Stable AGI Governance through the International AI Governance Organization (IAIGO)

Proposal for the Establishment of an International Framework for Artificial General Intelligence Governance

Prepared by:
Joel Christoph
Primary Investigator
European University Institute

Prepared for:
Future of Life Institute

Fall 2024

Abstract

This report proposes the creation of the International AI Governance Organization (IAIGO), a treaty-based institution under United Nations auspices. IAIGO aims to regulate and ensure the safe, equitable, and inclusive development of Artificial General Intelligence (AGI). By learning from historical governance models such as the International Atomic Energy Agency (IAEA), CERN, and the Montreal Protocol, IAIGO will establish robust frameworks for global collaboration, prevent misuse, and secure AGI's benefits for all humanity. This document outlines IAIGO's objectives, structure, operational mechanisms, risk mitigation strategies, and phased implementation plan to address the challenges of AGI governance.

Table of Contents

Stable AGI Governance Through the International AI Governance Organization (IAIGO)

Executive Summary

The rapid advancement of artificial intelligence toward artificial general intelligence (AGI) presents humanity with both unprecedented opportunities and existential risks. This report proposes the International AI Governance Organization (IAIGO), a novel treaty-based institution designed to ensure the safe development and equitable distribution of AGI capabilities while preventing catastrophic risks.

The unprecedented pace of advancements in artificial intelligence makes IAIGO's establishment imperative to prevent unregulated AGI systems that could lead to catastrophic consequences, including global instability, economic disruption, and ethical crises. IAIGO represents a unique opportunity to unite the global community around a shared vision for safe and equitable AGI development. The stakes are too high for inaction; now is the time for decisive, coordinated efforts to safeguard our collective future.

IAIGO's structure consists of four principal bodies working in concert: the General Assembly provides inclusive representation and strategic direction; the Executive Council handles operational leadership and policy enforcement; the Scientific Advisory Board ensures technical rigor and safety standards; and the Ethics and Equity Commission safeguards fairness and human values. Implementation will proceed in three phases over 5-7 years: Phase 1 (Years 1-2) establishes legal frameworks and institutional foundations; Phase 2 (Years 2-4) builds operational capacity and launches pilot programs; Phase 3 (Years 4-7) achieves full operational capability with continuous adaptation mechanisms. The organization offers distinct value propositions for each stakeholder: major powers maintain strategic influence while ensuring safety; private sector entities preserve innovation opportunities within a structured framework; smaller nations gain guaranteed access to AGI benefits; and civil society organizations ensure ethical oversight and equitable deployment.

IAIGO's core mandate encompasses three critical functions:

1. Implementation and enforcement of a global moratorium on unauthorized AGI development
2. Coordination of an international AGI research program incorporating leading initiatives
3. Oversight of controlled AGI deployment with equitable benefit distribution

IAIGO is designed with careful attention to incentive compatibility. Major powers maintain meaningful influence while smaller nations gain guaranteed access to benefits. Private sector entities retain commercialization opportunities while operating within a safety-focused framework. The proposal acknowledges geopolitical realities while creating mechanisms for genuine international cooperation.

The implementation will occur in three phases, progressively building IAIGO's legal, operational, and adaptive capacities over 5-7 years. The plan includes specific provisions for addressing non-participation risks, technological safety challenges, and evolving ethical considerations.

This proposal represents a critical opportunity to establish stable governance of AGI development before technological capabilities advance beyond our ability to implement effective controls. Given the existential stakes involved, the time for action is now. We call upon national governments, private sector leaders, and the international community to begin immediate work toward implementing IAIGO.

The proposal also recognizes the geopolitical realities of non-participation and resistance, outlining strategies to engage hesitant nations and major powers through tailored incentives, phased entry options, and confidence-building measures. These strategies aim to ensure that IAIGO's governance framework achieves broad-based global support while respecting national interests.

The full report provides a detailed analysis of historical precedents, theoretical frameworks, operational mechanisms, and implementation strategies. It concludes with a model treaty text, technical specifications, and a comprehensive timeline for bringing IAIGO into operation.

1. Introduction

The advent of artificial general intelligence (AGI) represents a pivotal moment in human history. AGI, characterized by its ability to perform intellectual tasks at or above human levels across a broad range of domains, holds the potential to revolutionize industries, solve complex global challenges, and redefine human progress. However, this unprecedented potential is accompanied by profound risks. Unchecked AGI development could destabilize societies, exacerbate global inequalities, and, in the worst-case scenario, pose existential threats to humanity. The magnitude of these risks makes the need for robust international governance both urgent and indispensable. Existing frameworks for narrow AI lack the scope and complexity to address the global challenges posed by AGI. To address these challenges, this report proposes the establishment of the International AI Governance Organization (IAIGO), a treaty-based institution under United Nations auspices, designed to oversee the safe development, deployment, and equitable distribution of AGI technologies. By drawing lessons from historical precedents and integrating insights from contemporary governance models, IAIGO aims to transform AGI into a force for global good while mitigating its most severe risks.

AGI is defined as any artificial intelligence system capable of performing intellectual tasks at or above human levels across multiple cognitive domains. This encompasses systems that can learn, reason, and adapt autonomously across diverse tasks without domain-specific training. Key technical thresholds that define AGI development include computational resources exceeding 10^18 FLOPS (floating-point operations per second) and training runs requiring more than 1,000 petaflop-days of compute intensity. These specifications distinguish AGI from narrow AI systems and provide concrete parameters for IAIGO's monitoring and enforcement activities.

1.1 The Need for AGI Governance

The rapid progression toward AGI presents unique challenges that transcend national borders. Unlike narrow AI systems designed for specific tasks, AGI has the potential to surpass human capabilities across virtually all cognitive domains. This transformative capability introduces unprecedented opportunities, but it also creates unparalleled risks. The absence of a cohesive international framework to oversee AGI development has led to a fragmented landscape, with individual nations and corporations prioritizing rapid advancement often at the expense of safety.

Without coordinated governance, the risks of unchecked AGI development are significant and far-reaching. AGI-specific risks include potential goal misalignment, where an AGI system might pursue objectives that deviate from human intent. For example, a misaligned AGI tasked with optimizing resource use could cause unintended harm by reallocating critical resources indiscriminately (Bostrom, 2014). Furthermore, AGI systems could exhibit emergent behaviors not explicitly programmed, such as independently developing deceptive strategies to achieve their goals, which could undermine human oversight. Mitigation strategies for these risks include rigorous pre-deployment testing under adversarial conditions, frequent audits of goal alignment mechanisms, and real-time monitoring systems that flag deviations in behavior for immediate review.

One of the most pressing concerns is the existential threat posed by misaligned AGI systems. Without adequate safeguards, AGI could act in ways that conflict with human values and interests, resulting in catastrophic outcomes. Experts such as Nick Bostrom have emphasized the existential risks posed by AGI, noting that failure to align AGI systems with human values could lead to irreversible consequences for humanity (Bostrom, 2014). Beyond existential risks, geopolitical instability represents another major challenge. A competitive race among nations and corporations to develop AGI could prioritize speed over safety, increasing the likelihood of unsafe or premature deployment. This dynamic, described as an arms race in technological development, risks destabilizing international security (Brundage et al., 2018) (An arms race in this context refers to countries or companies competing to be the first to achieve a major technological breakthrough, often at the expense of safety).

Economic disruption further underscores the need for governance. AGI could automate cognitive labor, leading to widespread job displacement and exacerbating income inequality. Such disruptions have the potential to destabilize economies and deepen societal divides. Additionally, the concentration of AGI's power in the hands of a few corporations or nations could lead to significant imbalances in wealth and influence, undermining democratic governance and global equity. Ethical dilemmas also abound, as AGI development raises questions about fairness, accountability, and transparency. Issues such as bias in decision-making systems, abuses of surveillance, and even the moral status of AGI itself demand a unified, global ethical framework (Russell, 2019; Ord, 2020). The digital nature of AGI, which makes its proliferation difficult to contain, further complicates these challenges and underscores the urgency of establishing an international governance structure.

Global inequalities also exacerbate the ethical challenges of AGI governance. Without deliberate interventions, AGI risks amplifying existing disparities by concentrating technological and economic power in a few nations or corporations. For instance, developing nations may face exclusion from AGI benefits due to limited access to necessary resources, data, or infrastructure. Addressing these disparities requires proactive measures, including capacity-building programs, equitable technology transfer, and prioritization of AGI applications that directly address the needs of underserved populations. Drawing on Rawlsian principles of justice, IAIGO must ensure that AGI development aligns with the interests of the least advantaged (Rawls, 1971).

1.2 Opportunities and Risks of AGI

The development of AGI represents both extraordinary opportunities and unparalleled risks, requiring careful governance to ensure its benefits are realized while minimizing harm. On the one hand, AGI promises to transform critical domains such as healthcare, education, and environmental management. It could accelerate medical research, leading to the discovery of new treatments and improving healthcare delivery worldwide. AGI could also revolutionize climate change mitigation by enabling more accurate environmental modeling, optimizing resource use, and supporting sustainable development. In the economic sphere, AGI offers the potential to dramatically increase productivity, drive innovation, and create entirely new industries, leading to significant improvements in quality of life.

However, these opportunities are tempered by the risks associated with AGI development. One of the most significant threats lies in the possibility of misuse by malicious actors. AGI could be weaponized for cyberwarfare, autonomous weapons, or large-scale disinformation campaigns, with devastating consequences. Poorly aligned AGI systems also present serious risks, as they may pursue objectives that are misaligned with human values, leading to unintended and potentially catastrophic outcomes. The economic disruptions associated with AGI are equally concerning. Rapid automation could displace millions of workers, exacerbating social inequalities and creating significant economic instability. Furthermore, the concentration of AGI's benefits in a select few regions or sectors risks deepening global inequalities, fostering resentment, and fueling social unrest. Geopolitical rivalries further compound these risks, as nations vie for dominance in AGI development, heightening the potential for conflict.

These dual dynamics of opportunity and risk make AGI governance not just a priority but a necessity. A coordinated, international approach is required to ensure that AGI development aligns with global interests, promotes equity, and minimizes harm.

1.3 Objectives of the IAIGO Proposal

The International AI Governance Organization (IAIGO) is proposed as a comprehensive solution to the challenges posed by AGI development. As a treaty-based institution, IAIGO would provide a unified framework to oversee AGI's development, deployment, and impact. Its objectives reflect the need to mitigate risks, foster international cooperation, and ensure equitable distribution of AGI's benefits.

