Research Report GA6
Lili Koerber and Jona Welchert
AMUN 2025

The Sixth Committee of the General Assembly
(Legal)
"Discussing Regulatory Mechanisms to Ensure the Safe Development of AI Technology and Minimizing Potential Risks"

Artificial intelligence (AI) has moved from a far-off dream of science fiction to a force in motion at the heart of the world today. In the guise of recommendation engines calibrated to personal tastes on streaming sites, or smart diagnostic technology at the hospital, the fingerprints of AI are everywhere. Within a decade, the power of machine learning, natural language processing, and generative models expanded at a pace few could have imagined. But this rapid progress has been accompanied by the growing awareness of issues: algorithmic bias, safety of autonomous systems, and potential job displacement.

Compared to previous waves of technological innovation, AI presents challenges that cut across borders, political systems, and economic spheres. The pace of its development raises basic questions: How can we have AI develop safely? What role is played by government? Can international institutions cope? And most importantly, how can society reap the benefits of AI as much as possible and contain harm?

This essay argues that AI safe development needs to be regulated multi-facetly—a policy approach that combines binding law, adaptive standards, global coordination, and ethical standards. Pure laissez-faire policy induces catastrophic misuse, but over-regulation can strangle innovation. The challenge, therefore, is to design mechanisms that are both strict and responsive, forward-looking and open to alteration, and globally harmonized but locally adaptable.

Regulation of breakthrough technology is rarely simple. History has proved that the world tends to respond to crises rather than pre-emptively responding. Nuclear energy control was considered necessary only after the ghastly potential of the atom bomb was revealed in 1945, for example. Air safety regulation became institutionalized as a result of repeated accidents over the first half of the 20th century. The modern regulatory framework of the

drug industry has evolved after disasters like the thalidomide disaster of the 1960s, where inadequate regulation had caused widespread harm.

Those are mere illustrations. Reactive regulation is expensive, and expense typically comes in the form of loss of lives or irreparable destruction. Trust can be lost but decades are spent regaining public confidence. Finally, regulation does not necessarily quash innovation—nuclear power still generates grids, air travel is now the safest method of transportation, and medication still saves millions of lives annually.

AI, in contrast to these earlier instances, evolves at electronic speed. Timely regulation is therefore particularly hazardous. Pre-emptive regulation, and not a luxury, is an imperative. Early experiments such as the Asilomar AI Principles (2017), the OECD AI Principles (2019), and the European Union's ongoing drafting of the AI Act are efforts at installing guardrails before ghastly crashes call for tighter controls.

Artificial intelligence is neither a lone technology nor a collection of technologies but a set of approaches and applications that cut across nearly all aspects of human endeavor, from health and finance to education, transportation, and security. The diversity breeds an intensified problem of exploration and control of the risks. AI is unlike most other current technologies in that it is not only performing pre-coded functions after it has been installed. It learns, adapts, and even produces outcomes that its developers themselves are not able to adequately explain. Thus, the danger it poses is not just technical constraints but gnaws into the ethical, economic, security, and even existential features of human life. Some of the highest-profile challenges come from the manner in which AI systems learn and adapt to repeat human biases.

Algorithms are conceived to be value-free tools, but they reflect the data they are being trained on, and human data is always socially, culturally, and historically biased. Predictive policing algorithms, for instance, were targeting minority groups since the data it was derived from reflected decades of discriminatory policing. Recruitment software had another cautionary tale: trained on résumés taken in overwhelming proportion from male candidates for jobs in the tech sector, some algorithms learned to reject women's applications, infusing gender bias into processes intended to be neutral. When AI systems are deployed at scale in areas such as hiring, credit determination, or health eligibility, these biases have the potential to entrench systemic disadvantage. Ethical concerns transcend equity: when an individual is injured, either by a discriminatory or an incomprehensible machine choice, responsibility is typically undefined, and the injured individual has no way to pursue a remedy. Directly related to these ethical concerns are safety concerns.

