مبادرات محمد بن راشد آل مكتوم العالمية
Mohammed Bin Rashid
Al Maktoum Global Initiatives

كلية محمد بن راشد للإدارة الحكومية
MOHAMMED BIN RASHID
SCHOOL OF GOVERNMENT

Mohammed Bin Rashid School Of Government
POLICY BRIEF

Policy Brief No. 68 July 2025

A Strategic AI Safety Imperative: Towards a UAE Artificial Intelligence Safety Institute

Cyrus Hodes and Fadi Salem

Summary

The United Arab Emirates (UAE) stands at a transformative inflection point in its artificial intelligence development journey. While the nation has made significant strides in AI innovation, deployment, solidified by strategic partnerships, a critical gap persists in translating these advancements into global AI security leadership. This policy brief builds the case for establishing a UAE Artificial Intelligence Safety Institute (AISI), both as an urgent national priority and a strategic complement to the Emirates' newly cemented role as the Middle East's AI infrastructure hub.

The UAE's strategic investment in AI was further amplified in mid 2025 by landmark agreements with the United States that position the UAE as a global AI infrastructure leader. The two countries agreed to an investment of $150 billion on AI infrastructure, which includes access to cutting-edge semiconductors and binding security commitments to prevent technology diversion, creates both new opportunities and responsibilities. With Abu Dhabi positioned to host hyperscale AI compute infrastructure capable of serving nearly half the world's population within a 2,000-mile radius, the Emirates now faces imperative to pioneer security frameworks that protect not just national interests but the integrity of global AI ecosystems. The proposed AISI would leverage three strategic advantages: 1) The UAE's established cybersecurity leadership, demonstrated through its world-class threat detection and mitigation capabilities; 2) its unique geopolitical position as a trusted partner to competing AI powers; and 3) The unprecedented scale of sovereign investments in AI infrastructure through vehicles like MGX and G42.

By focusing on advanced AI security research, verification mechanisms, and international coordination, the AISI would address critical gaps in the global AI safety landscape while protecting the UAE's $1 trillion investments in next-generation technologies. Such step would align with the security obligations of the US-UAE AI Acceleration Partnership, which mandates robust safeguards for American-sourced technologies while enabling the UAE to emerge as a global hub for developing transnational AI safety standards. As frontier AI capabilities advance exponentially, the window for establishing governance frameworks that keep pace with infrastructure growth is closing rapidly. The AISI represents not just a strategic necessity for securing the UAE's AI ambitions, but an opportunity to redefine global norms for safe, beneficial AI development in an era of intensifying technological competition.

The UAE's Trajectory towards AI Leadership

Over the past decade, the UAE has cemented its position as a global leader in artificial intelligence through visionary policymaking, strategic investments, and cross-sector partnerships. This trajectory reached a historic inflection point in mid 2025 with the US-UAE AI Acceleration Partnership, which secured the country's role as the Middle East's AI infrastructure epicenter while aligning its technological ambitions with global security imperatives.

The nation's AI leadership journey began with institutional firsts: the 2017 creation of the world's first Ministry of Artificial Intelligence and the National AI Strategy 2031, which targeted a 335 billion dirham GDP boost through AI integration across healthcare, education, and transportation. Since then, the commitment to AI advancement has intensified significantly, marked by the creation of several institutional agencies that practically contribute to this vision, including the UAE Council for Artificial Intelligence and Blockchain . The national strategy was complimented with various strategic activities at the local levels across the seven Emirates that make up the country. Among many others, this included the establishment of the Artificial Intelligence and Advanced Technology Council (AIATC) in Abu Dhabi in 2024 . This was accompanied by substantial financial commitments that demonstrate determination to lead in the AI era. For example, in 2024, the AIATC announced the creation of MGX, a technology investment company with Mubadala and Group 42 (G42) serving as foundational partners. MGX's investment strategy focuses on three critical areas: AI infrastructure (including data centers and connectivity), semiconductors (including chip design and manufacturing), and AI core technologies and applications . Industry analysts project that MGX could amass up to $100 billion in assets under management within a few years , positioning it as one of the world's most significant AI-focused investment vehicles.

Sustained investment across the AI ecosystem in the UAE triggered the rise of several powerful and globally prominent homegrown AI companies. One key player is Group 42 (G42), the UAE-based artificial intelligence development holding company founded in 2018, which has rapidly established itself as a major player in the global AI landscape. The company received a transformative boost in 2024 when Microsoft announced a $1.5 billion investment in G42 , demonstrating international confidence in the UAE's AI capabilities.

The scale of the UAE's international AI influence is perhaps best illustrated by its participation in the Stargate project announced in January 2025 . This initiative, described as "the largest AI infrastructure project by far in history," aims to invest $500 billion to build AI infrastructure in the United States. Abu Dhabi-based MGX has teamed up with US technology firms Oracle and OpenAI, along with Japan's Softbank, to form this venture, announced by the US President as a landmark initiative, underscoring the UAE's growing influence in shaping global AI infrastructure development.

