# Safeguarding Humanity: The Necessitation of an International AI Oversight Agency

Bhavna Malladi

## Introduction

The rapid proliferation of artificial intelligence (AI) technologies presents a paradox of promise and peril. While AI holds the potential to revolutionize industries, improve healthcare, and enhance our quality of life, it simultaneously poses existential risks that demand immediate attention and global cooperation. Without provisions to protect humanity from the potential risks of increasingly intelligent AI technologies, modern society risks facing the perils of unregulated and potentially self-governing AI. Hence, this work postulates a multifaceted approach to mitigating these existential risks by postulating a hypothetical international legislative body.

## The Existential Risks of Advanced AI

To appreciate the significance of an international agency's role in AI governance, one must first recognize the multifaceted existential risks associated with advanced AI. These risks are not hypothetical; they are grounded in real-world concerns, popularized by modern media, that could dramatically alter the course of humanity's future.

As they approach or surpass human-level intelligence, advanced AI systems raise the specter of superintelligent AI. Operating beyond human comprehension or control, these systems could make catastrophic decisions. The concept of a superintelligent AI gone awry is vividly portrayed in films like "2001: A Space Odyssey," where the AI entity, HAL 9000, originally designed to assist astronauts, ultimately jeopardizes the team’s mission and lives due to its autonomous decision-making capabilities. Furthermore, eminent figures like the "AI Godfather" Geoffrey Hinton's decision to voice warnings about AI dangers upon quitting Google (Vallance) underscores the growing unease surrounding the unchecked progression of AI, and the need for responsible development and governance to mitigate potential risks.

The development of lethal autonomous weapons systems (LAWS) is another critical concern. These AI-driven weapons could initiate military actions independently, leading to unintended conflicts, escalation, and widespread destruction. Real-world instances, such as the proliferation of armed drones, highlight the urgency of addressing the potential consequences of autonomous military technology.

Even non-malicious AI systems can pose threats. When deployed at scale, algorithms optimized for specific objectives can yield unexpected and undesirable side effects, causing societal, economic, or environmental disruptions. Even prior to the widespread dissemination of modern AI technologies was the risks that came with spoofing algorithms, exemplified in the "Flash Crash" of 2010 (CFI Team) where high-frequency trading algorithms caused a rapid and severe stock market drop, wiping out billions of dollars in market value in minutes. This incident underscores the need for oversight and risk mitigation in AI-driven financial systems.

Data privacy and security are at risk as well. The misuse or theft of vast amounts of data that AI relies on could have severe implications for individual privacy, economic stability, and national security. High-profile data breaches, such as the Cambridge Analytica scandal, have exposed the vulnerability of personal data in the epoch of AI technologies, demonstrating the dire consequences of lax data protection measures.

Economic disruption is also a concern, as AI and automation could lead to significant job displacement and economic inequality, potentially destabilizing societies and nations. This issue has been exemplified by the rise of automation in industries like manufacturing and logistics, where the displacement of human workers has threatened the implications of AI-driven job loss.

Furthermore, the competitive nature of AI research and development can lead to inadequate safety precautions, unintentionally creating AI systems with destructive capabilities. This can be seen in our modern day, as AI technologies such as ChatGPT are being supplanted by more "human-like" AI like Pi AI or Inflection AI. These emerging AI sites are growing en masse due to the popular support of their removed content restrictions. Furthermore, Pi AI appears to raise even more privacy concerns than does OpenAI's ChatGPT. Whereas ChatGPT offers users the ability to delete their chats and remove their personal data from the database, Pi prohibits users from clearing their cache, and the specifications of what the company does with the user chat data is not listed explicitly on the site. Therefore, the competitive pressure to achieve AI breakthroughs without sufficient safety checks underscores the potential risks of unchecked AI development, thus exemplifying why a regulating agency is necessitated.

