, and provenance verification. This may include watermarking techniques, cryptographic content signatures, creation of multilingual and multi-cultural misinformation dataset, active detection AI of factually incorrect content, and collaborative moderation frameworks.

Equally important is the need for coordinated response strategies. Whether responding to misinformation during elections, public health crises, or climate-related disasters, countries must

work together to warrant that their AI systems and response protocols reinforce instead of undermining trust and public safety [36].

4.3.7 Joint Efforts in High-Impact Domains

Several application domains are particularly well suited for cross-regional collaboration. In healthcare, for instance, initiatives to build regional data repositories, akin to the UK Biobank, could improve population health modeling and support AI-enabled diagnostics across Southeast Asia and the Pacific. In education, shared AI infrastructure can help deliver equitable learning opportunities and support AI literacy in multiple languages. In agriculture and environmental monitoring, cross-border data sharing can support regional food security and climate resilience.

Even data from widely used social media platforms can be leveraged cautiously as a proxy for understanding local norms and behaviors. When governed ethically, such data can support the development of culturally sensitive models and inform the design of AI systems attuned to local priorities.

4.3.8 Talent Development and Institutional Partnerships

Sustained international progress in responsible AI requires investment in people as much as in technology. Building equitable talent pipelines involves supporting institutions in underserved regions, facilitating student and faculty exchanges, and creating joint training programs. These partnerships can help build research capacity, foster mutual understanding, and bring future AI leaders who are equipped to operate in globally interconnected environments.

By emphasizing co-supervision, shared research agendas, and multilingual educational resources, such efforts can promote both excellence and inclusivity in the global AI ecosystem.

4.3.9 Cross-Regional Themes and Shared Priorities

Across the countries represented in this workshop—including India, Vietnam, Thailand, and Australia—several common themes emerge that reinforce the need for regional cooperation.

First, localization is essential. AI systems must be designed with a deep understanding of local contexts, languages, and infrastructure. Solutions that work in one country may not translate effectively elsewhere without adaptation. Second, inclusive capacity-building is critical. Broad-based education and digital inclusion efforts, especially for rural and marginalized populations, are necessary to democratize AI benefits. Third, collaborative ecosystems that span government, academia, civil society, and industry can support innovation while embedding mechanisms for transparency and accountability.

In addition, ethical alignment must be grounded in local values while remaining interoperable with global frameworks. Culturally sensitive governance structures are necessary to balance the rights of individuals and communities with the promise of technological advancement. Finally, scalable infrastructure, including shared compute power, public datasets, and open-source tools, can

support a more equitable AI landscape by lowering barriers for small actors and underrepresented regions.

By weaving these themes into the fabric of responsible AI research and evaluation, international collaborations can help build AI systems that advance not only technological goals, but also societal well-being across interconnected and diverse global communities.

4.4 U.S./NSF's Role and Mutual Benefits

The U.S. National Science Foundation (NSF) is uniquely positioned to play a catalytic role in fostering international collaborations on responsible AI. With its global credibility, long-standing commitment to fundamental research, and robust funding mechanisms [26], NSF can support high-impact partnerships that advance both U.S. strategic interests and the responsible deployment of AI technologies worldwide.

NSF affiliation adds credibility and trust to collaborative efforts, particularly in regions where international projects may face skepticism or require endorsement from reputable institutions. This institutional standing can ease access to localized datasets, facilitate regulatory permissions, and encourage participation from governmental and academic stakeholders in partner countries. Moreover, NSF's emphasis on ethical and responsible research aligns well with many countries' growing interest in AI that serves the public good [42], thereby laying a solid foundation for mutually beneficial engagement.

International collaboration in AI evaluation offers clear reciprocal advantages. For the U.S., such partnerships provide access to diverse environments and user populations, which are invaluable for testing the robustness, fairness, and generalizability of AI systems. Whether evaluating AI tools in low-resource healthcare settings, language-rich environments, or dense urban infrastructures, these varied testbeds allow researchers to uncover failure modes and biases that may not emerge in U.S.-only contexts. This improves not only the reliability of AI but also its readiness for global deployment.

Conversely, partner countries benefit from collaboration through access to cutting-edge research tools, technical expertise, and capacity-building resources. NSF-funded initiatives can support technology transfer, co-development of evaluation tools, and the training of researchers and practitioners. These contributions help strengthen local innovation ecosystems and promote self-sufficiency in the development and evaluation of AI systems.

A practical and scalable collaboration model would involve NSF funding the U.S. side of research activities, while partner countries support their domestic contributions. This distributed investment model promotes joint ownership and sustainability, and it builds on successful precedents in other domains of international science cooperation. Importantly, it allows for flexible alignment with national research priorities and funding structures.

To operationalize such partnerships, several mechanisms can be considered. NSF could expand support for U.S. graduate students, postdoctoral researchers, and faculty to engage in collaborative fieldwork abroad, working alongside local partners on contextually grounded projects. Memoranda of Understanding (MOUs) between NSF and counterpart agencies in partner countries can help formalize collaboration in specific thematic areas, like AI for crowd management, localized autonomous driving systems [146], or health data collection for underserved populations. These joint efforts would not only advance scientific discovery but also reinforce shared values around equity, safety, and human-centered design in AI [57].