AI performs adequately in carefully controlled laboratory environments but often does not perform adequately in open real environments. Autonomous cars are a good case in point. While touted as a method of reducing accidents caused by human error, autonomous cars have already resulted in fatal crashes, not because of ill intent but because perception software based on small data sets was unable to handle novel driving conditions. Medical devices are not different. Machine learning-based diagnosis equipment was found to be absolutely accurate for some diseases, but made others wrong when tested on people outside the category they were trained on. These are bad technical errors, not humble ones, but life-and-death errors. Things become a lot more high-stakes when applying this in a war zone. Killer robots, to borrow the term of foreboding, allow room for accident, miscalculation, or malicious misuse with devastating humanitarian effect. Beyond the immediate threats to security are economic threats that play out on a social stage.

AI offers efficiency and growth but not evenly. Automation risks displacing tens of millions of employees, from backroom assistants to routine roles, from call centers to bank clerks. Pessimists warn that jobs will never come back, but optimists forecast new jobs will emerge, as they have in past industrial revolutions, but unevenly and disruptively. More stable workers, made redundant by machines, will struggle to adjust to newly created ones, particularly if they lack the technical skills needed. In the absence of tough retraining programs and cushions, AI-based automation has the potential to increase inequality and social disintegration. To this is added the concentration of AI potential in the hands of a few multinational conglomerates. The massive investment required to design models up to date with the latest—massive datasets and gargantuan computation—is within the reach of only a few institutions. This focus threatens to create monopoly frameworks reducing competition, stifling innovation, and providing undue power to private interests whose agendas may not be aligned with society as a whole. Perhaps even more alarming are the security issues inherent in the dual-use potential of AI.

The same methods which can accelerate drug discovery or streamline logistics can be applied for nefarious ends. Generative AI has already been shown to be capable of creating realistic deepfakes, videos, and sound files, weapons that could be used to spread disinformation on an unprecedented scale. When institutions are already at risk of being distrusted, such capabilities can potentially destabilize democratic institutions and drive polarization. Cybersecurity risks also intensify. AI technology can be used to make better malware, automate phishing, and attack vulnerabilities more quickly and effectively than human hackers. Perhaps most unsettling is the possibility of misuse of life sciences. Medicine advancing systems and proteins modeling systems would be able to be used to make bio-weapons in the wrong hands. While nuclear technology relies on scarce material and

advanced infrastructure, AI can be disseminated as code, and the traditional containment methods are of much less value. Once advanced models are made available in the public domain, they cannot be withdrawn easily, posing existential risks to international security. And then, most warn, the greatest dangers might not yet be recognizable but could be existential in character.

The likelihood of highly advanced AI systems with autonomous reasoning and self-enhancement poses an alignment problem: how do we ensure that the goals pursued by these systems stay aligned with human intention and values? A mechanized AI created to optimize efficiency, for example, can pursue policies that are technically efficient but socially disastrous, merely because it does not understand the larger context of human priorities. These are speculative possibilities, yet the rapidity of progress makes them impossible to dismiss. Skill sets once forecast decades ahead—e.g., mass language ability or autonomous solving—have already arisen sooner than foreseen. The concern is not suddenly hostile robots, but that poorly aligned objectives, fueled by an autonomous system at scale, could create outcomes disastrous for human existence. Combined, these risks illustrate that AI is not simply another technology but a force capable of reshaping economies, societies, and potentially the very future of human civilization itself.

Ethical breakdowns, safety threats, economic disruption, security vulnerabilities, and survival hazards constitute a continuum of challenges that cannot be easily divided but must be addressed in relation to the others. All must be handled carefully, and considered collectively, they require a regulatory stance as delicate and sophisticated as the technology. Regulatory strategies for artificial intelligence can be imagined in a spectrum that goes from legally enforceable, strict regimes to softer, voluntary approaches.