Beyond these partnerships with technology leaders, the UAE has also built sophisticated local AI capabilities. For example, the Technology Innovation Institute (TII) has developed the UAE's Falcon large language model series, establishing indigenous capacity in one of the most transformative AI technologies. G42 has invested in massive computing infrastructure, including purchasing supercomputers from Cerebras capable of 4 exaflops of computing power . This comprehensive approach—spanning investment, infrastructure, research, and applications—demonstrates the UAE's determination to position itself at the forefront of the global AI revolution. In 2023, the Mohammed Bin Zayed University for AI (MBZUAI) has launched the Institute of Foundation Models to lead generative AI research and foster regional AI innovation by bringing together top scientists and experts in the UAE.

This investment framework gained unprecedented momentum through the 5GW AI Data Center Campus unveiled during the US President's state visit to the UAE in May 2025. Spanning 10 square miles in Abu Dhabi and developed by G42 with US hyperscalers, the campus represents the largest AI infrastructure project outside the United States . Phase 1's 1GW capacity will host American cloud providers offering latency-optimized services to nearly half the global population within a 2,000-mile radius, effectively transforming the UAE from an AI investor into a global infrastructure guarantor. The project's scale-equivalent to powering 25 million Nvidia B200 GPUs-positions Abu Dhabi as the computational gateway for the Global South while alleviating US data-center grid constraints.

In 2025, G42 published its Frontier AI Safety Framework , introducing a multi-layered approach to AI risk management designed to ensure that advanced AI systems are developed, tested, and deployed responsibly. The framework establishes frontier capability thresholds, independent governance mechanisms, and deployment safeguards to identify and mitigate AI risks before they become critical. This initiative aligns with global best practices and contributes to ongoing AI safety efforts worldwide, reflecting the UAE's growing attention to responsible AI development alongside its ambitious deployment goals.

The UAE's commitment to responsible AI development extends beyond private sector initiatives through a comprehensive government-led ethical framework. The federal authorities have established robust governance structures beginning with the UAE Charter for the Development and Use of Artificial Intelligence (2024), which mandates algorithmic fairness audits, human oversight requirements, and strict compliance with Abu Dhabi's AI ethics principles for all public-sector AI deployments . This charter operationalizes the AI Ethics Principles and Guidelines published by the Minister of State for AI Office; which in turn was influenced by Dubai's AI Ethics Principles, piloted and disseminated across the local Dubai Government agencies since 2018. The Federal level AI Ethics Principles and Guidelines is an 88-page framework detailing eight core mandates: fairness, accountability, transparency, explainability, robustness, human-centered design, environmental sustainability, and privacy preservation , . The guidelines require impact assessments for high-risk systems, including documentation of training data provenance and bias mitigation strategies for any AI influencing healthcare, employment, or judicial decisions.

At the multilateral level, the UAE's International Stance on AI Policy (2024) codified six strategic pillars - advancement, cooperation, community, ethics, sustainability, and security - creating binding requirements for AI developers to implement "ethical by design" architectures aligned with the OECD AI Principles . This policy framework introduced mandatory transparency registers for public-sector AI systems and established the world's first AI Seal certification program, which evaluates systems against 127 technical criteria spanning algorithmic accountability, third-party auditing capacity, and societal impact metrics . Complementing these technical standards, the Dubai AI Policy for Government Entities (2024) mandates real-time monitoring of AI systems' compliance with the UAE's constitutional protections of privacy and equality, requiring fail-safe mechanisms for any AI impacting Emirati citizens' fundamental rights .

These governmental initiatives create a layered governance ecosystem where private-sector led safety protocols (e.g. G42's) interface with national requirements for independent auditing and oversight. For instance, the Frontier AI Safety Framework's capability thresholds must align with the UAE Charter's prohibition on autonomous systems making "critical decisions" in judicial, medical, or military contexts without human validation loops. This integration of corporate and national standards positions the UAE as one of few nations with AI governance spanning constitutional principles, sectoral regulations, and international cooperation frameworks - a model now influencing GCC-wide AI policy harmonization efforts, as well as other emerging economies across MENA, Africa and Central Asia .

The Global AI Safety Institute Landscape

As artificial intelligence capabilities have rapidly advanced, governments worldwide have recognized the need for specialized institutions to evaluate and ensure the safety of frontier AI systems. This recognition has led to the establishment of AI Safety Institutes (AISIs) across major economies, creating a new institutional dimension of international technology governance that the UAE would benefit from engaging with to maintain its leadership position.