Through these engagements, NSF can serve as a global convenor of responsible AI research, ensuring that the benefits of technological progress are equitably distributed, culturally relevant, and aligned with democratic values. In doing so, the U.S. can strengthen its role as a trusted partner in shaping the global future of AI.

5. Conclusions

The NSF US–Southeast Asia Responsible AI workshop highlighted the risks, challenges and opportunities in advancing AI that is safe, equitable, and socially beneficial [28]. Key themes emerging from the workshop included the critical role of global interdisciplinary collaboration, the need to bridge the gap between research and practice, and the value of risk-based governance and evaluation frameworks. Responsible AI is not only a technical endeavor but also a social one: it requires engaging diverse stakeholders, appreciating cultural contexts, and embedding accountability throughout the AI lifecycle [55]. By fostering collaboration between technologists, human-centered computing experts, policymakers, social scientists, and community advocates, we can advance both the development of responsible AI and the evaluation of responsible AI, creating frameworks and tools that are actionable and adaptable across diverse regional and global contexts [32].

Looking ahead, sustaining responsible AI will depend on flexible and forward-looking strategies, including certification and compliance mechanisms, participatory evaluation practices, and globally coordinated standards [31]. These efforts should aim to promote transparency, fairness, and inclusivity while supporting innovation. The workshop underscores that responsible AI is a collective pursuit: success requires commitment from researchers, industry, regulators, and civil society to co-create AI systems that serve public interest, mitigate harm, and reflect ethical and societal values [61]. By translating research insights into practical tools, policies, and guidelines, this collaborative effort can strengthen both the development and evaluation of responsible AI, shaping AI in ways that are accountable, culturally aware, and aligned with long-term public good.

Acknowledgement

This work is supported in part by National Science Foundation IIS 2437592 and AI Singapore.

References:

[1] https://indiaai.gov.in/

[2] https://www.swaayattrobots.com/

[3] https://esanjeevani.mohfw.gov.in

[4] https://agriwelfare.gov.in/en/DigiAgriDiv

[5] https://indiaai.gov.in/article/ai-in-agriculture-in-2025-transforming-indian-farms-for-a-sustainable-future

[6] S. Avin, H. Belfield, M. Brundage, G. Krueger, J. Wang, A. Weller, M. Anderljung, I. Krawczuk, D. Krueger, J. Lebensold, *et al.*, "Filling gaps in trustworthy development of AI," *Science*, vol. 374, no. 6573, pp. 1327–1329, 2021.

[7] Y. Bengio, T. Maharaj, L. Ong, S. Russell, D. Song, M. Tegmark, L. Xue, Y.-Q. Zhang, S. Casper, W. S. Lee, *et al.*, "The Singapore consensus on global AI safety research priorities," *arXiv preprint* arXiv:2506.20702, 2025.

[8] R. Bommasani, S. R. Singer, R. E. Appel, S. Cen, A. F. Cooper, L. A. Gailmard, I. Klaus, M. M. Lee, I. D. Raji, A. Reuel, *et al.*, "The California report on frontier AI policy," *arXiv preprint* arXiv:2506.17303, 2025.

[9] N. K. Corrêa, C. Galvão, J. W. Santos, C. Del Pino, E. P. Pinto, C. Barbosa, D. Massmann, R. Mambrini, L. Galvão, E. Terem, *et al.*, "Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance," *Patterns*, vol. 4, no. 10, 2023.

[10] M. A. Akbar, A. A. Khan, S. Mahmood, S. Rafi, and S. Demi, "Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges," Software: Practice and Experience, early access, n.d., doi: 10.1002/spe.3216.

[11] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," arXiv preprint arXiv:1606.06565, 2016. [Online]. Available: http://arxiv.org/abs/1606.06565

[12] Z. Arnold, D. S. Schiff, K. J. Schiff, B. Love, J. Melot, N. Singh, L. Jenkins, A. Lin, K. Pilz, O. Enweareazu, and T. Girard, "Introducing the AI Governance and Regulatory Archive (AGORA): An analytic infrastructure for navigating the emerging AI governance landscape," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2024, pp. 39–48.

[13] Z. Ashktorab, G. Bansal, Z. Buçinca, K. Holstein, J. Hullman, A. M. Smith-Renner, T. Wu, and W. Zhang, "Trust and reliance in evolving human-AI workflows (TREW)," in Extended Abstracts of the 2024 CHI Conf. Human Factors in Computing Systems, 2024, pp. 1–6, doi: 10.1145/3613905.3636319.

[14] S. Avin, H. Belfield, M. Brundage, G. Krueger, J. Wang, A. Weller, M. Anderljung, I. Krawczuk, D. Krueger, J. Lebensold, T. Maharaj, and N. Zilberman, "Filling gaps in trustworthy development of AI," Science, vol. 374, no. 6573, pp. 1327–1329, 2021, doi: 10.1126/science.abi7176.

[15] A. M. Barrett, D. Hendrycks, J. Newman, and B. Nonnecke, "Actionable guidance for high-consequence AI risk management," unpublished, 2022.

[16] G. Berman, N. Cooper, W. H. Deng, and B. Hutchinson, "Troubling taxonomies in GenAI evaluation," arXiv preprint arXiv:2410.22985, 2024. [Online]. Available: http://arxiv.org/abs/2410.22985

[17] J. B. Biddle, J. P. Nelson, and O. E. Olugbade, "How can we know if you are serious? Ethics washing, symbolic ethics offices, and the responsible design of AI systems," Canadian Journal of Philosophy, pp. 1–17, 2025, doi: 10.1017/can.2025.9.