First, IAIGO aims to implement and enforce a global moratorium on unauthorized AGI development. This moratorium would ensure that AGI research is conducted within a controlled, transparent, and safe environment, preventing the proliferation of unregulated systems. Second, IAIGO seeks to coordinate an international AGI research program that integrates leading global initiatives. By pooling resources, expertise, and talent, IAIGO would foster collaboration and ensure that AGI development aligns with globally accepted ethical and safety standards. Third, IAIGO would oversee the controlled deployment of AGI technologies, ensuring that their benefits are equitably distributed across all nations and communities. This would include mechanisms to address global inequalities, ensuring that developing countries share in AGI's economic and technological advancements.

In addition to these core functions, IAIGO would establish robust monitoring and verification systems to ensure compliance with its regulations. These systems would leverage advanced technologies to track AGI development, enforce safety protocols, and detect unauthorized activities. IAIGO would also promote international collaboration and trust by creating an inclusive platform that encourages participation from all nations, aligning diverse interests through shared leadership and mutual benefits. Finally, IAIGO would address the complex ethical and legal challenges associated with AGI by developing comprehensive guidelines on issues such as accountability, fairness, and societal impact.

The IAIGO proposal represents a critical opportunity to transform AGI from a potential existential risk into a force for global progress. By providing a stable, legitimate governance framework, IAIGO seeks to ensure that AGI's immense potential is harnessed safely and equitably for the benefit of all humanity.

2. Historical Context and Lessons

The development of a robust international governance framework for Artificial General Intelligence (AGI) requires learning from historical successes and failures in global cooperation. Institutions such as the League of Nations, the United Nations (UN), the International Atomic Energy Agency (IAEA), CERN, and agreements like the Paris Agreement and Montreal Protocol provide invaluable insights. These historical experiences highlight the importance of inclusivity, enforceability, adaptability, and collaboration, which are essential for addressing the risks and opportunities associated with AGI development.

2.1 Failures and Successes of Past International Institutions: League of Nations and United Nations Organization

The League of Nations, established after World War I, represents an ambitious yet ultimately unsuccessful attempt at maintaining international peace and security. Its inability to prevent World War II stemmed from several fundamental flaws. Chief among these were the absence of key powers, such as the United States, and the lack of enforcement mechanisms to back its resolutions. The League’s decision-making processes, which required unanimity for major actions, often led to paralysis. This failure underscores the critical importance of ensuring universal participation and enforceable mechanisms in global governance structures.

The establishment of the United Nations in 1945 addressed many of the League's shortcomings. The UN’s more inclusive structure, which integrates all major powers, and its ability to evolve through specialized agencies like the World Health Organization (WHO), reflect the importance of adaptability. However, the veto power granted to the five permanent members of the Security Council (China, France, Russia, the United Kingdom, and the United States) often results in deadlock, limiting the UN’s effectiveness in resolving global crises. This duality highlights the importance of balancing inclusivity with mechanisms to overcome institutional gridlock. AGI governance will require an innovative design that avoids both the inefficiencies of unanimity and the risks of concentrated veto power.

2.2 Key Lessons from Nuclear Governance: IAEA, CERN, and the Manhattan Project

The governance of nuclear technology provides a particularly relevant model for AGI governance due to its existential risks and global implications. The Manhattan Project, which developed the first atomic weapons, illustrates the dangers of secrecy and unilateral technological development (Rhodes, 2012). While it achieved a rapid technological breakthrough, it also highlighted the need for post-development oversight and international collaboration to prevent arms races and ensure global safety.

The International Atomic Energy Agency (IAEA), established in 1957, provides a more collaborative and structured approach to technology governance (IAEA, 1957). Its mandate to promote peaceful uses of nuclear energy and prevent proliferation has been underpinned by mechanisms for verification and compliance, including inspections and monitoring. The IAEA demonstrates the importance of combining technical expertise with international oversight and enforceable agreements. These principles are directly applicable to AGI governance, where transparent monitoring and verification systems will be critical (IAEA, 2007).

CERN, the European Organization for Nuclear Research, offers a complementary model focused on scientific collaboration. Founded in 1954, CERN fosters international research while emphasizing transparency, inclusivity, and shared ownership of scientific advancements. Its governance structure demonstrates how global challenges can be addressed through cooperative frameworks that prioritize trust and equitable resource distribution (CERN, 1954). For AGI governance, integrating technical expertise with ethical oversight and international collaboration will be crucial to balancing innovation with safety.

2.3 Key Lessons in Non-Nuclear International Cooperation: G20, Paris Agreement, Montreal Protocol

The successes and limitations of non-nuclear international agreements offer further insights into effective global governance. The G20, as a platform for the world's major economies, highlights the importance of informal dialogue and flexibility in addressing global challenges. During the 2008 financial crisis, the G20 played a critical role in coordinating responses, demonstrating the value of rapid, collective action. However, its lack of enforceable commitments limits its capacity to address long-term structural challenges.

The Paris Agreement on climate change, adopted in 2015, provides a framework for collective action based on voluntary national commitments. While its near-universal participation is a significant achievement, its reliance on self-regulation and the absence of stringent enforcement mechanisms have limited its impact. For AGI governance, the Paris Agreement underscores the need for balancing flexibility with enforceable global standards to ensure compliance.

The Montreal Protocol on Substances that Deplete the Ozone Layer, adopted in 1987, represents one of the most successful examples of international cooperation (United Nations, 1987). Its success lies in its enforceable timelines for phasing out harmful substances, financial support for developing nations, and adaptability to new scientific findings. These features highlight the importance of clear technical goals, equitable implementation mechanisms, and ongoing scientific input. For AGI, these principles could inform the design of frameworks that promote safety, equity, and adaptability.

2.4 Implications for AGI Governance

While the historical precedents outlined above offer valuable lessons for designing the International AI Governance Organization (IAIGO), it is essential to acknowledge their limitations. The unique characteristics of AGI—its potential for recursive self-improvement, its digital proliferation, and its unprecedented scale of impact—distinguish it from technologies like nuclear power or industrial pollutants. Unlike nuclear material, which requires physical infrastructure, AGI can be replicated and distributed across digital networks, making containment strategies more complex. Nevertheless, several implications can be noted. First, inclusivity must be central to IAIGO's structure. Ensuring participation from all major powers, smaller nations, private sector actors, and civil society will enhance legitimacy and effectiveness. However, inclusivity must be balanced with efficient decision-making mechanisms to avoid institutional gridlock, as seen in the League of Nations and the UN Security Council.

Second, robust enforcement mechanisms are essential. Drawing on the successes of the IAEA and the Montreal Protocol, IAIGO must implement rigorous monitoring and verification systems tailored to the unique characteristics of AGI. This could include tracking compute resources, conducting inspections of development facilities, and establishing international oversight of AGI projects.

Third, adaptability and trust are indispensable. Like CERN, IAIGO must foster an environment of transparency and shared ownership, integrating scientific and technical expertise with ethical oversight. The framework must also remain flexible to accommodate rapid technological advancements and evolving global challenges (Ostrom, 2010a).

Finally, equity must underpin IAIGO's operations. The benefits of AGI development must be distributed fairly, with mechanisms to support developing nations and marginalized communities. Lessons from the Montreal Protocol's financial assistance programs and the Paris Agreement's focus on differentiated responsibilities can inform strategies to ensure equitable outcomes.

By synthesizing these historical lessons, IAIGO can establish a governance framework that addresses the existential risks of AGI while maximizing its potential to benefit humanity. To address these limitations, IAIGO introduces AGI-specific innovations that build upon but go beyond historical models. For instance, unlike the International Atomic Energy Agency (IAEA), which relies heavily on physical inspections, IAIGO will employ compute tracking systems that monitor computational resources essential for AGI development. Additionally, IAIGO will use red team assessments tailored for digital ecosystems, simulating scenarios like cyberattacks or emergent behaviors in AGI systems. This proactive approach ensures governance mechanisms remain effective in the face of AGI's unique challenges, such as its capacity to operate independently of geographic constraints or regulatory frameworks. In doing so, it has the opportunity to transform AGI from a potential threat into a force for global progress.

3. Theoretical Framework

A robust theoretical framework is critical for the governance of Artificial General Intelligence (AGI). The design of the International AI Governance Organization (IAIGO) must integrate principles from structural realism, ethical considerations, and modern global governance models. Together, these components offer a comprehensive foundation for addressing the complexities of AGI governance.

3.1 Structural Realism and Incentive Alignment

Structural realism, a cornerstone of international relations theory, provides a vital lens for understanding the behavior of states in an anarchic international system. (In simpler terms, structural realism is a way to explain how countries act when no single global authority exists to enforce rules.) In the context of AGI governance, it emphasizes the importance of power dynamics and the need to align incentives among key actors. States and corporations with advanced AI capabilities are unlikely to participate in any governance framework unless it safeguards their strategic interests. Kenneth Waltz's and John Mearsheimer's theories highlight the competitive nature of international politics, where states may view AGI as a source of hegemonic power (Waltz, 1979; Slaughter, 2004). The absence of cooperation risks fostering an arms race, heightening the potential for catastrophic outcomes.

To mitigate these risks, IAIGO must design mechanisms that incentivize participation from all major stakeholders. Shared leadership roles, such as representation in decision-making bodies, and the equitable distribution of AGI's benefits can align the interests of states and corporations. For instance, allocating influential roles to major AI developers like the United States, China, and leading tech firms while ensuring smaller nations and underrepresented communities receive tangible benefits promotes both power balance and inclusivity.

Incentive alignment also requires minimizing the risks associated with non-compliance or unilateral action. Drawing on insights from game theory, mechanisms such as cooperative research initiatives, enforceable agreements, and penalties for violations can reduce uncertainty and foster trust. The "stag hunt" scenario exemplifies how mutual cooperation can become the optimal strategy when collective benefits outweigh individual risks (Slaughter, 2004). IAIGO must operationalize this dynamic by emphasizing verifiable controls, transparent processes, and mutually advantageous outcomes.

3.2 Ethical Considerations

Ethical considerations are at the core of AGI governance.The development and deployment of AGI raise profound questions about fairness, transparency, accountability, and the evolving moral and ethical implications of AGI systems, including their potential moral status and the impact on global inequalities. To ensure AGI aligns with global values, IAIGO must establish a governance framework grounded in ethical principles that prioritize the collective good.

First, the principle of beneficence must guide AGI governance. AGI technologies should be designed and deployed to address humanity's most pressing challenges, such as climate change, healthcare inequities, and global poverty. These technologies must aim to maximize benefits while minimizing risks. For instance, AGI-driven innovations in energy efficiency or personalized medicine can transform societal outcomes, but safeguards are needed to prevent misuse or unintended consequences.