The movement toward establishing AISIs gained significant momentum during the AI Safety Summit at Bletchley Park in November 2023. At this watershed event, both the United Kingdom and the United States announced the creation of their respective institutes, setting a precedent that other nations would quickly follow. The summit produced the "Bletchley Declaration on AI Safety" agreed to by 28 countries including the UAE, which committed signatories to a global approach to addressing AI risks and to "deepening our understanding of the emerging risks of frontier AI". This declaration formed the foundation for an ongoing international dialogue on AI safety governance that continues to evolve.

The momentum continued at the AI Seoul Summit in May 2024, where international leaders agreed to form a network of AI Safety Institutes comprising institutes from the UK, US, Japan, France, Germany, Italy, Singapore, South Korea, Australia, Canada, and the European Union . The summit also produced a safety pledge signed by 16 AI companies, including the UAE's Technology Innovation Institute and G42 . This pledge committed signatories to setting out which risks would be "deemed intolerable" and to publish safety guidelines explaining how they would measure risks.

The formal inauguration of the International Network of AI Safety Institutes in November 2024, marked a critical step toward global coordination in AI safety governance. During this launch, participating nations committed over $11 million to advance AI safety research, with the stated goal that the AI safety institutes and offices are "technical organisations that aim to advance AI safety, help governments and society understand the risks posed by advanced AI systems, and suggest solutions to address those risks in order to minimize harm".

These institutes are defined by three key characteristics related to research, standards and cooperation: 1) they are safety-driven, focusing on assessing and mitigating risks associated with advanced AI systems; 2) they are government-backed, typically operating in coordination with governmental or public bodies; and 3) they are independent entities with a high degree of technical expertise. Their primary functions include evaluating AI systems, ensuring that development complies with safety standards, and mitigating potential risks before they materialize.

As this global infrastructure for AI safety evaluation and governance takes shape, the UAE has maintained its engagement through participation in international declarations and through private sector initiatives. However, as of March 2025, while countries across the Global North and increasingly the Global South establish their own dedicated AI Safety Institutes, the UAE still lacks a comparable institution. This represents a notable gap in the UAE's otherwise comprehensive approach to AI leadership, particularly as institutes from both traditional technology powers and emerging economies join the International Network of AI Safety Institutes (Table 1).

Country/Region	AI Safety Institute Name	Creation Date
United Kingdom	UK AI Safety Institute	November 2023
	UK AI Security Institute	February 2025
United States	US AI Safety Institute	November 2023
	Center for AI Standards and Innovation (CAISI)	June 2025
Japan	Japan AI Safety Institute	February 2024
Singapore	Singapore AI Safety Institute	Early 2024
European Union	EU AI Office	Early 2024
South Korea	South Korea AI Safety Institute	2024
Canada	Canada AI Safety Institute	2024
France	Institut National pour l'Évaluation et la Sécurité de l'Intelligence Artificielle	February 2025
Australia	Australia AI Safety Institute	Planned, no date
Kenya	Involvement in International Network	No specific date
India	India AI Safety Institute	January 30, 2025
China	China AI Safety and Development Association (CnAISDA)	February 2025
Germany/Italy	Part of EU initiatives	Planned, no date
Australia/Canada	Part of International Network	Planned, no date

The UAE's absence from the global AI Safety Institute Network represents an area for action, especially given the country's existing strategic ties with many member countries. Nations such as the United States, United Kingdom, China, and India already collaborate with the UAE on several AI and technology initiatives, while others such as Kenya, for example, partners with UAE-based G42—backed by Microsoft—to expand cloud-computing infrastructure. In addition to its local imperative, establishing a UAE AI Safety Institute would strengthen global coordination on AI safety through already-established bilateral and multilateral partnerships.

The UAE's Unique Geopolitical Position across the AI Landscape

The UAE occupies a distinctive and strategically valuable position in the global AI landscape due to its ability to maintain productive relationships with all key AI players, including the United States and China. This balanced approach has allowed the UAE to access technologies, investments, and partnerships from both spheres of influence, creating unique opportunities (and challenges) that directly influence its approach to AI governance.

The May 2025 AI partnership with the United States significantly enhanced the UAE's strategic position in the global AI landscape. By securing formal access to advanced US AI technologies under a comprehensive security agreement, the UAE has demonstrated its ability to navigate complex geopolitical waters while establishing itself as a trusted partner for the world's leading AI power.

At the same time, the UAE maintains broader economic ties with China that remain significant. China is the UAE's top trading partner, with UAE estimates valuing non-oil trade at $82 billion in 2023, a 34 percent increase from 2021. Beyond trade, the UAE continues to work with Chinese firms in areas such as 5G networks and other technologies.

This strengthened relationship with the US, coupled with the UAE's existing ties to China as its top trading partner, uniquely positions the country as a credible neutral ground for AI security cooperation and standards development between competing technological spheres. An AI Safety Institute based in the UAE could leverage this position to facilitate dialogue on safety protocols that transcend geopolitical divisions while reinforcing the UAE's reputation as a responsible steward of advanced technologies.