[18] B. Blili-Hamelin et al., "Position: Stop treating 'AGI' as the north-star goal of AI research," in Proc. 42nd Int. Conf. Mach. Learn. (ICML), Position Paper Track, 2025. [Online]. Available: https://openreview.net/forum?id=1RlrtH6ydW

[19] B. Blili-Hamelin et al., "Stop treating 'AGI' as the north-star goal of AI research," arXiv preprint arXiv:2502.03689, 2025, doi: 10.48550/arXiv.2502.03689.

[20] R. Bommasani et al., "Foundation model transparency reports," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), vol. 7, 2024, pp. 181–195.

[21] R. Bommasani et al., "The California report on frontier AI policy," arXiv preprint arXiv:2506.17303, 2025, doi: 10.48550/arXiv.2506.17303.

[22] M. Chiodo and D. Müller, "The problem of algorithmic collisions: Mitigating unforeseen risks in a connected world," arXiv preprint arXiv:2505.20181, 2025, doi: 10.48550/arXiv.2505.20181.

[23] M. Coeckelbergh, "LLMs, truth, and democracy: An overview of risks," Science and Engineering Ethics, vol. 31, no. 1, Art. no. 1, 2025, doi: 10.1007/s11948-025-00529-0.

[24] N. K. Corrêa et al., "Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance," Patterns, vol. 4, no. 10, Art. no. 100857, 2023, doi: 10.1016/j.patter.2023.100857.

[25] F. S. de Sio, G. Mecacci, S. Calvert, D. Heikoop, M. Hagenzieker, and B. van Arem, "Realising meaningful human control over automated driving systems: A multidisciplinary approach," Minds and Machines, vol. 33, no. 4, pp. 587–611, 2023, doi: 10.1007/s11023-022-09608-8.

[26] N. Dreksler et al., "What the public thinks about AI and the implications for governance," Brookings TechTank, https://www.brookings.edu/?p=1812651&post_type=article&preview=1&_ppp=2d45508bb2 2025

[27] M. Feffer, A. Sinha, W. H. Deng, Z. C. Lipton, and H. Heidari, "Red-teaming for generative AI: Silver bullet or security theater?" in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), vol. 7, 2024, pp. 421–437.

[28] F. V. Giarmoleo, I. Ferrero, M. Rocchi, and M. Pellegrini, "What ethics can say on artificial intelligence: Insights from a systematic literature review," Business and Society Review, early access, n.d., doi: 10.1111/basr.12336.

[29] J. A. Goldstein and G. Sastry, "The PPOu framework: A structured approach for assessing the likelihood of malicious use of advanced AI systems," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), vol. 7, 2024, pp. 503–518.

[30] L. Hammond et al., "Multi-agent risks from advanced AI," arXiv preprint arXiv:2502.14143, 2025, doi: 10.48550/arXiv.2502.14143.

[31] E. Hermann, "Deploying artificial intelligence in services to aid vulnerable consumers," Journal of the Academy of Marketing Science, early access, 2023, doi: 10.1007/s11747-023-00986-8.

[32] J. Kay, A. Kasirzadeh, and S. Mohamed, "Epistemic injustice in generative AI," arXiv preprint arXiv:2408.11441, 2024. [Online]. Available: http://arxiv.org/abs/2408.11441

[33] N. Kolt, M. Anderljung, J. Barnhart, A. Brass, K. Esvelt, G. K. Hadfield, L. Heim, M. Rodriguez, J. B. Sandbrink, and T. Woodside, "Responsible reporting for frontier AI development," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), vol. 7, 2024, pp. 768–783.

[34] R. S. S. Kumar, D. O. Brien, K. Albert, S. Viljöen, and J. Snover, "Failure modes in machine learning systems," arXiv preprint arXiv:1911.11034, 2019. [Online]. Available: http://arxiv.org/abs/1911.11034

[35] M. Li, W. Bickersteth, N. Tang, J. Hong, L. Cranor, H. Shen, and H. Heidari, "A closer look at the existing risks of generative AI: Mapping the who, what, and how of real-world incidents," arXiv preprint arXiv:2505.22073, 2025, doi: 10.48550/arXiv.2505.22073.

[36] H. A. Love et al., "The future of work in the age of automation: Proceedings of a workshop on Norbert Wiener's 21st century legacy," IEEE Trans. Technol. Soc., pp. 1–23, 2024, doi: 10.1109/TTS.2024.3476041.

[37] B. Lund, Z. Orhan, N. R. Mannuru, R. V. K. Bevara, B. Porter, M. K. Vinaih, and P. Bhaskara, "Standards, frameworks, and legislation for artificial intelligence (AI) transparency," AI and Ethics, pp. 1–17, 2025, doi: 10.1007/s43681-025-00661-4.

[38] J. Maclure and A. Morin-Martel, "AI ethics' institutional turn," Digital Society, vol. 4, no. 1, Art. no. 1, 2025, doi: 10.1007/s44206-025-00174-x.

[39] L. Methnani, V. Dignum, and A. Theodorou, "PRISM: A pragmatic framework for evaluating counterfactual explanations in XAI," unpublished.