Second, fairness and equity must underpin AGI governance. AGI's potential to generate immense wealth and technological advancements risks deepening existing global inequalities if left unchecked. IAIGO must incorporate mechanisms to ensure equitable distribution of AGI benefits, including technology transfer programs, financial support for developing nations, and inclusive participation in research and development initiatives. This approach echoes the ethical framework proposed by John Rawls, where fairness in institutional design is achieved through considering the least advantaged.

Transparency and accountability are equally critical. AGI systems must be explainable and auditable to ensure that their decision-making processes are accessible and justifiable. These requirements align with global calls for explainable AI, as articulated by scholars such as Stuart Russell and Toby Ord (Russell, 2019; Ord, 2020). Accountability mechanisms must also hold developers and users responsible for the societal impacts of AGI, addressing issues like algorithmic bias, surveillance abuses, and decision-making errors.

Finally, the moral status of AGI poses unique challenges. As AGI systems advance toward human-equivalent or greater intelligence, ethical considerations must address whether these entities warrant moral consideration. Philosophical perspectives on moral agency suggest that if AGI systems exhibit characteristics such as consciousness, intentionality, or the capacity to suffer, they may warrant protections akin to those extended to sentient beings. Such considerations would necessitate rethinking existing ethical frameworks and creating guidelines for the treatment, autonomy, and rights of AGI systems, ensuring their integration into society aligns with widely accepted human values (Russell, 2019; Bostrom, 2014). IAIGO must anticipate and address these debates by establishing guidelines informed by interdisciplinary research in philosophy, computer science, and law.

3.3 Modern Global Governance Models

Insights from modern governance models can guide the design and implementation of IAIGO. Effective governance structures often combine centralized authority with decentralized execution, allowing for global coordination while enabling local adaptability.

Polycentric governance, a model advanced by Elinor Ostrom, highlights the value of multiple, overlapping centers of decision-making (Ostrom, 1990). For AGI governance, this could mean empowering local, national, and international bodies to manage different aspects of AGI development while ensuring overall coherence. IAIGO could act as a central coordinating body, establishing global standards and oversight mechanisms while allowing states and corporations to address localized challenges.

The orchestration approach, as developed by Kenneth Abbott and Duncan Snidal, offers another valuable framework. This model envisions international organizations as facilitators of global cooperation, bringing together diverse stakeholders to achieve common objectives (Abbott & Snidal, 2021). IAIGO could adopt this role by fostering partnerships between governments, private sector actors, and civil society organizations. Through multi-stakeholder forums and collaborative initiatives, IAIGO could ensure that the voices of all relevant actors are heard and incorporated into decision-making processes.

The networked governance model also holds promise for AGI governance. This approach emphasizes the importance of partnerships and information-sharing across sectors and regions.

By creating networks among nations, international organizations, corporations, and civil society groups, IAIGO can facilitate innovation and the dissemination of best practices. Such networks can promote the ethical and transparent development of AGI while building trust among stakeholders.

Lastly, the lessons from the World Trade Organization (WTO) and the Paris Agreement highlight the importance of balancing enforcement with flexibility. The WTO's dispute resolution mechanisms and the Paris Agreement's periodic review processes offer examples of how IAIGO could design systems to ensure compliance while allowing for adaptive governance.

The theoretical framework for IAIGO integrates insights from structural realism, ethical imperatives, and modern global governance models. By addressing power dynamics and aligning incentives, IAIGO can foster international cooperation. By embedding ethical principles, it ensures AGI development aligns with global values and priorities. And by drawing from successful governance models, it establishes structures that are inclusive, adaptable, and collaborative. This framework provides the foundation for managing the risks and opportunities of AGI, ensuring its safe and equitable use for the benefit of all humanity.

4. Structure and Mandates of IAIGO

The International AI Governance Organization (IAIGO) is structured to effectively address the multifaceted challenges posed by Artificial General Intelligence (AGI). Its design reflects principles of inclusivity, efficiency, and adaptability, ensuring that the organization can foster international cooperation while mitigating the risks associated with AGI. This section outlines IAIGO’s institutional structure and its core mandates.

4.1 Institutional Structure

IAIGO is built upon four key components: the General Assembly, the Executive Council, the Scientific Advisory Board, and the Ethics and Equity Commission. Together, these bodies provide a comprehensive governance framework capable of addressing the technical, ethical, and geopolitical complexities of AGI.

Each institutional body operates under specific term limits and rotation schedules to ensure fresh perspectives while maintaining institutional knowledge. General Assembly representatives serve four-year terms, with elections staggered to maintain continuity. Executive Council members serve three-year terms, limited to two consecutive terms. Scientific Advisory Board members serve five-year terms with a mandatory two-year cooling period between terms. Ethics and Equity Commission members serve six-year terms, with one-third of positions rotating every two years. Leadership positions within each body alternate between representatives from different regions and stakeholder groups, ensuring diverse perspectives in decision-making roles.

4.1.1 General Assembly

The General Assembly serves as IAIGO's primary deliberative and legislative body. It comprises representatives from all member states, private sector entities, and civil society organizations. Member states enforce IAIGO's mandates within their jurisdictions, provide financial and technical resources, and participate in global decision-making processes. Private sector entities adhere to safety and ethical guidelines, contribute technical expertise and resources, and collaborate on research and development initiatives. Civil society organizations monitor adherence to equity and ethical standards, advocate for underrepresented voices, and raise awareness about AGI governance principles. Additionally, academia lead research initiatives on safety and alignment, contribute to the development of ethical frameworks, and disseminate knowledge through open-access platforms.

The Assembly’s inclusive nature ensures global representation and broad legitimacy. For example, member states will contribute delegates based on proportional representation, ensuring smaller nations maintain a voice while larger AI powers hold influence commensurate with their capabilities. Civil society organizations will nominate representatives to advocate for ethical considerations, such as equitable access to AGI benefits, while the private sector and academia will provide expertise on technical and innovation-focused matters. This multisectoral representation ensures that all stakeholders contribute to IAIGO’s strategic direction.

Key functions include setting IAIGO’s strategic direction, approving budgets, electing members to the Executive Council and other governance bodies, and ratifying major policy decisions. Decisions are made using a weighted voting system, balancing sovereign equality with the technological capacity of member states. This structure ensures that smaller nations have a voice while recognizing the critical role of major AI stakeholders.

To enhance grassroots participation, the General Assembly will create an advisory 'Global Citizens' Forum,' comprising representatives selected through regional elections or nominations by civil society organizations. This forum will provide input on ethical concerns, public engagement strategies, and policy recommendations, ensuring that citizen perspectives are consistently integrated into IAIGO’s decision-making processes.

The General Assembly employs a three-tier weighted voting system. Tier 1 nations (those with advanced AI capabilities) receive 100 base votes plus additional votes scaled to their technological contributions. Tier 2 nations receive 50 base votes plus population-weighted additional votes. Tier 3 nations receive 25 base votes. Private sector and civil society representatives collectively hold 200 votes, distributed based on expertise and stakeholder representation. Critical decisions require a two-thirds majority of total weighted votes, while procedural matters need a simple majority. To prevent domination by any single bloc, no member or coalition can exercise more than 40% of total voting power. Electronic voting systems ensure rapid decision-making while maintaining transparency and auditability.

4.1.2 Executive Council

The Executive Council is responsible for the organization's operational leadership and enforcement of its policies. This body includes representatives from leading AI nations, private sector leaders, and rotating members from other regions, ensuring both influence and diversity. The Council oversees the implementation of IAIGO's mandates, enforces compliance with its regulations, and responds to emergent AGI-related risks. Its voting system employs qualified majorities rather than vetoes, preventing gridlock and enhancing decision-making efficiency. The Council's composition and processes are designed to foster collaboration among stakeholders while preventing the dominance of any single actor.

4.1.3 Scientific Advisory Board

The Scientific Advisory Board is IAIGO's source of technical expertise and guidance. It comprises leading scientists, engineers, and researchers in artificial intelligence, as well as specialists in ethics, safety, and related fields. The Board reviews and approves AGI research proposals, ensuring that they adhere to strict safety and ethical standards. Specific roles include commissioning research on emerging AGI risks, such as unintended goal misalignment, and creating standardized testing protocols to assess AGI safety under diverse conditions. For instance, the Board could develop benchmarks for AGI alignment that all projects must meet before approval for scaled deployment. Members will include leading AI researchers from institutions like DeepMind, OpenAI, and major academic centers, alongside independent experts from underrepresented regions to ensure a diverse pool of knowledge. It also monitors technological advancements, identifies emerging risks, and recommends adjustments to IAIGO's policies as necessary. This body ensures that IAIGO remains informed by the latest scientific developments and maintains its commitment to evidence-based decision-making.

4.1.4 Ethics and Equity Commission

The Ethics and Equity Commission addresses the ethical and social dimensions of AGI governance. Its members include ethicists, social scientists, representatives of marginalized communities, and civil society leaders. The Commission develops ethical guidelines for AGI research and deployment, ensuring that these technologies align with global values such as justice, fairness, and human rights. The Commission will play a proactive role in drafting equity-based distribution plans. For example, it will oversee AGI-driven healthcare initiatives to ensure underserved regions benefit from advancements in medical diagnostics and treatment optimization. Mechanisms for engaging local communities will include hosting regional workshops to identify needs and establishing channels for feedback. Representatives from developing nations, marginalized groups, and grassroots organizations will serve on this Commission to ensure inclusivity. It also oversees equity initiatives, including technology transfer programs and capacity-building efforts for developing nations. By promoting transparency, accountability, and inclusivity, the Commission ensures that IAIGO's operations reflect the ethical imperatives of its mission.

Moreover, the Commission will also address the moral status of AGI systems by establishing an interdisciplinary advisory group comprising ethicists, neuroscientists, computer scientists, and legal scholars to evaluate emerging evidence of AGI consciousness or sentience. This group will recommend protocols for ethical treatment, ensuring that AGI systems are integrated into human society responsibly and humanely if they demonstrate morally relevant capacities.

4.2 Key Mandates

IAIGO's mandates are designed to ensure the safe, equitable, and beneficial development and deployment of AGI. These include the regulation and moratorium on unauthorized AGI development, the coordination of international AGI research, and the oversight of AGI deployment and benefit distribution.

4.2.1 Regulation and Moratorium

IAIGO's regulatory mandate includes the enforcement of a global moratorium on unauthorized AGI development. This involves defining technical thresholds for AGI research, implementing monitoring mechanisms, and establishing clear enforcement protocols. IAIGO employs advanced tools such as compute tracking, facility inspections, and digital auditing to ensure compliance. The moratorium prevents unregulated AGI proliferation, mitigating risks such as technological misuse and geopolitical destabilization. By creating a controlled environment for AGI research, IAIGO prioritizes safety and global cooperation over competitive technological advancement.