The UAE highlights a case study for countries seeking to balance relations with both the United States and China in the digital age . The UAE's ability to navigate these complex geopolitical waters has significant implications not only for its own technological development but for the broader landscape of global AI governance.

This UAE's balanced position was articulated by UAE Assistant Minister for Advanced Science and Technology, who stated that navigating between two great powers is what other third-party countries are "trying to learn from the UAE model". He observed that "countries who are able to navigate that the best, are the ones who get the most out of the opportunities that are there globally, whether economic or other areas that involve science and tech" . This approach to international relations creates unique opportunities for the UAE in the realm of AI governance.

The strategic importance of this position cannot be overstated. As global tensions over technology leadership intensify, countries like the UAE that maintain productive relationships with both powers can potentially serve as bridges, facilitating dialogue and cooperation where direct engagement might be politically difficult. In the context of AI safety specifically, an institute based in the UAE could potentially evaluate systems from both Western and Eastern developers, helping to establish common safety standards that transcend geopolitical divisions.

Furthermore, participation in the AI Safety Institute Network would enhance the UAE's role as an active and constructive contributor to the global AI governance ecosystem. By aligning with international efforts to standardize safety and compliance requirements, a UAE-based AISI would help ensure that domestic AI companies can more easily navigate global regulatory landscapes, reducing the cost and complexity of compliance. At the same time, it would give the UAE a seat at the table in shaping how these standards evolve, reinforcing its reputation as a trusted and pragmatic partner in advancing responsible AI development across borders.

The UAE Cybersecurity Landscape: Foundation for AI Safety

The UAE has established robust capabilities in cybersecurity that provide a strong foundation for expanding into AI safety. The UAE's Cyber Security Council has demonstrated advanced capabilities in detecting and neutralizing threats, including those increasingly powered by artificial intelligence technologies. This existing infrastructure and expertise represent valuable assets that could be leveraged in establishing an AI Safety Institute focused initially on AI-enabled cyber threats.

The UAE Cyber Security Council has demonstrated remarkable effectiveness in protecting the nation's digital infrastructure, successfully countering malicious ransomware attacks targeting strategic sectors including government and private entities, with emergency cyber-response systems proactively intercepting and neutralizing approximately 200,000 cyber-attacks daily in early 2025 . Building on these defensive capabilities, the UAE announced plans at IDEX 2025 to establish a state-of-the-art Cybersecurity Centre of Excellence through a strategic partnership between the Tawazun Council, UAE Cyber Security Council, and Lockheed Martin, enabled through the Tawazun Economic Programme to enhance digital security capabilities and develop local expertise . This initiative aligns with the UAE Cabinet's approval of the comprehensive National Cybersecurity Strategy, based on five main pillars of governance, protection, innovation, establishing and building, and partnership, which seeks to establish a cohesive governance framework for cybersecurity and ensure a secure digital environment. Additionally, the Cabinet approved the API-First Policy, which includes requirements and procedures for ministries and federal entities in their technological systems to ensure rapid interconnection and integration with other systems while regulating relationships between providers and users and enhancing public-private sector partnerships in government services.

Particularly relevant to AI safety, the Council noted the detection of "complex hacking attempts supported by artificial intelligence technologies". Such advancements present new challenges by targeting critical digital infrastructure in ways that traditional security measures may struggle to address. The Council highlighted a growing trend in cyber threats leveraging AI, not just in areas like deepfake creation and social engineering but also in enhancing the sophistication of malicious software, particularly ransomware. The Council warned that "with the growing availability of sophisticated AI-based tools among criminal groups and terrorist organizations, a new wave of precise and multifaceted cyber-attacks is expected to target entities unprepared with adequate detection and response tools". This assessment underscores the urgent need for specialized capabilities to evaluate and counter AI-enabled threats to national security and critical infrastructure.

The UAE possesses an advanced digital infrastructure capable of "handling diverse types of cyber-attacks with high agility, preemptively responding to threats efficiently and within record time" as per the assessment of Dr. Mohamed Al-Kuwaiti, Head of the UAE Cyber Security Council. This existing capability provides a strong foundation upon which an AI Safety Institute could build specialized expertise in evaluating and mitigating risks from frontier AI systems.

The interconnection between cybersecurity and AI safety is becoming increasingly apparent as AI systems grow more powerful and potentially autonomous. An AI Safety Institute could naturally extend from the UAE's cybersecurity expertise to address emerging challenges related to AI-enabled cyber threats, providing a natural starting point for developing more comprehensive AI safety capabilities.