[40] M. Mitchell, A. Ghosh, A. S. Luccioni, and G. Pistilli, "Fully autonomous AI agents should not be developed," arXiv preprint arXiv:2502.02649, 2025, doi: 10.48550/arXiv.2502.02649.

[41] E. Moss, "Trust is not enough: Accuracy, error, randomness, and accountability in an algorithmic society," Commun. ACM, 2023. [Online]. Available: https://cacm.acm.org/opinion/trust-is-not-enough-accuracy-error-randomness-and-accountability-in-an-algorithmic-society/

[42] L. Nazareno and D. S. Schiff, "The impact of automation and artificial intelligence on worker well-being," Technology in Society, vol. 67, p. 101679, 2021, doi: 10.1016/j.techsoc.2021.101679.

[43] E. Nieuwenhuizen, "Algorithm registers: A box-ticking exercise or meaningful tool for transparency?" Information Polity, 2025, doi: 10.1177/15701255241297107.

[44] M. O'Shaughnessy, "Five policy uses of algorithmic explainability," arXiv preprint arXiv:2302.03080, 2023, doi: 10.48550/arXiv.2302.03080.

[45] B. Rakova, R. Shelby, and M. Ma, "Terms-we-serve-with: Five dimensions for anticipating and repairing algorithmic harm," Big Data & Society, vol. 10, no. 2, p. 20539517231211553, 2023, doi: 10.1177/20539517231211553.

[46] M. Rauh et al., "Gaps in the safety evaluation of generative AI," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), vol. 7, 2024, pp. 1200–1217.

[47] A. Resseguier and F. Ufert, "AI research ethics is in its infancy: The EU's AI Act can make it a grown-up," Research Ethics, 2023, doi: 10.1177/17470161231220946.

[48] D. Schiff, "Future trends in global AI governance," SSRN Scholarly Paper 5343797, 2025, doi: 10.2139/ssrn.5343797.

[49] D. Schiff, "Strategies for harmonizing fragmented AI ethics frameworks, standards, and regulations," SSRN Scholarly Paper 5343799, 2025, doi: 10.2139/ssrn.5343799.

[50] D. Schiff, B. Rakova, A. Ayesh, A. Fanti, and M. Lennon, "Explaining the principles to practices gap in AI," IEEE Technol. Soc. Mag., vol. 40, no. 2, pp. 81–94, 2021, doi: 10.1109/MTS.2021.3056286.

[51] D. S. Schiff, S. Kelley, and J. Camacho Ibáñez, "The emergence of artificial intelligence ethics auditing," Big Data & Society, vol. 11, no. 4, p. 20539517241299732, 2025, doi: 10.1177/20539517241299732.

[52] D. S. Schiff, K. J. Schiff, and P. Pierson, "Assessing public value failure in government adoption of artificial intelligence," Public Administration, vol. 100, no. 3, pp. 653–673, 2022, doi: 10.1111/padm.12742.

[53] Singapore Conference on AI, "The Singapore consensus on global AI safety research priorities," Government of Singapore, 2025. [Online]. Available: https://www.scai.gov.sg/2025/scai2025-report

[54] Meta AI, "System cards, a new resource for understanding how AI systems work," Meta AI Blog, 2022. [Online]. Available: https://ai.facebook.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/

[55] S. Vallor and T. Vierkant, "Find the gap: AI, responsible agency and vulnerability," Minds and Machines, vol. 34, no. 3, p. 20, 2024, doi: 10.1007/s11023-024-09674-0.

[56] H. Wallach et al., "Evaluating generative AI systems is a social science measurement challenge," arXiv preprint arXiv:2411.10939, 2024. [Online]. Available: http://arxiv.org/abs/2411.10939

[57] L. Wiese, S. S. Rathinam, M. Oschinski, B. DeWitt, and D. S. Schiff, "AI ethics and governance in the job market: Trends, skills, and sectoral demand," IEEE Trans. Technol. Soc., early access, 2025, doi: 10.1109/TTS.2025.3567143.

[58] Z. Xiao, W. H. Deng, M. S. Lam, M. Eslami, J. Kim, M. Lee, and Q. V. Liao, "Human-centered evaluation and auditing of language models," in Extended Abstracts of the 2024 CHI Conf. Human Factors in Computing Systems, 2024, pp. 1–6, doi: 10.1145/3613905.3636302.

[59] A. Younas and Y. Zeng, "Epistemic modesty in cross-cultural AI ethics: Against forced relevance and coerced inclusion," PhilArchive, 2025. [Online]. Available: https://philarchive.org/archive/YOUEMI-2

[60] E. Papagiannidis, P. Mikalef, and K. Conboy, "Responsible artificial intelligence governance: A review and research framework," *J. Strategic Inf. Syst.*, vol. 34, no. 2, p. 101885, 2025, doi: 10.1016/j.jsis.2024.101885.

[61] M. Sadek, E. Kallina, T. Bohné, C. Mougenot, R. A. Calvo, and S. Cave, "Challenges of responsible AI in practice: Scoping review and recommended actions," *AI & Society*, vol. 40, no. 1, pp. 199–215, 2025, doi: 10.1007/s00146-024-01877-4.

[62] P. Radanliev, O. Santos, A. Brandon-Jones, and A. Joinson, "Ethics and responsible AI deployment," *Front. Artif. Intell.*, vol. 7, p. 1377011, 2024, doi: 10.3389/frai.2024.1377011.