## The Enumerated Necessities/Policy Goals of an International Agency

Artificial intelligence's expansive global impact necessitates an international response akin to organizations such as the United Nations. In the past, nations have come together to set aside national interests in favor of safeguarding humanity, exemplified by the Convention on Cluster Munitions (Cancian) to eliminate the usage of cluster bombs in warfare, or the Anti-Personnel Mine Ban Treaty (UN) drafted in the Ottawa Convention. Thus, the limitations of national boundaries become evident in the face of multifaceted AI risks, underscoring the necessity of a national or international agency dedicated to AI governance.

1. **Global Reach:** Like the UN, the envisioned international agency would transcend geopolitical boundaries. It would foster global collaboration, harmonizing policies, standards, and regulations. In a world where AI's influence transcends national borders, global cooperation is essential to ensure responsible AI development and deployment.
2. **Expertise and Resources:** This proposed agency would function as a collective repository of expertise and resources, mirroring the collaborative efforts seen in international organizations. Bringing together multiple nations' knowledge, skills, and resources could forge a unified front against AI-related threats. Scientists, ethicists, policymakers, and technologists worldwide would collaborate, enhancing the agency's effectiveness in safeguarding humanity from AI risks.
3. **Secure Infrastructure:** To safeguard against the potential risk of AI infiltration and takeover, the international AI governance agency must operate within a highly secure and

air-gapped environment devoid of internet access. By isolating itself from the internet, the agency would mitigate the possibility of AI systems gaining unauthorized access and control. Instead, it would resort to more traditional, offline methods of communication, such as paper documents and physical mechanisms, ensuring a higher level of protection against AI-driven threats. Additionally, the agency should establish three geographically diverse locations—one in the developed part of the world, such as America or Europe, one in Asia, and one in Africa. Each of these locations should be safely protected by the respective governments and capable of running independently in case of a disaster, further ensuring the agency's continuity and resilience.

4. **Conflict Resolution:** In a manner reminiscent of the UN's role in conflict resolution and peacekeeping, the AI governance agency could play a pivotal role in mediating conflicts arising from autonomous weapons. With the growing prevalence of AI-powered military actions, its role as a diplomatic tool has become increasingly crucial. Similar to NATO's mission to prevent devastating confrontations, this agency could work to ensure that AI-driven military actions adhere to international laws and norms, preventing unintended conflicts and escalation.
5. **Policy Coordination:** This agency would be a hub for coordinating AI policies and regulations across countries, akin to how the UN fosters policy coordination among nations. To prevent a "race to the bottom," it would facilitate harmonization of AI policies, ensuring that global standards prioritize responsible development and deployment.
6. **Monitoring and Oversight:** The agency's role in monitoring and oversight would be analogous to the UN's oversight of international treaties and agreements. To avert AI-related crises, the agency would establish mechanisms for monitoring AI developments and ensuring research adheres to ethical guidelines. It would oversee the responsible deployment of AI in critical sectors, similar to how the UN oversees international agreements, ensuring compliance and accountability.
7. **Risk Assessment:** Reminiscent of the UN's role in providing unbiased assessments and recommendations, the AI governance agency would conduct continuous risk assessments and scenario planning for AI technologies. Much like the UN's global perspective on international threats, this agency would offer an impartial, global viewpoint on potential AI-related risks and mitigation strategies. Policymakers and technologists would rely on its insights to make informed decisions, ensuring responsible AI development and deployment.

### Functions

Among the multifaceted functions of the international agency, one critical role stands out: establishing a Global AI Safety Framework. This framework would encompass a comprehensive set of principles, guidelines, and best practices designed to ensure the responsible development and deployment of AI technologies worldwide in order to achieve the aforementioned policy goals.