Protecting Strategic National Investments

The UAE has made enormous investments in AI infrastructure, companies, and applications, with initiatives like MGX and partnerships with global technology leaders representing tens of billions of dollars in commitments. The security and long-term value of these investments depend in part on ensuring that powerful AI systems are developed and deployed safely. A dedicated institution for evaluating AI safety would help protect these investments by reducing the risk of AI accidents, misuse, or unintended consequences that could damage public trust or trigger restrictive regulation.

Leading AI private sector actors in the UAE clearly recognize this imperative. For example, this was demonstrated in G42's publication of its Frontier AI Safety Framework in February 2025. The framework sets clear protocols for risk assessment, governance, and external oversight to ensure the safe and responsible development of advanced AI models. A UAE AISI would complement and strengthen such private initiatives, providing independent evaluation capabilities that could validate corporate safety claims and strengthen public confidence in UAE-developed AI systems.

The strategic importance of establishing a UAE AI Safety Institute has been dramatically amplified by the May 2025 agreements with the United States. With the UAE now committed to developing a 5GW AI campus along with investments in US-based data centers of comparable scale, the Emirates' stake in the global AI ecosystem has increased exponentially. It is estimated that Middle Eastern capital flowing into AI infrastructure could approach one trillion dollars, representing one of the largest technology investments in history. This unprecedented scale of investment creates a proportionate imperative for robust security frameworks and evaluation capabilities. An AI Safety Institute would provide the independent oversight necessary to protect these investments by ensuring that the powerful AI systems developed and deployed on this infrastructure meet the highest safety standards. Moreover, the institute would help fulfill the UAE's commitments under the new agreement to implement "robust security assurances" to prevent the misuse or diversion of US technology , creating a virtuous cycle where access to advanced technology is secured through demonstrable safety governance.

Beyond its functional utility for the AI economy within the UAE, globally, establishing an AI Safety Institute would add a new dimension to the UAE's soft power and technological standing on the global stage. As one of the world's most ambitious nations in AI development, the UAE has an opportunity to extend its influence in global AI governance by becoming a leader in AI safety as well. This would reinforce the country's set objectives in advancing responsible AI and ensuring it benefits humanity.

An additional strategic benefit of establishing a UAE AI Safety Institute is the opportunity to deepen cooperation with key international partners, particularly the United States. As concerns grow globally around the misuse or proliferation of advanced AI capabilities—especially model weights—demonstrating a commitment to rigorous, independent safety oversight can serve as a powerful gesture of transparency and alignment with shared security priorities. A national AISI could act as a trusted institution capable of managing sensitive technologies responsibly, thereby reinforcing the UAE's credibility as a reliable partner in AI development and governance. This would support ongoing collaboration with U.S. stakeholders and potentially open doors to further technological and research partnerships grounded in mutual trust.

On the other hand, the absence of a national AI Safety Institute could slow the momentum of the UAE's AI development and hinders its influence in global AI governance fora. As global gatherings on AI safety increasingly look to engage with institutions that represent national capabilities in this space, the lack of a dedicated body may result in the UAE not being well represented while identifying relevant experts or official delegates. In turn, this risks reducing influence in key discussions that shape the future of AI governance, despite its significant investments and contributions to the field. Establishing a national AISI would ensure the UAE having a 'seat on the table' with clear institutional presence and voice in these global conversations.

The geopolitical significance of AI leadership continues to grow, with major powers investing heavily in both development capabilities and governance frameworks. By establishing an AISI, the UAE would further signal its intention to shape not only how AI is developed responsibly but how it is governed—a critical dimension of technological leadership in the 21st century. This would align with the UAE's broader strategy of diversifying influence beyond traditional economic and diplomatic channels into emerging technological domains.

The UAE as a Global AI Infrastructure Hub

Unlike other parts of the world, that face power and energy limitations, the GCC, with easy access to energy and capital, is positioning itself as a global AI hub. Within this context, the US-UAE AI partnership represents a transformative moment

in global AI development, positioning the UAE as an emerging powerhouse in AI infrastructure. This infrastructure leadership creates both an opportunity and responsibility for the UAE to simultaneously lead in AI security governance.

The new US-UAE agreement includes specific security requirements that align with and accelerate the need for a dedicated AI Safety Institute. As part of the partnership, the UAE and the US will align security regulations, to includes safeguards against the diversion of US technology. For example, the US Commerce Secretary emphasized that the data center agreement incorporates "robust security assurances" to prevent the misuse of US technology from its intended purposes . These commitments could be components of the UAE AI Safety Institute's framework, serving as an institutional embodiment of these security guarantees.

The scale of the planned infrastructure—a 5GW AI campus that will host US hyperscalers serving nearly half the global population—creates a unique opportunity to embed security governance at the foundation of this new AI ecosystem . Rather than retrofitting security protocols onto existing systems, the UAE has the unprecedented opportunity to establish AI-aligned security standards and evaluation frameworks from the ground up as this massive infrastructure is developed. A UAE AI Safety Institute could lead this effort, working in coordination with both US technology partners and local stakeholders to ensure that safety considerations are integrated throughout the AI development lifecycle.