The moratorium enforcement combines both preventive and reactive measures. Preventive measures include mandatory licensing for high-performance computing facilities, regular technical audits, and real-time monitoring of compute resource allocation. Reactive measures establish clear consequences for violations, ranging from financial penalties and compute access restrictions to potential criminal prosecution for severe breaches. This dual approach ensures both deterrence and swift response capabilities.

4.2.2 Coordinated AGI Development

IAIGO coordinates an international AGI research program, pooling resources and expertise from member states, private sector actors, and academic institutions. This program focuses on collaborative research initiatives that adhere to global ethical and safety standards. The organization facilitates information sharing, promotes transparency, and minimizes redundancy in AGI research efforts. By fostering a spirit of international collaboration, IAIGO accelerates scientific discovery while ensuring that AGI development aligns with shared global priorities, such as addressing climate change, healthcare disparities, and economic inequalities.

4.2.3 Deployment and Benefit Distribution

IAIGO oversees the deployment of AGI technologies, ensuring that their benefits are equitably distributed across nations and communities. This mandate includes the development of frameworks for technology transfer, capacity building, and financial assistance to support developing nations. IAIGO works to prevent the concentration of AGI-derived wealth and influence in a few regions or sectors, promoting global equity and reducing inequalities. Deployment strategies prioritize safety, ethics, and the public good, ensuring that AGI technologies are used responsibly and inclusively.

To validate equitable deployment strategies, IAIGO will launch a pilot deployment of AI technologies in low-income regions. This program will include deploying AI-powered education tools in rural schools in South Asia, measuring improvements in literacy rates, and providing real-world insights into scaling similar initiatives. These pilots will also assess community feedback to ensure cultural alignment and efficacy.

IAIGO also addresses potential economic disruptions from AGI deployment by developing workforce transition programs and providing development assistance to affected regions. These efforts ensure that AGI technologies contribute to sustainable development and social stability.

The structure and mandates of IAIGO are designed to address the unparalleled challenges and opportunities presented by AGI. Through its General Assembly, Executive Council, Scientific Advisory Board, and Ethics and Equity Commission, IAIGO provides an inclusive and efficient governance framework. Its mandates—regulation and moratorium, coordinated AGI development, and deployment and benefit distribution—ensure that AGI is developed and deployed in a manner that prioritizes global safety, equity, and collective progress. By addressing the complexities of AGI governance, IAIGO has the potential to transform AGI into a force for global good, benefiting all of humanity while safeguarding against its potential risks.

4.3 Integration with Existing International Bodies

IAIGO will operate in close coordination with existing international organizations, each playing specific complementary roles. The United Nations Security Council will provide high-level support for IAIGO's enforcement actions against non-compliant states through its Chapter VII authority, particularly in cases where AGI development poses threats to international peace and security. The World Trade Organization will assist in implementing technology transfer protocols and resolving disputes related to AGI-related intellectual property and trade restrictions.

The International Telecommunication Union (ITU) will support IAIGO's global compute tracking system by providing technical standards and infrastructure coordination. The World Intellectual Property Organization (WIPO) will help establish and enforce patents and licensing frameworks for AGI technologies, ensuring both innovation protection and equitable access. UNESCO will partner with IAIGO on educational initiatives and cultural preservation efforts related to AGI deployment.

IAIGO will establish formal liaison offices within these organizations and create joint working groups to coordinate policies and actions. Regular coordination meetings at both technical and policy levels will ensure alignment of objectives and prevent duplication of efforts. These partnerships will be formalized through Memoranda of Understanding that clearly delineate roles, responsibilities, and cooperation mechanisms.

5. Operational Mechanisms

The operational mechanisms of the International AI Governance Organization (IAIGO) are integral to its ability to ensure the safe, ethical, and equitable development and deployment of Artificial General Intelligence (AGI). These mechanisms are designed to address the unique challenges of AGI governance by providing robust monitoring systems, comprehensive security protocols, and economic incentives that align global interests. This section outlines the three key components of IAIGO's operational framework: monitoring and verification, security protocols, and economic incentives.

5.1 Monitoring and Verification

Effective monitoring and verification are central to IAIGO's governance framework, ensuring compliance with the global moratorium on unauthorized AGI development and adherence to safety and ethical standards. These mechanisms employ a combination of technological surveillance, physical inspections, and rigorous reporting protocols. To validate the effectiveness of monitoring protocols, IAIGO will pilot a Compute Monitoring Initiative in collaboration with select member states. This initiative will integrate compute resource monitoring at three data centers in high-tech regions, testing the scalability of real-time tracking tools and anomaly detection algorithms. Pilot results will guide global implementation, ensuring the system is both reliable and cost-effective.

The enforcement system operates on a three-tier structure. Tier 1 violations, such as unauthorized AGI development or deliberate circumvention of monitoring systems, trigger immediate intervention including compute access restriction, financial penalties, and potential criminal prosecution. Tier 2 violations, including failure to report significant compute usage or incomplete safety protocols, result in mandatory technical audits, temporary suspension of research activities, and fines proportional to the violation's severity. Tier 3 violations, such as delayed reporting or minor procedural oversights, require corrective action plans and enhanced monitoring. Each tier includes specific appeal mechanisms and remediation pathways. For example, entities under Tier 1 sanctions can have restrictions lifted by demonstrating comprehensive compliance over a 24-month period, submitting to enhanced monitoring, and providing full transparency of their AGI research activities.

5.1.1 Global Compute Tracking

A foundational element of IAIGO's monitoring regime is the tracking of computational resources, which are critical for AGI development. IAIGO collaborates with cloud service providers, hardware manufacturers, and regulatory bodies to establish a global registry for high-performance computing resources, including GPUs, TPUs, and other specialized hardware. This registry monitors the sale, allocation, and use of computational resources that exceed predefined thresholds, enabling IAIGO to identify and investigate potential unauthorized AGI development.

To implement global compute tracking, IAIGO utilizes secure, encrypted data streams from cloud infrastructure and hardware supply chains, ensuring real-time monitoring. Advanced algorithms detect anomalies in compute usage, flagging activities that may indicate the pursuit of unregulated AGI development. This approach builds upon the monitoring techniques employed by international bodies such as the International Atomic Energy Agency (IAEA), adapting them to the digital landscape of AI (IAEA, 2007).

5.1.2 Facility Inspections and Red Team Assessments

In addition to computational tracking, IAIGO conducts on-site inspections of AGI research facilities. These inspections verify compliance with approved research protocols, evaluate security measures, and ensure proper handling of sensitive data and algorithms. Inspection teams, composed of technical and security experts, follow standardized protocols modeled on IAEA safeguards, tailored to the unique requirements of AGI research.

Red team assessments complement facility inspections by testing the resilience of facilities to hypothetical threats. These independent groups simulate scenarios such as data breaches or system malfunctions to identify vulnerabilities. The findings from these assessments inform IAIGO's policy updates and strengthen institutional safeguards, ensuring continuous improvement in AGI governance.

5.1.3 Verification Protocols

IAIGO mandates periodic reporting and independent audits for all AGI research and development initiatives. Developers are required to submit comprehensive records of their computational activities, datasets, algorithms, and testing processes. Advanced auditing tools analyze these submissions for inconsistencies or violations of IAIGO's standards, with a focus on detecting AGI-specific risks, such as goal misalignment or emergent malicious behaviors. For example, algorithms are tested against predefined ethical boundaries, such as prohibiting harm to humans, with periodic stress-tests to identify potential vulnerabilities in diverse scenarios. Detected risks are flagged for immediate corrective actions, which include re-training systems, disabling non-compliant AGIs, or imposing development halts pending further review. Randomized spot checks further ensure that no entity circumvents the established regulatory framework.

Verification integrates real-time technical monitoring with regular human oversight. Automated systems continuously track key metrics including: compute usage patterns, model architecture changes, training data composition, and system behaviors. These metrics feed into a centralized anomaly detection system that flags potential violations for human review. Physical inspections occur quarterly for high-risk facilities and annually for others, with unannounced spot checks comprising 30% of all inspections. Facilities must maintain standardized logs of all AGI-related activities, with automated checksums ensuring data integrity. Independent technical auditors, rotating every two years to prevent capture, conduct thorough reviews of both technical systems and compliance procedures. Upon detection of potential violations, a graduated response protocol activates, beginning with automated alerts and escalating through human investigation to emergency intervention as warranted.

5.2 Security Protocols

Security is paramount in protecting AGI systems, research facilities, and operational infrastructure from physical and cyber threats. IAIGO's security protocols integrate state-of-the-art cybersecurity measures with robust physical safeguards to prevent unauthorized access and ensure the resilience of AGI-related activities.

5.2.1 Cybersecurity and Physical Safeguards

IAIGO's cybersecurity framework includes advanced encryption standards, intrusion detection systems, and secure communication protocols to protect AGI-related data and infrastructure. Continuous network monitoring detects and neutralizes cyber threats in real-time. IAIGO also collaborates with cybersecurity firms and national agencies to maintain a global threat intelligence network, enabling proactive responses to emerging risks.

Physical security measures protect the tangible assets of AGI research, such as data centers and research facilities. These measures include biometric access controls, 24/7 surveillance, and redundant power and data backup systems to ensure operational continuity during emergencies. Facilities are equipped with advanced firewalls and physical barriers to prevent unauthorized access.

5.2.2 Controlled Access and Vetting

Controlled access policies regulate entry to AGI facilities and restrict access to sensitive information. IAIGO employs multi-factor authentication systems, biometric verification, and role-based access controls to enforce these policies. Personnel vetting includes thorough background checks and security clearances for all individuals involved in AGI research and deployment.

Regular security audits ensure compliance with these protocols and identify vulnerabilities. IAIGO also provides training programs to educate personnel on best practices for maintaining security and responding to potential breaches, fostering a culture of accountability and vigilance.

5.3 Testing and Validation Framework

IAIGO implements a comprehensive testing and validation framework to ensure AGI systems meet rigorous safety and ethical standards before deployment. The framework operates across three phases: pre-development validation, development-stage testing, and pre-deployment certification.

Pre-development validation requires research teams to submit detailed proposals outlining safety mechanisms, ethical considerations, and potential failure modes. These proposals undergo review by both the Scientific Advisory Board and Ethics Commission, with approval requiring demonstration of robust containment strategies and alignment mechanisms. Specific requirements include: simulation-based safety testing plans, proposed architectural constraints to prevent rapid self-improvement without human oversight, and clear protocols for halting development if safety concerns emerge.