[63] I. H. Sarker, "LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling," *Discover Artif. Intell.*, vol. 4, no. 1, p. 40, 2024, doi: 10.1007/s44163-024-00097-4.

[64] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet, "Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–35, 2024, doi: 10.1145/3649446.

[65] Infocomm Media Development Authority (IMDA) and AI Verify Foundation, *Model AI Governance Framework for Generative AI*. Singapore, Jun. 6, 2024. [Online]. Available: https://aiverifyfoundation.sg/wp-content/uploads/2024/06/Model-AI-Governance-Framework-for-Generative-AI-6-June-2024.pdf

[66] AI Verify Foundation, "AI Verify Foundation—Building Trustworthy AI." Accessed Aug. 31, 2025. [Online]. Available: https://aiverifyfoundation.sg/

[67] Infocomm Media Development Authority (IMDA), *Starter Kit for Safety Testing of LLM-Based Applications (Draft for Public Consultation)*. Singapore, May 28, 2025. [Online]. Available: https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/large-language-model-starter-kit.pdf

[68] Infocomm Media Development Authority (IMDA), *AI Playbook for Small States*. Singapore, Sep. 22, 2024. [Online]. Available: https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2024/09/ai-playbook-for-small-states/imda-ai-playbook-for-small-states.pdf

[69] Y. Kim *et al.*, "Medical hallucinations in foundation models and their impact on healthcare," *arXiv preprint arXiv:2503.05777*, 2025. [Online]. Available: https://arxiv.org/abs/2503.05777

[70] R. Hatem, B. Simmons, and J. E. Thornton, "A call to address AI 'hallucinations' and how healthcare professionals can mitigate their risks," *Cureus*, vol. 15, no. 9, 2023, doi: 10.7759/cureus.44722.

[71] B. Munir, "Hallucinations in legal practice: A comparative case law analysis," *Int. J. Law, Ethics, Technol.*, 2025.

[72] M. Dahl *et al.*, "Large legal fictions: Profiling legal hallucinations in large language models," *J. Legal Anal.*, vol. 16, no. 1, pp. 64–93, 2024, doi: 10.1093/jla/lau023.

[73] H.-T. Ho, D.-T. Ly, and L. V. Nguyen, "Mitigating hallucinations in large language models for educational application," in *Proc. IEEE Int. Conf. Consumer Electronics-Asia (ICCE-Asia)*, 2024, pp. 1–4, doi: 10.1109/ICCE-Asia59899.2024.10788512.

[74] H. Elsayed, “The impact of hallucinated information in large language models on student learning outcomes: A critical examination of misinformation risks in AI-assisted education,” *Northern Rev. Algorithmic Res., Theor. Comput., Complex.*, vol. 9, no. 8, pp. 11–23, 2024.

[75] S. Zhang, et al., “Which agent causes task failures and when? On automated failure attribution of LLM multi-agent systems,” *arXiv preprint arXiv:2505.00212*, 2025. [Online]. Available: https://arxiv.org/abs/2505.00212

[76] S. Lee, et al., “LLM Attributor: Interactive visual attribution for LLM generation,” in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, vol. 39, no. 28, 2025.

[77] J. Bae, et al., “Training data attribution via approximate unrolling,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 66647–66686, 2024.

[78] F. Zhang and N. Nanda, “Towards best practices of activation patching in language models: Metrics and methods,” in *Proc. 12th Int. Conf. Learning Representations (ICLR)*, 2024.

[79] P. W. Koh and P. Liang, “Understanding black-box predictions via influence functions,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2017, pp. 1885–1894.

[80] H. Wang, S. Tan, and H. Wang, “Probabilistic conceptual explainers: Trustworthy conceptual explanations for vision foundation models,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2024, pp. 51502–51522.

[81] V. Arya, R. K. Bellamy, P. Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, and S. Mourad, “AI explainability 360 toolkit,” in *Proc. 3rd ACM India Joint Int. Conf. Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, Jan. 2021, pp. 376–379.

[82] P. Rajpurkar, et al., “AI in health and medicine,” *Nature Med.*, vol. 28, no. 1, pp. 31–38, 2022, doi: 10.1038/s41591-021-01614-0.

[83] M. R. King, “The future of AI in medicine: A perspective from a chatbot,” *Ann. Biomed. Eng.*, vol. 51, no. 2, pp. 291–295, 2023, doi: 10.1007/s10439-022-03074-2.

[84] P. Lee, S. Bubeck, and J. Petro, “Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine,” *New Engl. J. Med.*, vol. 388, no. 13, pp. 1233–1239, 2023, doi: 10.1056/NEJMsr2214184.

[85] C. Milana and A. Ashta, “Artificial intelligence techniques in finance and financial markets: A survey of the literature,” *Strategic Change*, vol. 30, no. 3, pp. 189–209, 2021, doi: 10.1002/jsc.2406.

[86] B. Chen, Z. Wu, and R. Zhao, “From fiction to fact: The growing role of generative AI in business and finance,” *J. Chinese Econ. Bus. Stud.*, vol. 21, no. 4, pp. 471–496, 2023, doi: 10.1080/14765284.2023.2254359.