Furthermore, the UAE's established strength in cybersecurity provides a natural foundation for expanding into this new domain. The UAE Cyber Security Council has already demonstrated advanced capabilities in countering AI-enabled threats. With the massive influx of AI infrastructure, the scope and complexity of these threats will increase proportionally, requiring specialized expertise in AI security evaluation and mitigation. An AI Safety Institute would bridge the gap between 'traditional' cybersecurity and the novel challenges posed by frontier AI systems, positioning the UAE as a global leader in both AI infrastructure and AI security.

Potential Structure and Functions of a UAE AI Safety Institute

Based on examination of existing AI Safety Institutes worldwide and the UAE's specific context, a UAE AI Safety Institute would benefit from a design that leverages existing strengths while addressing the unique challenges and opportunities of the country's position in the global AI landscape. This section outlines a potential structure and set of functions for such an institution.

Institutional Positioning

Aligned with global practices, the UAE AI Safety Institute could be established as an independent entity. It could be connected to related local institutional agencies, such as the Artificial Intelligence and Advanced Technology Council (AIATC) in Abu Dhabi and the AI Council in Dubai , . This positioning would provide appropriate coordination framework with governmental authorities while maintaining sufficient independence to perform credible evaluations. Similar to international practices, the institute would be granted clear statutory authority to evaluate AI systems deployed in the UAE and to collaborate with international partners on safety standards development.

Given the UAE's unique position between competing AI global powers, the AISI should maintain operational independence from both foreign governments and commercial interests. This independence would enhance the credibility of its evaluations and enable it to serve as a trusted arbiter in assessing systems developed in different geopolitical contexts.

Potential Initiatives Under the UAE National AISI

Based on international benchmarking and contextual analysis, the following provide a list of priorities, as well as locally and globally-focused roles and initiatives that can be positioned under a national AISI:

Local AI Safety Roles

1. **Risk Assessment Framework:** Create and maintain a UAE-specific framework for assessing AI risks that reflects national priorities while sustaining compatibility with international approaches. Such a framework should address risks ranging from near-term concerns like deepfakes and influence operations to longer-term issues like autonomous replication or misalignment, providing a structured approach to evaluating and mitigating potential harms from AI systems deployed in the UAE.

2. **Safety Certification Program:** Practically implement the "UAI Seal of Approval" envisioned in the UAE's National AI Strategy , providing gradual certification levels for AI systems based on rigorous safety evaluations. Such certification could become a valuable signal of quality and trustworthiness for AI systems deployed in the UAE and potentially the broader region, incentivizing developers to prioritize safety in their design and implementation processes.

Potential Global AI Safety Initiatives:

3. **AI Security Research:** Develop the research foundation for establishing the first ASL-4-5 data centers in the world, for example by building on the securing model weights paper by RAND. This initiative would involve creating infrastructure with confidential computing capabilities, implementing advanced cyber red-teaming, and establishing physically and digitally isolated data centers. The UAE, with its substantial resources and commitment to technological leadership, is well-positioned to pioneer these advanced security measures.

4. **Verification Mechanisms:** Position the UAE as an international player in AI verification, leveraging strong synergies with security research. This would include developing expertise in cryptographic techniques, proof-of-work systems, and new verifiable hardware designs that could ensure AI systems operate as intended and remain secure against tampering or manipulation. The development of these verification mechanisms would complement the UAE's existing cybersecurity capabilities while addressing a critical gap in the global AI safety ecosystem.

5. **International Coordination:** Engage actively with the International Network of AI Safety Institutes and other global governance bodies to share insights, methodologies, and evaluations. Position the UAE as a bridge between different approaches to AI governance, leveraging its unique geopolitical position to facilitate dialogue and cooperation between Western and Eastern approaches to AI safety. This function would enhance the UAE's contribution to and influence in shaping global AI governance frameworks.

Conclusion

The rapid evolution of artificial intelligence demands nothing less than a paradigm shift in global governance—a challenge that intersects perfectly with the UAE's transformative ambitions in the AI economy. As the UAE is emerging as a regional AI infrastructure epicenter, it faces both an unprecedented opportunity and a strategic imperative to lead in securing the very systems powering this technological revolution.

The proposed UAE Artificial Intelligence Safety Institute (AISI) represents the critical next phase in this journey: an institutional mechanism to convert infrastructural scale into effective governance frameworks, ensuring the nation's $1.4 trillion AI investments drive global progress rather than peril.