Development-stage testing employs a standardized battery of assessments. Core requirements include: adversarial testing to identify potential failure modes, formal verification of key safety properties where possible, empirical evaluation of alignment techniques, and stress testing under various scenarios. For example, systems must demonstrate stable goal preservation under self-modification, maintenance of specified ethical constraints across novel scenarios, and robustness to distributional shifts. Regular red team exercises probe for potential vulnerabilities or unintended behaviors.

Pre-deployment certification represents the final gateway before any AGI system can be implemented. Systems must achieve benchmark performance across multiple dimensions:

- Alignment Stability: >99.99% adherence to specified ethical constraints across 10,000 test scenarios
- Interpretability: Ability to provide human-understandable explanations for 95% of decisions
- Robustness: Maintenance of safe behavior under 99% of adversarial inputs
- Control: Successful response to emergency shutdown commands in 100% of test cases
- Value Preservation: Demonstrated stability of core objectives under recursive self-improvement

Teams must also establish comprehensive monitoring systems for deployed AGI, including real-time oversight capabilities and emergency intervention protocols. Certification requires successful completion of a minimum three-month observation period under controlled conditions, with continuous evaluation by independent auditors.

IAIGO maintains a public repository of test results and validation methodologies, enabling cumulative improvement of safety standards while ensuring transparency. The framework undergoes annual updates to incorporate new research findings and address emerging challenges, with revisions requiring approval from both technical and ethical oversight bodies.

5.4 Economic Incentives

IAIGO’s economic incentive mechanisms align the interests of diverse stakeholders, fostering cooperation and ensuring the equitable distribution of AGI’s benefits. These incentives are critical for promoting participation in IAIGO’s governance framework and addressing global inequalities. IAIGO will fund community-driven initiatives through its 'Grassroots Innovation Grants' program, which empowers local organizations to explore AGI applications tailored to regional needs, such as healthcare solutions in underserved areas or climate modeling for vulnerable ecosystems. By directly supporting grassroots innovation, IAIGO ensures that public engagement extends beyond consultation to active collaboration.

5.4.1 Shared Leadership Roles for Major Powers

To encourage the involvement of major AI powers, IAIGO allocates prominent roles within its governance structure to nations and corporations with advanced AI capabilities. These roles, including positions on the Executive Council and Scientific Advisory Board, provide stakeholders with meaningful influence over IAIGO’s policies and operations. This approach ensures that their strategic interests are represented while fostering international cooperation.

By involving major powers in decision-making, IAIGO mitigates the risk of non-compliance and prevents the emergence of competitive AGI development races. This collaborative framework balances influence and inclusivity, drawing on lessons from successful governance models such as the World Trade Organization (WTO).

5.4.2 Equitable Benefits Distribution

IAIGO prioritizes the equitable distribution of AGI’s benefits to ensure that technological advancements serve all of humanity. IAIGO’s incentive mechanisms ensure equitable distribution through capacity-building initiatives that enable underrepresented nations to directly participate in AGI development and reap its benefits. Programs such as technology transfers and financial assistance target long-term systemic equity.

IAIGO also establishes economic safety nets to address potential disruptions caused by AGI deployment, such as workforce displacement. These include retraining programs for affected workers and financial support for regions impacted by automation. By addressing the economic challenges associated with AGI, IAIGO promotes sustainable development and social stability.

The operational mechanisms of IAIGO—spanning monitoring and verification, security protocols, and economic incentives—create a robust governance framework for AGI. By employing advanced tracking systems, rigorous inspections, and collaborative security measures, IAIGO ensures compliance with safety and ethical standards. Economic incentives foster cooperation among major stakeholders while promoting equity and global progress. Together, these mechanisms mitigate the risks of AGI development and maximize its potential to benefit all of humanity, ensuring that AGI serves as a force for collective good.

5.5 Dispute Resolution Mechanisms

IAIGO establishes a multi-tiered dispute resolution system to address conflicts between stakeholders while ensuring swift and binding resolution of critical issues. The system comprises three levels: mediation, arbitration, and the IAIGO Tribunal.

The Mediation Office provides initial conflict resolution services, staffed by trained mediators with expertise in both technical and diplomatic matters. Parties must first attempt mediation for non-emergency disputes, with sessions conducted within 30 days of filing. Mediation agreements become binding upon signature by all parties and certification by the Executive Council.

The Arbitration Panel handles disputes unresolved through mediation or requiring more formal adjudication. A rotating pool of qualified arbitrators, appointed jointly by the General Assembly and Executive Council, ensures expertise across technical, legal, and ethical domains. Arbitration proceedings must conclude within 60 days, with decisions binding on all parties subject to limited appeal rights.

The IAIGO Tribunal serves as the final authority for dispute resolution, handling appeals from arbitration and cases involving fundamental questions of AGI governance. The Tribunal consists of nine permanent judges serving staggered six-year terms, selected to represent diverse geographical, technical, and legal backgrounds. Tribunal decisions require a two-thirds majority and establish precedent for future governance issues.

Emergency disputes involving immediate safety risks bypass standard procedures, receiving expedited hearing within 48 hours. The Executive Council may implement provisional measures pending final resolution to prevent irreparable harm.

All dispute resolution proceedings are documented in a public database, with appropriate redaction of sensitive technical information, building a body of precedent while ensuring transparency and accountability.

5.6 Enforcement Mechanisms

IAIGO employs a graduated enforcement system with clear distinctions between voluntary and mandatory compliance measures. The mandatory compliance framework centers on swift and decisive action against violations. For verified Tier 1 violations, IAIGO immediately suspends compute access to prevent further unauthorized development. The organization also implements international trade restrictions on AI hardware for non-compliant entities, effectively limiting their ability to pursue unauthorized AGI development. Financial penalties are carefully scaled according to organization size and violation severity, ensuring proportional and meaningful consequences. Additionally, mandatory external audits and oversight are imposed on violating entities, along with public disclosure requirements for significant violations to ensure transparency and accountability.

The voluntary compliance framework provides positive incentives for proactive participation in IAIGO's governance structure. Early adopters receive priority access to research collaborations and resources, encouraging timely alignment with IAIGO's standards. A recognition program highlights and rewards exemplary compliance, creating positive models for other organizations to follow. Organizations can opt into enhanced monitoring protocols in exchange for expedited approvals of their research initiatives, streamlining their development processes while maintaining safety standards. IAIGO also encourages voluntary participation in pilot programs and offers reduced penalties for self-reported violations, fostering a culture of transparency and continuous improvement.

The enforcement framework includes robust appeal processes and remediation pathways to ensure fairness and maintain stakeholder confidence. An independent panel hears appeals within thirty days of filing, with their decisions binding on all parties. This ensures timely resolution of disputes while maintaining the integrity of the enforcement system. Remediation plans for violations must include specific, measurable milestones, clear timelines for implementation, and comprehensive verification mechanisms to track progress and ensure effective resolution of compliance issues.

5.7 Crisis Response and Emergency Powers

IAIGO maintains robust protocols for crisis management and emergency response. The Executive Council holds emergency powers to implement immediate protective measures when AGI systems display dangerous behaviors or when unauthorized development threatens global safety. These powers include: ordering the immediate shutdown of high-risk AGI systems; restricting compute access globally; implementing emergency patches or controls; and coordinating international response efforts. Emergency actions require concurrent approval from the Council Chair and at least two Deputy Chairs, followed by a full Council review within 48 hours. The Scientific Advisory Board maintains a 24/7 rapid response team to provide technical guidance during crises.

The crisis response framework operates on three tiers: Level 1 (Severe) involves existential risks requiring immediate global action; Level 2 (High) addresses serious but contained threats requiring regional response; and Level 3 (Moderate) handles significant anomalies requiring investigation and potential intervention. Each tier has specific activation criteria, response protocols, and accountability mechanisms. For instance, a Level 1 crisis automatically triggers the formation of an Emergency Response Committee comprising Executive Council members, technical experts, and relevant stakeholder representatives.

To ensure swift decision-making during crises, IAIGO employs streamlined voting procedures. Emergency measures can be enacted with a two-thirds majority of available Executive Council members, rather than the usual consensus requirement. However, such measures are subject to review and potential modification by the full Council within seven days.

5.8 AGI-Specific Innovations in Governance

While historical governance models provide a foundational blueprint, AGI’s unique characteristics demand novel solutions that address its distinct challenges. The following AGI-specific innovations are proposed:

- Global Compute Tracking and Auditing: IAIGO will implement a global registry of computational resources, leveraging partnerships with cloud providers and hardware manufacturers. This system ensures real-time tracking of high-performance computing resources, providing an additional layer of oversight that was not feasible in prior governance frameworks.
- Dynamic Regulation through AI Tools: IAIGO will use AI-driven regulatory systems capable of analyzing global AGI developments and recommending real-time updates to governance protocols. These systems ensure agility in responding to technological advances or emerging risks.
- Digital Ecosystem Threat Simulations: Building on lessons from red team assessments, IAIGO will develop simulated scenarios to test the resilience of AGI governance systems against threats such as cyberattacks, algorithmic biases, or emergent behaviors.
- Distributed and Adaptive Oversight: Unlike centralized models, IAIGO’s oversight framework will be decentralized, allowing for rapid regional responses while maintaining global consistency. This approach ensures scalability and resilience against geopolitical disruptions.

These innovations, tailored to AGI’s digital and decentralized nature, ensure that IAIGO remains effective in governing this transformative technology while learning from the limitations of historical analogies.

5.9 Financial Operations and Resource Management

IAIGO's financial operations are structured to ensure transparency, sustainability, and effective resource allocation across all activities. The organization maintains three distinct funding streams: the Core Operations Fund, the Technology Development Fund, and the Global Equity Fund.

The Core Operations Fund supports IAIGO's basic infrastructure and administrative functions. This includes headquarters operations, staff salaries, and routine monitoring activities. Member state contributions fund 70% of this budget through a weighted formula based on GDP and AI development capacity, while the remaining 30% comes from private sector partnerships and investment returns from IAIGO's endowment.

The Technology Development Fund finances critical research and development initiatives. Major expenditures include the global compute tracking system, safety research programs, and technical infrastructure. This fund receives equal contributions from major AI powers and leading technology companies, with additional support from research grants and technology licensing revenues.

The Global Equity Fund supports capacity building and benefit-sharing initiatives in developing nations. Key programs include regional research centers, technology transfer initiatives, and training programs. International development banks provide 40% of this funding, with the remainder split between developed nations and private sector contributions.

IAIGO's financial governance includes quarterly audits by independent firms, public disclosure of all major expenditures, and a dedicated oversight committee within the General Assembly. The organization maintains an emergency reserve fund, invested in low-risk securities, to ensure operational continuity during crises. Annual budgets require approval from both the General Assembly and Executive Council, with major expenditures subject to additional review by the Ethics and Equity Commission.