[87] S. Bi, J. Xiao, and T. Deng, “The role of AI in financial forecasting: ChatGPT’s potential and challenges,” in *Proc. 4th Asia-Pacific Artif. Intell. Big Data Forum*, 2024, pp. 1–6.

[88] M. Imran and N. Almusharraf, “Google Gemini as a next generation AI educational tool: a review of emerging educational technology,” *Smart Learn. Environ.*, vol. 11, no. 1, p. 22, 2024.

[89] A. Harry, “Role of AI in education,” *Interdisciplinary Journal & Humanity (INJURITY)*, vol. 2, no. 3, 2023.

[90] K. Zhang and A. B. Aslan, “AI technologies for education: Recent research & future directions,” *Comput. Educ.: Artif. Intell.*, vol. 2, p. 100025, 2021.

[91] F. Santoni de Sio, T. Almeida, and J. Van Den Hoven, “The future of work: Freedom, justice and capital in the age of artificial intelligence,” *Crit. Rev. Int. Soc. Political Philos.*, vol. 27, no. 5, pp. 659–683, 2024, doi: 10.1080/13698230.2022.2122636.

[92] R. Ejjami, “AI-driven justice: Evaluating the impact of artificial intelligence on legal systems,” *Int. J. Multidiscip. Res.*, vol. 6, no. 3, pp. 1–29, 2024.

[93] A. R. Vargas-Murillo, et al., “Transforming justice: Implications of artificial intelligence in legal systems,” *Acad. J. Interdiscip. Stud.*, vol. 13, no. 2, p. 433, 2024, doi: 10.36941/ajis-2024-0042.

[94] N. Maleki, B. Padmanabhan, and K. Dutta, “AI hallucinations: A misnomer worth clarifying,” in *Proc. IEEE Conf. Artif. Intell. (CAI)*, 2024, pp. 1–8, doi: 10.1109/CAI59869.2024.10621126.

[95] S. Monteith, et al., “Artificial intelligence and increasing misinformation,” *Br. J. Psychiatry*, vol. 224, no. 2, pp. 33–35, 2024, doi: 10.1192/bjp.2023.55.

[96] R. Emsley, “ChatGPT: These are not hallucinations–they’re fabrications and falsifications,” *Schizophrenia*, vol. 9, no. 1, p. 52, 2023, doi: 10.1038/s41537-023-00397-1.

[97] C. Zhai, S. Wibowo, and L. D. Li, “The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review,” *Smart Learn. Environ.*, vol. 11, no. 1, p. 28, 2024, doi: 10.1186/s40561-024-00347-3.

[98] M. Gerlich, “AI tools in society: Impacts on cognitive offloading and the future of critical thinking,” *Societies*, vol. 15, no. 1, p. 6, 2025, doi: 10.3390/soc15010006.

[99] J. Y. Bo, S. Wan, and A. Anderson, “To rely or not to rely? Evaluating interventions for appropriate reliance on large language models,” in *Proc. CHI Conf. Human Factors Comput. Syst. (CHI)*, 2025, pp. 1–15, doi: 10.1145/3613904.3642392.

[100] V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, vol. 2156. Cham, Switzerland: Springer, 2019.

[101] A. Deshpande and H. Sharp, “Responsible AI systems: Who are the stakeholders?,” in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2022, pp. 287–293.

[102] P. Mikalef, I. O. Pappas, J. Krogstie, and M. Giannakos, “Thinking responsibly about responsible AI and the ‘dark side’ of AI,” *Eur. J. Inf. Syst.*, vol. 31, no. 3, pp. 257–268, 2022, doi: 10.1080/0960085X.2022.2064002.

[103] R. Baeza-Yates, “Introduction to responsible AI,” in *Proc. 17th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2024, pp. 1391–1392.

[104] I. H. Sarker, “LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling,” *Discover Artif. Intell.*, vol. 4, no. 1, p. 40, 2024, doi: 10.1007/s44163-024-00097-4.

[105] Y. Hong, Y. Li, X. Zhang, S. Wang, J. Wang, and J. Ni, “Statistical perspectives on reliability of artificial intelligence systems,” *Qual. Eng.*, vol. 35, no. 1, pp. 56–78, 2023, doi: 10.1080/08982112.2022.2105981.

[106] B. Li, X. Huang, H. Chen, X. Li, Y. Zhang, Y. Liu, J. Zhao, T. Liu, and J. Gao, “Trustworthy AI: From principles to practices,” *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–46, 2023, doi: 10.1145/3539600.

[107] H. Liu, J. Chen, C. Xu, Z. Zhang, Y. Chen, and B. Li, “Trustworthy AI: A computational perspective,” *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 1, pp. 1–59, 2022, doi: 10.1145/3512787.

[108] S. G. Ayyamperumal and L. Ge, “Current state of LLM risks and AI guardrails,” *arXiv preprint* arXiv:2406.12934, 2024. [Online]. Available: http://arxiv.org/abs/2406.12934

[109] T. Rebedea, C. Ristea, V. C. Ostafe, M. Rachatasumrit, D. S. Chaplot, A. Choudhary, and J. K. Kummerfeld, “NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails,” *arXiv preprint* arXiv:2310.10501, 2023. [Online]. Available: http://arxiv.org/abs/2310.10501

[110] A. Kumar, P. Bajaj, S. Kamath, J. Song, and D. Papailiopoulos, “Certifying LLM safety against adversarial prompting,” *arXiv preprint* arXiv:2309.02705, 2023. [Online]. Available: http://arxiv.org/abs/2309.02705

[111] Z. Ji, H. Lee, T. Yu, D. Jurafsky, and P. Liang, “Towards mitigating LLM hallucination via self reflection,” in Findings Assoc. Comput. Linguistics: EMNLP, 2023, pp. 1–15.