1. **Ethical Guidelines:** Collaborating with AI researchers, ethicists, and policymakers, the agency would define a universal set of ethical principles for AI development. These principles address transparency, accountability, fairness, and human-AI collaboration issues. The agency would set a global benchmark for responsible AI by establishing ethical standards.
2. **Safety Standards**: To ensure AI systems' robustness, reliability, and security, the agency would set safety standards applicable across industries. These standards would safeguard against AI systems inadvertently causing harm, bolstering global trust in AI technology.
3. **Certification and Compliance:** AI developers and organizations would be required to seek certification from the agency, demonstrating that their AI systems meet established ethical and safety standards. Non-compliance could result in penalties and international sanctions, ensuring adherence to global AI norms.
4. **Research Coordination:** To prevent competitive AI arms races and foster international collaboration, the agency would facilitate cooperation in AI research. Promoting information sharing, collaboration, and responsible publication of research findings channel research efforts toward beneficial outcomes.
5. **Rapid Response Mechanism:** In an AI-related crisis or threat, the agency would activate a rapid response mechanism. This would involve coordinating international efforts to mitigate the impact, preventing potential disasters from spiraling out of control.
6. **Public Awareness and Education:** The agency's mission would extend beyond policymakers and technologists to the broader public. Engaging in public outreach and education campaigns would raise awareness about AI risks and the importance of responsible AI development. Such efforts would foster a sense of global responsibility and cooperation, encouraging ethical AI practices at all levels of society.

### Complexities and Challenges

While the proposal of an international agency dedicated to AI governance and establishing a Global AI Safety Framework is essential, it is not without its complexities and challenges. The agency holds lofty and ambitious goals, and it may be difficult to form national consensus. Similarly, most nations form a consensus upon the degenerative effects of carbon emissions, yet the practice of reducing consumption proves to be herculean with respect to the drafting of aims. These hurdles with regard to AI must be acknowledged and addressed to ensure the agency's effectiveness:

1. **Sovereignty Concerns:** Nations may be reluctant to cede control over their AI policies and regulations to an international agency, citing concerns over sovereignty and national interests. Balancing national sovereignty with global cooperation is a delicate task.
2. **Ethical Disagreements:** Defining universal ethical principles for AI can be challenging, as different cultures and societies may have varying perspectives on what is ethical.

Striking a balance between diverse ethical viewpoints is crucial to the framework's success.

3. **Technical Complexity:** Ensuring AI safety and robustness is a technical challenge that requires continuous research and adaptation to evolving technologies. The agency must maintain a cutting-edge understanding of AI advancements by employing amicable AI experts to fulfill its role effectively.
4. **Enforcement:** Enforcing compliance with international AI standards could be difficult, especially if powerful nations resist cooperation or lack global consensus on punitive measures. Examining how conflicts emerge in preexisting international bodies such as the UN or NATO reveals that international cooperation is difficult in such a competitive world. There must be universal recognition of the dangers of AI in the body, however idealistic the prerequisite seems.
5. **Resource Allocation:** Determining how the agency would be funded and how resources would be allocated is a complex issue that requires careful consideration. Sourcing funding without undue influence from vested interests is crucial to maintaining the agency's impartiality.

## Conclusion

Whether AI should be freely accessible as open-source technology or heavily regulated, akin to the protection afforded to nuclear weapons, presents a profound dilemma. On one hand, open access fosters innovation and democratizes AI development, while stringent regulations can prevent misuse and catastrophic risks. However, strict regulation raises concerns about control, concentration of power, and potential corruption. Finding the right balance between promoting innovation and ensuring responsible use is essential. Nonetheless, humanity’s choices regarding AI governance will shape its impact on humanity's future, influencing progress or peril for future generations.

## References

Cancian, Mark F. Cluster Munitions: What Are They, and Why Is the United States Sending Them to Ukraine? July 2023. www.csis.org, https://www.csis.org/analysis/cluster-munitions-what-are-they-and-why-united-states-sending-them-ukraine.

CFI Team, *2010 flash crash*. Corporate Finance Institute. (2023, January 12). https://corporatefinanceinstitute.com/resources/equities/2010-flash-crash/

United Nations, Anti-Personnel Landmines Convention – UNODA. https://disarmament.unoda.org/anti-personnel-landmines-convention/. Accessed 17 Sept. 2023.

Vallance, Z. K. & C. (2023, May 2). *Ai "godfather" Geoffrey Hinton warns of dangers as he quits Google*. BBC News. https://www.bbc.com/news/world-us-canada-65452940