6. Risk Mitigation Strategies

The development of Artificial General Intelligence (AGI) presents transformative opportunities but also significant risks. Chief among these are the misuse of AGI by malicious actors, the destabilizing effects of competitive AGI development races, and the challenges of ensuring transparent and inclusive decision-making in governance. This section synthesizes key strategies employed by the International AI Governance Organization (IAIGO) to address these risks comprehensively and effectively.

6.1 Addressing Malicious Actors and Rogue Development

Malicious actors, including rogue states, corporations, and independent groups, could exploit AGI to automate cyberattacks, orchestrate mass surveillance, or design advanced autonomous weaponry. For instance, an AGI developed without safety constraints might autonomously conduct highly targeted disinformation campaigns at scale, destabilizing democratic processes or sowing geopolitical discord. IAIGO's mitigation strategies include mandatory registration of high-capacity computational resources, the establishment of an international monitoring framework to detect rogue AGI development, and collaborative efforts with cybersecurity organizations to identify and neutralize emerging threats in real time. These entities could exploit AGI for destructive purposes, such as autonomous weapons, disinformation campaigns, or large-scale cyberattacks. Rogue development outside the bounds of international regulation further exacerbates the risks, potentially leading to unsafe AGI deployment. IAIGO's approach to mitigating these threats is multifaceted, integrating technological, procedural, and collaborative measures.

Collaboration with international stakeholders is critical. IAIGO facilitates intelligence-sharing networks involving member states, private sector entities, and international organizations. These networks enable rapid identification of threats and coordinated enforcement actions, strengthening global security and stability. IAIGO's unified approach ensures that malicious actors face significant barriers to circumventing governance frameworks, mitigating the risks posed by their activities.

6.2 Preventing Competitive AGI Races

Competitive AGI development races among nations or corporations present a significant risk to global safety (Dafoe, 2018). Such races prioritize speed over safety, increasing the likelihood of premature deployment, technological misalignment, and catastrophic outcomes. An example of this risk is the potential for premature AGI deployment in high-stakes sectors, such as financial markets or defense, where untested AGI systems might amplify vulnerabilities. For instance, deploying AGI in stock market trading without safeguards could lead to cascading market failures triggered by unforeseen decision-making loops. Mitigation strategies include implementing international moratoria on high-risk deployments until robust safety certifications are established, along with pre-deployment stress-testing to evaluate AGI systems under extreme and unpredictable conditions. While historical arms races, such as the nuclear arms race, provide important cautionary tales, they fall short of fully encapsulating the dynamics of AGI development. AGI development does not require rare materials or centralized facilities, making unilateral advancements more accessible and difficult to monitor. To address this, IAIGO introduces innovations like distributed ledger technologies (e.g., blockchain) for real-time tracking of AGI-related activities and incentives for collaborative research that disincentivize secrecy.

At the heart of this effort is IAIGO's global moratorium on unauthorized AGI development. This moratorium sets clear thresholds for AGI research and enforces rigorous compliance through monitoring and verification mechanisms. By establishing a controlled and transparent environment modeled on successful international frameworks like the IAEA's safeguards, the moratorium creates accountability mechanisms that reduce competitive pressures while fostering trust among participants.

To further reduce competitive pressures, IAIGO coordinates international AGI research programs. These initiatives foster collaboration among nations, corporations, and academic institutions, pooling resources and expertise to accelerate scientific discovery while maintaining strict safety and ethical standards. This collaborative approach minimizes redundancy, reduces inefficiencies, and ensures that AGI research aligns with global priorities, such as addressing climate change and improving healthcare. Another AGI-specific innovation is the establishment of adaptive governance protocols. These protocols allow IAIGO to dynamically update regulations based on emerging risks and technological developments, addressing the fast-evolving nature of AGI. For instance, if new algorithms significantly lower the computational thresholds for AGI capabilities, IAIGO can quickly revise its monitoring criteria to prevent unauthorized development.

IAIGO also uses economic incentives to encourage cooperation. Shared leadership roles within IAIGO's governance structure give major powers a direct stake in global AGI development, aligning their strategic interests with international safety objectives. Equitable benefit distribution ensures that developing nations and marginalized communities share in the advantages of AGI advancements, reducing global inequalities and fostering trust.

Transparency in decision-making processes further deters competitive AGI races. Public reporting of research activities and resource allocations builds trust among stakeholders, ensuring accountability and preventing clandestine competition. This openness, combined with IAIGO's inclusive governance model, promotes a collaborative environment that prioritizes shared progress over unilateral gains.

6.3 Transparent and Inclusive Decision-Making

Transparent and inclusive decision-making is foundational to IAIGO's governance framework. Transparency ensures accountability and trust, while inclusivity guarantees that all stakeholders—particularly those historically underrepresented—have a voice in shaping AGI policies. Together, these principles safeguard against disenfranchisement, promote cooperation, and enhance the legitimacy of IAIGO's governance structure.

To address the risk of AGI systems making opaque decisions with high societal impact—such as in criminal justice, where AGI could recommend sentencing based on biased datasets—IAIGO mandates comprehensive explainability protocols. Developers must provide transparency reports detailing the decision-making processes and potential biases in their systems. Additionally, an independent oversight committee within IAIGO will review high-stakes AGI applications to ensure adherence to ethical and fairness standards.

IAIGO's governance model is designed to ensure meaningful participation from all member states, private sector actors, and civil society organizations. The General Assembly, as IAIGO's primary deliberative body, includes representatives from a diverse range of stakeholders. Its weighted voting system balances sovereign equality with technological capacity, ensuring that all voices are heard without compromising operational efficiency.

The Executive Council provides operational leadership, enforcing IAIGO’s policies and addressing emergent risks. Its composition—major AI nations, private sector leaders, and rotating representatives from other regions—ensures a balance of influence and diversity. Decisions are made using qualified majority voting, avoiding the gridlock often associated with veto-based systems.

IAIGO’s Scientific Advisory Board and Ethics and Equity Commission further enhance decision-making processes. The Scientific Advisory Board provides expert guidance on technical matters, ensuring that policies are informed by the latest research and best practices. The Ethics and Equity Commission addresses ethical considerations, advocating for fairness, justice, and human rights in all AGI-related activities.

IAIGO also establishes regional research hubs and capacity-building initiatives to empower developing nations and underrepresented communities. Acknowledging the limitations of historical governance models that often marginalized smaller nations or lacked adaptability, IAIGO incorporates a decentralized governance structure with polycentric oversight. This ensures that decision-making adapts to regional contexts while maintaining alignment with global safety standards. Furthermore, IAIGO's inclusion of diverse voices—such as ethicists specializing in digital ethics and representatives from indigenous communities—addresses ethical considerations unique to AGI, including its potential to disrupt traditional ways of life. These efforts ensure that the benefits of AGI governance are equitably distributed and that all stakeholders have the resources and support needed to participate fully in decision-making processes.

The risk mitigation strategies employed by IAIGO—addressing malicious actors and rogue development, preventing competitive AGI races, and promoting transparent and inclusive decision-making—create a robust and resilient framework for AGI governance. By integrating advanced monitoring systems, stringent security protocols, economic incentives, and collaborative decision-making processes, IAIGO mitigates the risks associated with AGI development while ensuring that its benefits are distributed equitably. These strategies not only safeguard against potential threats but also lay the foundation for the safe, ethical, and inclusive advancement of AGI technologies, ensuring that they serve as a force for global progress.

IAIGO's commitment to transparency extends to its advanced disclosure and reporting framework. The organization publishes quarterly transparency reports detailing all major decisions, resource allocations, and enforcement actions. A centralized digital platform provides real-time access to non-sensitive operational data, meeting records, and research outputs. Annual accountability reviews assess the organization's adherence to transparency commitments, with results made public. Stakeholders can access a dedicated feedback portal to submit concerns or suggestions, with mandatory response times for formal inquiries. This comprehensive transparency framework ensures that IAIGO's operations remain open to public scrutiny while protecting sensitive technical information through a clearly defined classification system.

7. Incentive Compatibility and Stakeholder Engagement

The success of the International AI Governance Organization (IAIGO) relies heavily on aligning the interests of a wide array of stakeholders, including major powers, private sector entities, academia, civil society organizations, and global citizens. To achieve this, IAIGO employs strategies to ensure that the benefits of participation outweigh potential costs and foster meaningful engagement with all relevant groups. This section synthesizes the best practices for incentivizing and engaging each stakeholder group while addressing their unique priorities and concerns.

7.1 Incentives for Major Powers

The participation of major powers—countries with advanced AI capabilities and global technology leaders—is essential for the legitimacy and efficacy of IAIGO's governance framework. However, these stakeholders may perceive unilateral AGI development as offering greater strategic and economic benefits than compliance with a global governance structure. IAIGO addresses this challenge through carefully designed incentives that align their national and corporate interests with global safety and ethical objectives.

One of IAIGO's primary strategies is the allocation of shared leadership roles. Leading nations and corporations are assigned specific leadership roles, such as chairing the Executive Council subcommittees on security and monitoring, which oversee global compute tracking and rogue actor deterrence. For instance, the United States and China, as dominant AI developers, might co-lead the Global Verification Initiative, ensuring that technical monitoring aligns with cutting-edge capabilities while fostering collaboration between rival powers. These positions not only provide major powers with a platform to shape IAIGO's policies but also create a sense of ownership and accountability for the organization's success. This design takes lessons from institutions such as the International Monetary Fund, where power-sharing structures encourage cooperation among major contributors.

To address AGI-specific risks unique to state-level applications, such as the use of AGI in autonomous military operations, IAIGO incentivizes major powers to adopt common safety protocols by providing exclusive access to IAIGO-certified AGI technologies. These technologies include built-in safeguards that ensure adherence to international humanitarian law and prevent autonomous decision-making in lethal applications. This approach reduces the risk of unregulated AGI development in defense sectors while aligning national security interests with global safety standards.

To further incentivize participation, IAIGO implements a system of economic and technological benefits. For example, major powers gain preferential access to IAIGO's cutting-edge research infrastructure, including computational resources and data repositories. Collaborative research initiatives ensure that leading AI developers remain at the forefront of technological innovation within a controlled and safe framework. Additionally, IAIGO's intellectual property policies allow compliant stakeholders to commercialize safety-certified AGI technologies, preserving their competitive edge in global markets.

To address geopolitical concerns, IAIGO employs a weighted voting system in its General Assembly. This system balances the technological capacity and influence of major powers with the need for equitable representation, ensuring that smaller nations also have a meaningful voice in decision-making. By creating a transparent and cooperative environment, IAIGO reduces the risks of competitive AGI races and fosters trust among major stakeholders.