One end has hard regulation, in which governments pass legislation backed by legally enforceable sanctions. The European Union's AI Act is perhaps the most sensational example of this kind. It establishes complex risk categorizations, exacting a heavy burden on developers of high-risk systems, and threatening severe penalties for non-compliance. The charm of this approach is its accountability; companies cannot disregard regulations without tangible consequences. But the very same inflexibility that makes such regulation potentially great in theory can also render it poor in practice. AI is evolving at a pace unlike most other technologies, and rules crafted at a particular moment may soon make themselves redundant, with regulators playing catch-up on the latest advancements. At the opposite end of the spectrum are soft regulatory mechanisms.

These include such as guidelines, standards, and best practices that organizations take on voluntarily. The United States National Institute of Standards and Technology (NIST), for

example, has developed guidelines as a means to help companies control risk without requiring them to comply. The benefit of this is flexibility. Guidelines can be modified more quickly than legislation and designed to differ from situation to situation. But so is their vulnerability: without law, implementation is at the mercy of goodwill, and businesses faced with a squeeze to cut cost or beat competitors may ignore voluntary standards. International agreements and cooperation arrangements lie between these poles.

AI, by definition, is an international technology. A powerful model learned in one country can be applied in a moment worldwide, and misuse in one region can spread worldwide. It is this reality that has spawned calls for the kind of agreements in the form of nuclear non-proliferation treaties to attempt to create mutual standards for safety and accountability. Such efforts are, however, faced with grave challenges. Geopolitical rivalry, issues of national security, and discrepancies in political regimes serve to make alignment of interests problematic between nations. However, without international cooperation to some extent, self-regulation is likely to prove inadequate to address transnational challenges. Public-private partnerships are another means forward, one trying to harness the strengths of governments with the capacities of the private sector.

An example can be seen with the United Kingdom's AI Safety Institute. It collaborates with big companies to test and examine new-generation AI models without having complete control or autonomy. The collaborations recognize that governments are usually too short on technical expertise to regulate, but nor can businesses be trusted to self-police in the interest of the public. If well designed, these collaborations can leverage the abilities of both sides. Given the complexity of the risks in AI and the limitations of any single regulatory model, a system of layers appears to be needed.

Building on the excellence already being worked on around the globe, this report suggests a multi-stranded approach. It is underpinned by a risk-levelled categorization framework, in order that regulation is proportionate: light responsibilities for tools of low risk but strong obligations for systems with the potential to inflict serious harm. Before being deployed in sensitive domains such as healthcare, transport, or national defense, AI needs to undergo mandatory safety trials on the model of clinical trials in medicine. This would ensure that systems are subjected to a wide range of scenarios and stress-tested for dependability before affecting individuals' lives. Transparency needs to be another cornerstone of the framework.

Developers would be forced to make public documentation on their models, including details about training data, known limitations, and potential misuse environments. Public disclosures would not only benefit regulators but also permit the public and external researchers to scrutinize claims. High-risk systems ought to be subject to mandatory third-party independent audit by reputable third parties authorized to verify compliance. Whistleblower protection is also needed; employees who warn authorities of harmful or unethical practices need to be safeguarded from retaliation, or warning signs will never reach regulators. Finally, because AI is global in nature, there needs to be an international watchdog agency, à la organizations like the International Atomic Energy Agency. This agency

would be able to harmonize standards, monitor risks, and facilitate cooperation across borders. The challenge is how to make such mechanisms operate without strangling innovation.

Over-regulation, if too heavy-handed or inflexible, carries the danger of forcing companies into less regulated jurisdictions or forcing research underground, where it can be little controlled. To avoid this, there are a number of things that can be done. Regulator sandboxes hold out some hope. Such controlled environments enable companies to test novel AI applications with regulatory supervision, facilitating innovation while the safety measures are tested at the same time. Adaptive regulation is another important strategy: rather than enshrining rules in stone, legislation and guidance need to incorporate mechanisms for periodic update as the technology advances. Well-balanced privacy laws also apply, protecting citizens from data exploitation while not overly restricting valid research that must make use of large datasets. Open-source AI also raises the tension between innovation and security.