The UAE's achievements-from pioneering the world's first AI ministry to orchestrating transcontinental ventures like the 5GW data campus and Stargate Project-have redefined its role from regional innovator to global infrastructure guarantor. True leadership in the AI era requires equal mastery over the invisible architectures of safety and trust. As hyperscalers begin operations within Abu Dhabi's 1GW AI hub, servicing nearly half humanity with latency-optimized intelligence, the Emirates must pioneer security frameworks that protect not just national interests but the integrity of global digital ecosystems. The AISI would anchor this effort, leveraging the UAE's proven cybersecurity capabilities and newly secured access to advanced US technologies to establish evaluation protocols for frontier systems.

Geopolitically, the Institute would crystallize the UAE's unique position as the indispensable bridge between competing technological spheres. By maintaining robust ties to all sides of the AI ecosystems while adhering to the strict security assurances of the 2025 US pact, the AISI could emerge as the world's first truly global safety arbiter—a neutral ground for evaluating systems developed under divergent governance models. This role amplifies the UAE's strategic positioning across major AI players, advancing beyond the position of being an infrastructure provider into the architect of transnational safety norms in the AI era.

Given the breakneck speed of AI developments, the window for action is narrow. Establishing the AISI now would institutionalize the security commitments embedded in recent agreements while positioning the UAE to shape-rather than follow-emerging global standards. More than a policy choice, this is a strategic necessity: the final piece in building an AI ecosystem where scale, sovereignty, and safety converge to redefine technological leadership for the 21st century. In embracing this challenge, the UAE would transcend its role as an AI power to become the global steward for safe intelligence.

References

1) https://ai.gov.ae/ai_council/
2) https://www.uae-embassy.org/news/abu-dhabi-launches-comprehensive-global-investment-strategy-artificial-intelligence
3) https://www.mubadala.com
4) https://www.g42.ai
5) https://www.mgx.ae/en
6) https://www.bloomberg.com/news/articles/2024-03-11/abu-dhabi-said-to-target-100-billion-aum-for-ai-investment-firm
7) https://www.mediaoffice.abudhabi/en/technology/abu-dhabi-government-accelerates-digital-strategy-with-landmark-microsoft-g42-partnership/
8) https://www.cnbc.com/2024/04/16/microsoft-to-invest-1point5-billion-in-emirati-ai-firm-g42-takes-minority-stake.html
9) https://openai.com/index/announcing-the-stargate-project/
10) https://www.cerebras.ai/press-release/cerebras-and-g42-unveil-worlds-largest-supercomputer-for-ai-training-with-4-exaflops-to-fuel-a-new-era-of-innovation
11) https://mbzuai.ac.ae/news/institute-of-foundation-models/
12) https://semianalysis.com/2025/05/16/ai-arrives-in-the-middle-east-us-strikes-a-deal-with-uae-and-ksa/
13) https://www.reuters.com/world/china/uae-set-deepen-ai-links-with-united-states-after-past-curbs-over-china-2025-05-15/
14) https://www.g42.ai/resources/publications/g42-frontier-ai-safety-framework
15) https://uaelegislation.gov.ae/en/policy/details/the-uae-charter-for-the-development-and-use-of-artificial-intelligence
16) https://ai.gov.ae/wp-content/uploads/2023/05/MOCAI-AI-Ethics-EN.pdf
17) https://www.mofa.gov.ae/en/mediahub/news/2024/10/28/28-10-2024-uae-technology
18) https://uaelegislation.gov.ae/en/policy/details/uae-s-international-stance-on-artificial-intelligence-policy
19) https://en.incarabia.com/uae-launches-ai-charter-to-advance-technology-standards-673188.html
20) https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf
21) Based on the findings of several MBRSG policy research projects.
22) https://www.gov.uk/government/publications/ai-safety-summit-bletchley-declaration
23) https://www.gov.uk/government/news/global-leaders-agree-to-launch-first-international-network-of-ai-safety-institutes-to-boost-understanding-of-ai
24) https://www.agbi.com/ai/2024/05/uae-joins-ai-safety-pledge-at-uk-south-korea-summit/
25) https://www.ansi.org/standards-news/all-news/2024/11/11-25-24-us-launches-international-ai-safety-network-with-global-partners
26) https://www.nist.gov/news-events/news/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international
27) https://www.csis.org/analysis/united-arab-emirates-ai-ambitions
28) https://www.semafor.com/article/11/20/2024/uae-sees-diplomacy-as-key-to-its-space-ai-ambitions
29) https://uaecabinet.ae/en/news/uae-cabinet-approves-national-cybersecurity-strategy-api-first-policy
30) https://www.wam.ae/article/bi9jgsd-idex-2025-uae-establish-cybersecurity-centre
31) https://gulfnews.com/uae/uae-cyber-security-council-calls-for-stronger-vigilance-amid-growing-ai-driven-cyber-attacks-1.500018585
32) https://semianalysis.com/2025/05/16/ai-arrives-in-the-middle-east-us-strikes-a-deal-with-uae-and-ksa/
33) https://thehill.com/policy/technology/5303724-trump-administration-uae-ai-data-center/
34) https://thehill.com/policy/technology/5303724-trump-administration-uae-ai-data-center/
35) https://www.cnn.com/2025/05/15/middleeast/trump-abu-dhabi-ai-center-latam-intl