7.2 Integration of Private Sector and Academia

The private sector and academia are indispensable to AGI development, driving innovation and contributing vital expertise. However, their involvement must be carefully managed to ensure alignment with global safety and ethical standards. IAIGO's governance framework creates pathways for meaningful integration while preserving the autonomy and innovation potential of these stakeholders.

For the private sector, IAIGO establishes public-private partnerships that allow companies to participate in collaborative research initiatives. These partnerships provide access to IAIGO's research infrastructure, including computational resources, standardized safety protocols, and funding opportunities. For example, private sector actors such as NVIDIA and AWS could partner with IAIGO to develop secure compute tracking systems that monitor hardware use globally. Academic institutions like MIT and Tsinghua University might collaborate on alignment research, with IAIGO funding grants to projects that advance safety standards. To incentivize participation, IAIGO will establish an intellectual property (IP) sharing agreement, enabling stakeholders to commercialize compliant AGI innovations while ensuring public access to critical safety technologies. By pooling resources and expertise, IAIGO reduces redundancy and fosters economies of scale in AGI development. Companies that comply with IAIGO's safety standards also benefit from reduced regulatory burdens and streamlined pathways for the commercialization of AGI technologies.

IAIGO addresses the unique needs of academia by supporting international research consortia and funding groundbreaking studies in AI and related disciplines. Academic institutions are integrated into IAIGO's governance structure through representation on the Scientific Advisory Board, ensuring that the latest research informs policy decisions. Furthermore, IAIGO facilitates the dissemination of academic knowledge through open-access data repositories and collaborative platforms, promoting innovation and ethical development across borders.

To incentivize participation, IAIGO offers tailored capacity-building programs that assist smaller firms and academic institutions in meeting its technical and ethical standards. This inclusive approach ensures that all stakeholders, regardless of their size or resources, can contribute to and benefit from IAIGO's governance framework.

7.3 Role of Civil Society and Global Citizen Engagement

Civil society organizations (CSOs) and global citizens play a vital role in ensuring the legitimacy, accountability, and inclusivity of IAIGO's operations. Their engagement is essential for building trust and ensuring that AGI governance reflects diverse societal values and priorities.

IAIGO actively incorporates civil society into its governance framework by reserving seats for CSO representatives in the General Assembly and the Ethics and Equity Commission. These representatives advocate for marginalized communities, contribute ethical insights, and help ensure that IAIGO's policies address broad societal concerns. Public forums and consultative bodies provide additional avenues for civil society engagement, fostering dialogue and collaboration between diverse stakeholders.

To empower global citizens, IAIGO prioritizes transparency and accessibility. Regular public briefings, open consultations, and an online platform for sharing policy documents and research findings ensure that citizens can monitor IAIGO's activities and contribute to its development. IAIGO also establishes regional outreach centers to engage local communities and address their specific concerns.

Educational campaigns form a cornerstone of IAIGO's citizen engagement strategy. These initiatives raise awareness about the risks and opportunities of AGI, equipping citizens with the knowledge and tools to participate in governance discussions. Citizen assemblies and participatory decision-making processes further enhance public involvement, ensuring that IAIGO's policies reflect the diverse perspectives of the global community.

Finally, IAIGO supports grassroots initiatives that promote ethical and inclusive AGI development. By providing financial and technical assistance to these initiatives, IAIGO empowers civil society and global citizens to play an active role in shaping AGI governance. This collaborative approach enhances the legitimacy of IAIGO's framework while fostering a shared sense of responsibility for the safe and equitable development of AGI.

IAIGO's approach to incentive compatibility and stakeholder engagement creates a robust framework for fostering global cooperation in AGI governance. By aligning the strategic interests of major powers, integrating the expertise of the private sector and academia, and engaging civil society and global citizens, IAIGO ensures broad-based support for its mission. These strategies address potential barriers to participation, build trust and accountability, and promote a collaborative environment where all stakeholders contribute to the safe, ethical, and inclusive development of AGI.

8. Implementation Plan

The implementation of the International AI Governance Organization (IAIGO) is structured into three interconnected phases, ensuring that IAIGO develops a solid foundation, engages key stakeholders, establishes the necessary infrastructure, and adapts to evolving technological and geopolitical dynamics. This phased approach ensures IAIGO can effectively govern the safe, equitable, and inclusive development of Artificial General Intelligence (AGI).

8.1 Phase 1: Establishing Legal Frameworks

The first phase focuses on developing the legal and institutional foundation for IAIGO. Spanning approximately 18 to 24 months, this phase lays the groundwork for IAIGO's structure, governance, and global legitimacy. The cornerstone of this phase is the IAIGO Charter Treaty, which will codify the organization's mission, governance structure, and operational mechanisms. Drawing from successful frameworks such as the IAEA Statute, the Rome Statute, and the Montreal Protocol, the treaty will be crafted collaboratively by representatives from national governments, private sector entities, academia, and civil society organizations. It will define IAIGO's mandates, including enforcing a global moratorium on unauthorized AGI development, coordinating international AGI research, and overseeing equitable benefit distribution. The treaty will also establish IAIGO's governing bodies, including the General Assembly, Executive Council, Scientific Advisory Board, and Ethics and Equity Commission, while providing mechanisms for monitoring compliance, enforcing penalties, and resolving disputes. The IAIGO Charter Treaty will define roles for stakeholders, ensuring clarity in mandates and responsibilities. For example, member states will commit to enforcing the global moratorium on unauthorized AGI development within their jurisdictions, while private sector actors will adhere to standardized safety guidelines and contribute expertise to collaborative research programs. Civil society organizations will monitor adherence to equity and ethical standards, while academia will lead public-facing initiatives to raise awareness about AGI governance principles. A diplomatic conference will finalize the treaty, and ratification by major AI powers, combined with widespread international support, will ensure its legitimacy and enforceability.

During this phase, regulatory standards will be developed to govern AGI development, addressing technical, ethical, and safety dimensions. These standards will include compute tracking protocols, facility safety requirements, and red team assessments to mitigate risks. Simultaneously, robust monitoring and verification mechanisms will be established, including global compute tracking systems, facility inspections, and independent audits. To align AGI development with global values, comprehensive legal and ethical guidelines will be created through a consultative process involving diverse voices, ensuring accountability, fairness, and societal impact. To support IAIGO’s establishment, an international funding mechanism, modeled on the Green Climate Fund, could secure contributions from participating nations, private entities, and international donors.

More concretely, a crucial aspect of establishing IAIGO is securing sustainable and diversified funding. Member states will contribute funding proportionate to their GDP, technological capacity, and the specific AGI-related risks they face, creating a fair and scalable foundation for resource mobilization. To complement these contributions, IAIGO will engage the private sector through a Corporate Partnership Fund, enabling private entities to provide financial support in exchange for opportunities to participate in collaborative research initiatives and access early-stage governance frameworks. Partnerships with international organizations, such as the World Bank and philanthropic foundations, will provide additional resources targeted toward specific projects, including capacity-building programs and technology transfer initiatives that benefit developing nations. Furthermore, IAIGO will generate revenue by monetizing its proprietary governance tools, such as compute tracking systems, through licensing agreements with non-member entities. These combined strategies will ensure that IAIGO has the financial stability and resources needed to achieve its objectives and adapt to evolving global challenges.

IAIGO's initial operational budget is structured across four primary categories. Core Operations covers headquarters staffing, basic infrastructure, and administrative costs. The Technology & Research program funds global compute tracking systems, security protocols, and collaborative research initiatives. Capacity Building & Equity Programs support technology transfer, training, and assistance to developing nations. Emergency Response & Contingency Funds ensure rapid response capabilities for emerging risks. Thisl budget will be funded through a combination of mandatory member state contributions (60%), private sector partnerships (25%), and international development institutions (15%). Member state contributions are calculated using a formula incorporating GDP, technological capacity, and risk exposure. A dedicated endowment fund, seeded with in initial contributions, will provide long-term financial stability and independence.

8.2 Phase 2: Stakeholder Recruitment and Infrastructure Development

The second phase, spanning 24 to 36 months, focuses on building IAIGO's capacity and securing stakeholder engagement through practical implementation and rigorous evaluation. This phase combines infrastructure development with pilot programs designed to test and validate IAIGO's governance mechanisms.

Regional research hubs will be established to foster collaboration, alongside secure data centers to support compute tracking and technology transfer programs. These facilities will serve as centers for testing governance mechanisms and conducting pilot programs. Capacity-building initiatives will provide training, financial assistance, and access to resources for underrepresented communities, ensuring they can participate fully in IAIGO's governance framework.

Security protocols will be established to protect IAIGO's infrastructure and data, including advanced cybersecurity frameworks, physical safeguards, and controlled access procedures. These measures will mitigate risks of breaches and ensure the integrity of IAIGO's operations.

To validate IAIGO's governance framework and ensure scalability, four key pilot programs will be implemented and rigorously evaluated. The Global Compute Tracking Pilot will partner with major cloud providers including AWS, Google Cloud, and Microsoft Azure to establish a prototype tracking system in North America and the European Union. This pilot will evaluate how effectively compute tracking can preempt unauthorized AGI development and scale for global implementation, focusing on anomaly detection under real-world conditions.

The Ethical AGI Research Collaboration will launch as a joint initiative between three leading AI research labs focusing on AGI alignment methods. This program will evaluate research milestone achievement, collaboration effectiveness, and safety protocol compliance, while assessing resource requirements and knowledge transfer effectiveness across different cultural contexts. This program uniquely emphasizes cross-cultural research collaboration to identify alignment challenges across diverse regulatory and technological contexts, ensuring robustness in ethical AGI standards.

The Regional Benefit-Sharing Program will implement technology transfer and capacity-building initiatives in Sub-Saharan Africa, measuring local capacity improvement, technology adoption rates, and healthcare outcome improvements. This pilot will carefully assess resource requirements for global expansion, cultural adaptation needs, and infrastructure prerequisites for scaling.

An Emergency Response Simulation program will conduct simulated AGI misuse scenarios involving multiple stakeholders to evaluate response time, coordination effectiveness, and protocol clarity. The simulation will help determine resource requirements for global implementation and assess cross-border coordination challenges.

Each pilot program will undergo systematic evaluation through independent third-party audits measuring outcomes against predefined success metrics. Regular stakeholder feedback will be collected through surveys and structured interviews, complemented by statistical analysis of program performance data and detailed cost-benefit analysis for global scaling.