On the positive side, open models democratize access, enabling researchers, small businesses, and civic organizations to monitor and leverage AI technologies. On the other hand, opening up powerful models with abandon raises the risk of misuse by evil agents who would misuse them for disinformation, cyberattacks, or worse. Maybe some middle ground has to be found where some part of models or data is opened up for real research reasons, but full access is opened and accompanied with shielding. Collectively, these methods lead us toward a regulatory landscape that is not one-dimensional but layered, dynamic, and responsive. Hard rules of accountability are held for where stakes are greatest, soft guidance provides room for manoeuvre, global conventions strive for global consensus, and alliances mediate technical expertise and state power. Under such a regime, risks can be managed without extinguishing the creative flame that makes AI one of the most revolutionary technologies of our time.

Sources:

Bostrom, N., **2014.** *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Russell, S. **2019.** *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Bullock, J.; Chen, Y.-C.; Himmelreich, J. (eds.) **2024.** *The Oxford Handbook of AI Governance*. Oxford: Oxford University Press.

Folberth, A., Jahnel, J., Bareis, J., Orwat, C. & Wadephul, C. **2022.** "Tackling problems, harvesting benefits – A systematic review of the regulatory debate around AI." *arXiv preprint* arXiv:2209.05468.

Erdélyi, O. J. & Goldsmith, J. **2020.** "Regulating Artificial Intelligence: Proposal for a Global Solution." *arXiv preprint* arXiv:2005.11072.

Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Tan, Z.-X., Wing, J., Tenenbaum, J. **2024.** "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems." *arXiv preprint* arXiv:2405.06624.

Comunale, M. & Manera, A. **2024.** *The Economic Impacts and the Regulation of AI: A Review of the Academic Literature and Policy Actions*. IMF Working Paper 2024/065.

Allen, D., Hubbard, S., Lim, W., Stanger, A., Wagman, S. & Zalesne, K. **2024.** "A Roadmap for Governing AI: Technology Governance and Power Sharing Liberalism." *Ash Center Occasional Paper Series*, Harvard Kennedy School. January 2024.

Finocchiaro, G. **2024.** "The Regulation of Artificial Intelligence." *AI & Society*, Volume 39, pp. 1961–1968.

Becker, N., Junginger, P., Martinez, L., Krupka, D., Beining, L. **2021.** "AI at Work – Mitigating Safety and Discriminatory Risk with Technical Standards." *Papers With Code*.

Smolensky, P., McCoy, R. T., Fernandez, R., Goldrick, M., Gao, J., et al. **2022.** "Neurocompositional computing: From the Central Paradox of Cognition to a new generation of AI systems." *arXiv preprint* arXiv:2205.01128.

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., DiCarlo, J., Ganguli, S., Hawkins, J., Koerding, K., Koulakov, A., LeCun, Y., Lillicrap, T., Marblestone, A., Pouget, A., Savin, C., Sejnowski, T., Solla, S., and Tolias, A. S. **2022.** "Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI Revolution." *arXiv preprint* arXiv:2210.08340.

Mohamed, S., Png, M.-T. & Isaac, W. **2020.** "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence." *arXiv preprint* arXiv:2007.04068.

Goldsmith, S. & Yang, J. T. **2024.** "AI and the Transformation of Accountability and Discretion in Urban Governance." *Data-Smart City Solutions*, Bloomberg Center for Cities, Harvard Kennedy School. October 2024.

Temper, M., Tjoa, S., David, L. **2025.** "Higher Education Act for AI (HEAT-AI): a framework to regulate the usage of AI in higher education institutions." *Frontiers in Education*, Volume 10, 2025.