36) https://digitalpolicyalert.org/event/16961-adoption-of-law-establishing-the-artificial-intelligence-and-advanced-technology-council
37) https://ai.gov.ae/ai_council/
38) https://ai.gov.ae/wp-content/uploads/2021/07/UAE-National-Strategy-for-Artificial-Intelligence-2031.pdf
39) ASL-4 and ASL-5 refer to advanced tiers within Anthropic's AI Safety Levels (ASL) framework, which categorizes AI systems based on their potential catastrophic risks and capabilities. Specifically, ASL-4 represents AI models with capabilities surpassing the best humans in critical areas, significantly increasing both autonomous and catastrophic risks, necessitating stringent security measures such as intelligence-grade physical and cybersecurity protections, sophisticated automated harm detection systems, external audits, and substantial progress in interpretability and alignment research to ensure safety. ASL-5 and beyond (ASL-5+) remain speculative and undefined due to their distance from current technological capabilities, but they are expected to involve even greater qualitative escalations in autonomy and potential for catastrophic misuse. More details: https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf
40) https://www.rand.org/pubs/research_reports/RRA2849-1.html
41) According to RAND, securing model weights is a critical dimension of AI security, requiring measures such as "centralizing all copies of weights to a limited number of access-controlled and monitored systems; reducing the number of people with authorization; hardening interfaces against weight exfiltration; engaging third-party red-teaming; implementing insider threat programs; and incorporating Confidential Computing to secure the weights and reduce the attack surface".

Author(s) and Citation

This Policy Brief was authored by:

Cyrus Hodes
Fellow of Practice,
Mohammed Bin Rashid School of Government

Fadi Salem
Senior Research fellow and Director of Policy Research Dept.,
Mohammed Bin Rashid School of Government

The authors express appreciation to the following experts for their reviews and input at different stages of the production of this Policy Brief:

Ima Bello | Future of Life Institute
Akmaral Orazaly | MBRSG
Renan Araujo | Institute for AI Policy and Strategy
Sheikh Abdur Rahim Ali

The views expressed in this report are those of the author(s) and do not necessarily reflect those of the trustees, officers, and other staff of the Mohammed Bin Rashid School of Government (MBRSG) and its associated entities and initiatives.

Acknowledgements

The authors wish also to express personal appreciation to the following individuals for their input to the different stages of producing this report and for providing essential input and assistance into the report and its related materials:

Eiman Almarzooqi | Shuaib Kunouth

Copyright Information

About MBRSG

The **Mohammed Bin Rashid School of Government** (MBRSG, Dubai) (formerly Dubai School of Government) is a research and teaching institution focusing on public policy in the Arab world. Established in 2005 under the patronage of HH Sheikh Mohammed bin Rashid Al Maktoum, Vice President and Prime Minister of the United Arab Emirates and Ruler of Dubai, in cooperation with the Harvard Kennedy School, MBRSG aims to promote good governance through enhancing the region's capacity for effective public policy.

Toward this goal, the Mohammed Bin Rashid School of Government also collaborates with regional and global institutions in delivering its research and training programs. In addition, the School organizes policy forums and international conferences to facilitate the exchange of ideas and promote critical debate on public policy in the Arab world. The School is committed to the creation of knowledge, the dissemination of best practice and the training of policy makers in the Arab world. To achieve this mission, the School is developing strong capabilities to support research and teaching programs, including:

- Applied research in public policy and management;
- Master's degrees in public policy and public administration;
- Executive education for senior officials and executives; and,
- Knowledge forums for scholars and policy makers.

The MBRSG Research Department focuses on the following seven priority policy areas:

1. Future Government and Innovation
2. Education Policy
3. Health Policy
4. Public Leadership
5. Social Policy, Wellbeing and Happiness
6. Sustainable Development Policy
7. Economic Policy

Scan the code to access MBRSG research:

For more information on research at the Mohammed Bin Rashid School of Government, please visit: http://www.mbrsg.ae/home/research.aspx

كلية محمد بن راشد للإدارة الحكومية
MOHAMMED BIN RASHID SCHOOL OF GOVERNMENT

Mohammed Bin Rashid School of Government
Convention Tower, Level 13, P.O. Box 72229, Dubai, UAE
Tel: +971 4 329 3290 - Fax: +971 4 329 3291
www.mbrsg.ae - info@mbrsg.ac.ae