[112] A. Martino, M. Iannelli, and C. Truong, “Knowledge injection to counter large language model (LLM) hallucination,” in Proc. Eur. Semantic Web Conf. (ESWC), Cham, Switzerland: Springer, 2023, pp. 1–16.

[113] S. M. Tonmoy, A. S. M. A. Hossain, A. S. M. I. Jony, M. M. Rahman, M. A. Rahman, and M. S. Islam, “A comprehensive survey of hallucination mitigation techniques in large language models,” arXiv preprint arXiv:2401.01313, 2024. [Online]. Available: https://arxiv.org/abs/2401.01313

[114] L. Zheng, S. Chi, Z. Zhang, S. Li, Y. Song, C. Du, Z. Shao, and E. P. Xing, “Judging LLM-as-a-judge with MT-bench and Chatbot Arena,” Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, pp. 46595–46623, 2023.

[115] J. Gu, Z. Wang, X. Zhang, Y. Chen, and M. Sun, “A survey on LLM-as-a-judge,” arXiv preprint arXiv:2411.15594, 2024. [Online]. Available: https://arxiv.org/abs/2411.15594

[116] J. Ye, H. Wu, T. Li, X. Zhou, and Y. Zhang, “Justice or prejudice? Quantifying biases in LLM-as-a-judge,” arXiv preprint arXiv:2410.02736, 2024. [Online]. Available: https://arxiv.org/abs/2410.02736

[117] Weavely, “Creating synthetic users for free with ChatGPT,” *Weavely Blog*, Accessed: Oct. 6, 2025. [Online]. Available: https://www.weavely.ai/blog/creating-synthetic-users-for-free-with-chatgpt

[118] Synthetic Users, “Introducing Shuffle v2,” *Synthetic Users – Science Posts*, Accessed: Oct. 6, 2025. [Online]. Available: https://www.syntheticusers.com/science-posts/introducing-shuffle-v2

[119] Y. Wang, H. Shi, L. Han, D. Metaxas, and H. Wang, “Blob: Bayesian low-rank adaptation by backpropagation for large language models,” in *Advances in Neural Information Processing Systems* (NeurIPS), vol. 37, 2024, pp. 67758–67794.

[120] F. Ye, M. Yang, J. Pang, L. Wang, D. Wong, E. Yilmaz, S. Shi, and Z. Tu, "Benchmarking LLMs via uncertainty quantification," in *Advances in Neural Information Processing Systems* (NeurIPS), vol. 37, 2024, pp. 15356–15385.

[121] N. Mehdiyev, M. Majlatow, and P. Fettke, "Quantifying and explaining machine learning uncertainty in predictive process monitoring: An operations research perspective," *Annals of Operations Research*, pp. 1–40, 2024, doi: 10.1007/s10479-024-06536-3.

[122] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in Neural Information Processing Systems* (NeurIPS), vol. 19, 2006.

[123] H. Wang, H. He, and D. Katabi, "Continuously indexed domain adaptation," in *Proc. Int. Conf. Mach. Learn.* (ICML), Nov. 2020, pp. 9898–9907.

[124] S. Hao, B. Hooi, J. Liu, K.-W. Chang, Z. Huang, and Y. Cai, "Exploring visual vulnerabilities via multi-loss adversarial search for jailbreaking vision-language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), 2025.

[125] S. Hao, Y. Wang, B. Hooi, J. Liu, M. Chen, Z. Huang, and Y. Cai, "Making every step effective: Jailbreaking large vision-language models through hierarchical KV equalization," in *Findings of the Conf. Empirical Methods in Natural Language Processing* (Findings of EMNLP), 2025.

[126] S. Hao, Y. Wang, B. Hooi, M.-H. Yang, J. Liu, C. Tang, Z. Huang, and Y. Cai, "Tit-for-tat: Safeguarding large vision-language models against jailbreak attacks via adversarial defense," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (ICCV) Workshops, 2025.

[127] W. You, B. Hooi, Y. Wang, Y. Wang, Z. Ke, M.-H. Yang, Z. Huang, and Y. Cai, "MIRAGE: Multimodal immersive reasoning and guided exploration for red-team jailbreak attacks," *arXiv preprint arXiv:2503.19134*, 2025. [Online]. Available: https://arxiv.org/abs/2503.19134

[128] C. Xu, Y. Wang, B. Hooi, Y. Cai, and S. Li, "How does watermarking affect visual language models in document understanding?" in *Proc. Conf. Language Modeling* (COLM), 2025.

[129] Z. Wang, B. Hooi, Y. Wang, M.-H. Yang, Z. Huang, and Y. Cai, "Text speaks louder than vision: ASCII art reveals textual biases in vision-language models," in *Proc. Conf. Language Modeling* (COLM), 2025.

[130] L. Mei, J. Yao, Y. Ge, Y. Wang, B. Bi, Y. Cai, J. Liu, M. Li, Z.-Z. Li, D. Zhang, et al., "A survey of context engineering for large language models," *arXiv preprint arXiv:2507.13334*, 2025. [Online]. Available: https://arxiv.org/abs/2507.13334

[131] Z. Wang, Y. Wang, and Y. Cai, "Cure or poison? Embedding instructions visually alters hallucination in vision-language models," *arXiv preprint arXiv:2508.01678*, 2025. Available: https://arxiv.org/abs/2508.01678

[132] S. Li, Y. Cai, and Y. Wang, "SemVink: Advancing VLMs' semantic understanding of optical illusions via visual global thinking," in *Proc. Conf. Empirical Methods in Natural Language Processing* (EMNLP), 2025.

[133] H. Ge, Y. Wang, M.-H. Yang, and Y. Cai, "MRFD: Multi-region fusion decoding with self-consistency for mitigating hallucinations in VLMs," in *Findings of the Conf. Empirical Methods in Natural Language Processing* (Findings of EMNLP), 2025.

[134] Infocomm Media Development Authority (IMDA) and AI Verify Foundation, "Model AI Governance Framework for Generative AI." 2025. [Online]. Available: https://aiverifyfoundation.sg/resources/mgf-gen-ai/

[135] AI Verify Foundation, "AI Verify testing framework—Assessing the responsible implementation of AI solutions against eleven internationally recognized AI governance principles." Accessed: Oct. 7, 2025. [Online]. Available: https://aiverifyfoundation.sg/what-is-ai-verify/

[136] AI Verify Foundation, "Project Moonshot: An open source LLM evaluation toolkit." Accessed: Oct. 7, 2025. [Online]. Available: https://aiverifyfoundation.sg/project-moonshot/

[137] Temasek, *Make Trust a Must: How AI Ethics and Governance is Heralding a New Era for Financial Services*. Temasek, 2021. [Online]. Available: https://www.temasek.com.sg/content/dam/temasek-corporate/news-and-views/resources/reports/how-ai-ethics-and-governance-is-heralding-a-new-era-for-financial-services.pdf

[138] Y. Zhang, J. O. Kephart, Z. Cui, and Q. Ji, "PhysPT: Physics-aware pretrained transformer for estimating human dynamics from monocular videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 2305–2317.

[139] H. Guo and Q. Ji, "Physics-augmented autoencoder for 3D skeleton-based gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 19627–19638.

[140] H. Wang and Q. Ji, "Epistemic uncertainty quantification for pre-trained neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 11052–11061.

[141] H. Wang, D. Joshi, S. Wang, and Q. Ji, "Gradient-based uncertainty attribution for explainable Bayesian deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 12044–12053.

[142] S. Liu, Y. Yao, J. Jia, S. Casper, N. Baracaldo, P. Hase, Y. Yao, C. Y. Liu, X. Xu, H. Li, et al., "Rethinking machine unlearning for large language models," Nature Mach. Intell., pp. 1–14, 2025.

[143] A. Suri, "The missing pieces in India's AI puzzle: Talent, data, and R&D," Carnegie Endowment for International Peace, 2025. Available: https://carnegieendowment.org/research/2025/02/the-missing-pieces-in-indias-ai-puzzle-talent-data-and-randd?lang=en. [Accessed: Sep. 8, 2025].

[144] Google Public Policy, "An AI opportunity agenda for India," n.d. [Online]. Available: https://static.googleusercontent.com/media/publicpolicy.google/en//resources/india_ai_opportunity_agenda_en.pdf. [Accessed: Sep. 8, 2025].

[145] Ministry of Electronics & IT, Govt. of India, "IndiaAI mission," Press Release ID 2012375, Mar. 2024. [Online]. Available: https://www.pib.gov.in/PressReleasePage.aspx?PRID=2012375. [Accessed: Sep. 8, 2025].

[146] Minus Zero, "Autonomous driving solutions," Minus Zero, 2025. [Online]. Available: https://minuszero.ai/. [Accessed: Oct. 7, 2025].

[147] A. Chinchure, et al., "TIBET: Identifying and evaluating biases in text-to-image generative models," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Lecture Notes in Computer Science, vol. 15137, A. Leonardis, E.

Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham, Switzerland: Springer, 2025, pp. 425–442, doi: 10.1007/978-3-031-72986-7_25.

[148] S. Wang, H. Zhang, K. Xu, X. Lin, S. Jana, C.-J. Hsieh, and J. Z. Kolter, "Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 29909–29921, 2021.

[149] J. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 1310–1320.

[150] ICLR, "ICLR 2025 Workshop: Reliable and Responsible Foundation Models," 2025. [Online]. Available: https://iclr.cc/virtual/2025/workshop/23973

[151] E. A. Rocamora, G. G. Chrysos, and V. Cevher, "Certified robustness under bounded Levenshtein distance," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[152] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6572

[153] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013. [Online]. Available: https://arxiv.org/abs/1312.6199

[154] J. Wang, "Adversarial examples in physical world," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2021, pp. 4925–4926.

[155] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018. [Online]. Available: https://arxiv.org/abs/1810.00069

[156] Anthropic, "Project V.E.N.D. 1: Fun experiment on the responsible launch of agents," *Anthropic Research*, Sep. 2025. [Online]. Available: https://www.anthropic.com/research/project-vend-1. [Accessed: Oct. 7, 2025].