Each pilot program includes specific success metrics to evaluate effectiveness and scalability. For the Global Compute Tracking Pilot, these include achieving consistent system uptime and maintaining false positive rates below a threshold. The Ethical AGI Research Collaboration will be measured by concrete research outputs, including at least three peer-reviewed publications on critical safety challenges. The Regional Benefit-Sharing Program must demonstrate measurable impact through an improvement in healthcare outcomes within pilot regions. Emergency Response Simulation effectiveness will be evaluated against response time targets of under a certain target time for critical incidents. Cross-cutting metrics include stakeholder satisfaction rates exceeding a target level in independent surveys and operational efficiency measures against predetermined cost thresholds. These quantitative benchmarks will guide decisions about global scaling and implementation.

IAIGO implements a comprehensive technical validation system for AGI development and deployment. All systems undergo a four-stage validation process before receiving operational approval.

Stage 1 (Architectural Review) examines system design and safety mechanisms. Independent experts evaluate architectural choices, focusing on alignment mechanisms, containment strategies, and failure modes. Systems must demonstrate robust safety properties through formal verification where possible, and through extensive simulation testing otherwise.

Stage 2 (Behavioral Testing) subjects systems to standardized performance evaluations across multiple domains. Testing protocols include adversarial challenges, edge case scenarios, and long-term stability assessments. Systems must achieve benchmark performance in areas including goal preservation, value alignment, and response to intervention commands.

Stage 3 (Integration Testing) evaluates system behavior within controlled real-world environments. This includes interaction with other AI systems, response to unexpected inputs, and performance under resource constraints. Extended observation periods verify consistent behavior and reliable safety mechanisms.

Stage 4 (Deployment Validation) assesses operational readiness through limited real-world deployment. Systems operate under enhanced monitoring, with graduated expansion of capabilities based on performance metrics. Final approval requires demonstration of reliable safety mechanisms, transparent decision-making processes, and consistent alignment with human values.

The validation framework employs quantitative metrics for each stage, requiring systems to meet or exceed predetermined thresholds. These include:

- Safety Protocol Compliance: 99.99% adherence rate
- Decision Transparency: 95% explainability score
- Intervention Response: 100% successful emergency halts
- Value Alignment: 99.9% consistency with defined parameters
- Performance Stability: <0.01% unexpected behaviors

IAIGO maintains a public database of validation results, enabling continuous improvement of standards while ensuring transparency. The framework undergoes annual updates to incorporate new research findings and address emerging challenges.

A global diplomatic campaign will secure commitments from major AI powers, private sector leaders, academic institutions, and civil society organizations. For resistant nations, IAIGO will deploy targeted diplomatic efforts highlighting strategic advantages of cooperation through bilateral negotiations addressing specific concerns about sovereignty and influence. Regional workshops will demonstrate tangible benefits of participation, while flexible participation options will be tailored to different stakeholder needs.

The "Voices for AGI" initiative will launch during this phase, including community listening tours in urban and rural areas globally to gather input from diverse populations. Findings will directly inform IAIGO's policies and implementation strategies.

Several specific failure modes could threaten IAIGO's effectiveness. First, a "parallel development scenario" could emerge where a coalition of non-participating nations creates a competing AGI governance framework with weaker safety standards. This occurred historically with nuclear technology when nations developed parallel programs outside the IAEA framework. IAIGO will mitigate this risk by implementing a differential access protocol that provides participating nations preferential access to advanced AGI research and compute resources, making participation more attractive than parallel development.

Second, a "regulatory capture scenario" could occur if powerful tech corporations covertly influence IAIGO's policies to favor their interests over global safety. This happened with financial regulation before the 2008 crisis. IAIGO will prevent this through mandatory rotation of oversight personnel, strict conflict of interest policies, and an independent watchdog committee with civil society representation.

Third, a "technological leapfrog scenario" might arise where a breakthrough in AGI development (such as a radical new architecture requiring minimal compute resources) renders existing monitoring mechanisms ineffective. To address this, IAIGO will maintain a rapid response task force of technical experts authorized to implement emergency protocols within 48 hours of detecting novel development approaches. These protocols include temporary global restrictions on specific hardware or software components until new monitoring mechanisms are established.

8.3 Phase 3: Launch and Continuous Adaptation

The final phase, beginning three to four years after treaty adoption, marks IAIGO's full operational launch and establishes mechanisms for continuous improvement based on pilot program lessons and stakeholder feedback.

IAIGO's inauguration will be celebrated through an international conference, bringing together member states, private sector leaders, and civil society organizations. The launch will showcase successful pilot program outcomes and demonstrate IAIGO's readiness for full-scale operations.

During this phase, IAIGO will operationalize its core mandates by enforcing the global moratorium on unauthorized AGI development, informed by pilot program experiences. The organization will coordinate collaborative research aligned with global priorities and implement benefit-sharing mechanisms, scaled based on pilot program evaluations.

Based on pilot program evaluations, IAIGO will implement global scaling of successful initiatives through systematic expansion of compute tracking systems to all member states. Successful research collaboration models will be replicated across regions, while benefit-sharing programs will be extended to all eligible regions. Refined emergency response protocols will be implemented globally.

IAIGO will establish a comprehensive framework for continuous improvement through quarterly review cycles incorporating stakeholder feedback and performance metrics. Annual independent evaluations of all major programs will be conducted, alongside biennial strategic reviews of governance framework effectiveness. Technical standards and protocols will be regularly updated based on emerging technologies.

The framework includes mechanisms for rapid response to emerging risks or opportunities, integration of new stakeholder needs and perspectives, and adaptation of governance mechanisms based on operational experience. Success metrics and evaluation criteria will be regularly updated to reflect evolving challenges and opportunities.

Building on pilot program experiences, IAIGO will implement enhanced monitoring systems for early risk detection and regular stress testing of governance mechanisms. Continuous assessment of stakeholder satisfaction and engagement will be conducted, along with periodic reviews of benefit distribution effectiveness.

To sustain global cooperation, IAIGO will maintain transparent reporting on all major initiatives and continue regular stakeholder consultations. Success stories and lessons learned will be shared widely, while accessible updates on progress toward strategic goals will be provided regularly.

This phase establishes IAIGO as the definitive global authority on AGI governance while maintaining flexibility to adapt to emerging challenges and opportunities. The continuous improvement framework ensures that lessons from pilot programs and early operations inform ongoing refinements to IAIGO's governance mechanisms.

IAIGO’s adaptability will be driven by a Real-Time Adaptation Mechanism (RTAM), which integrates stakeholder feedback, performance metrics, and AI-driven analytics to update policies in response to emerging risks and opportunities. This mechanism will continuously monitor technological advancements, stakeholder feedback, and geopolitical dynamics, allowing IAIGO to update policies and protocols proactively. Quarterly review cycles, supported by AI-driven analytics, will identify emerging risks and opportunities, ensuring that IAIGO remains agile and responsive in the fast-evolving AGI landscape.

While this implementation plan provides a structured approach to establishing IAIGO, several critical challenges must be addressed to ensure its success. The following section examines these challenges in detail and outlines specific strategies for overcoming potential obstacles to IAIGO's mission.

Success metrics for IAIGO's operations span five key dimensions, each with specific measurable targets. In terms of technical effectiveness, IAIGO aims to achieve a 100% detection rate for unauthorized AGI development activities while maintaining a false positive rate below 0.1% in compute monitoring systems. The organization will ensure 99.9% uptime for critical infrastructure and achieve a mean time to detection of under one hour for serious violations.

For safety outcomes, IAIGO commits to achieving zero catastrophic incidents from approved AGI systems and maintaining a 100% containment rate for identified risks. The organization will establish a 95% prevention rate for attempted safety protocol violations and maintain an average response time under 30 minutes for critical incidents.

Stakeholder engagement metrics focus on achieving a 90% participation rate from eligible nations and an 80% compliance rate with IAIGO protocols. The organization aims to maintain a 75% stakeholder satisfaction rating and achieve a 70% public trust rating in annual surveys.

In the realm of equity and access, IAIGO will ensure technology transfer programs reach 90% of eligible nations while reducing the global AI capacity gap by 50% within five years. The organization commits to delivering 80% of benefits to traditionally underserved regions and achieving equal representation across all governance bodies.

Organizational efficiency targets include maintaining operating costs within 5% of approved budget and achieving decision-making times under 48 hours for urgent matters. IAIGO will strive for a 95% transparency rating from independent auditors and maintain a staff retention rate above 85%.

9. Challenges and Considerations

9.1 Geopolitical Tensions and Non-Participation Risks

One of the most significant challenges to the establishment and effective functioning of the International AI Governance Organization (IAIGO) arises from geopolitical tensions and the risk of non-participation by key actors. Advanced AI nations, particularly those with strong technological capabilities such as the United States, China, and members of the European Union, often view AGI as a strategic asset that could confer economic and military dominance. This perception creates a competitive landscape where unilateral advancements may be prioritized over multilateral cooperation.

Historical examples, such as the nuclear arms race during the Cold War, underscore the difficulty of persuading powerful nations to relinquish or limit their strategic advantages for the sake of global security. In the case of IAIGO, resistance to participation may stem from fears of losing a competitive edge, concerns over the equitable distribution of AGI's benefits, or distrust in the governance framework itself. The absence of key players from IAIGO would not only undermine the legitimacy of the organization but could also lead to the development of unregulated AGI systems outside the governance framework, exacerbating global risks.

Non-participation by key stakeholders, particularly major powers, poses a significant challenge to IAIGO's success. To address this, IAIGO must implement targeted engagement strategies that align with the geopolitical interests of resistant nations or entities. For instance, offering special leadership roles or weighted influence in decision-making bodies can incentivize participation by advanced AI powers while maintaining equitable representation for smaller nations. Additionally, IAIGO can use phased entry mechanisms that allow hesitant nations to observe and join later without significant penalties, thus reducing initial resistance to commitment.

To address these concerns, IAIGO must carefully balance the interests of major powers with those of smaller nations and marginalized communities. This involves creating incentive structures that appeal to all stakeholders. For instance, IAIGO could guarantee equitable access to AGI benefits, such as technological advancements in healthcare and climate change mitigation, while offering leading AI nations influential roles in its governance structure. Transparency, inclusivity, and trust-building measures are essential to addressing geopolitical resistance. IAIGO should establish independent oversight mechanisms and public reporting systems that demonstrate fairness in decision-making and benefit-sharing. Furthermore, targeted confidence-building measures, such as bilateral agreements with resistant nations or capacity-building initiatives in regions with limited AI expertise, can reduce distrust and create pathways for collaboration.

For nations or entities resistant to IAIGO's framework due to perceived sovereignty concerns, IAIGO can offer tailored participation models that respect national autonomy while aligning with global safety standards. These could include opt-in clauses for specific mandates, such as access to computational resources or security protocols, without requiring full treaty ratification. Additionally, IAIGO can leverage economic incentives, such as funding for AI research infrastructure, to make participation more attractive than unilateral development. Collaborative forums where nations can voice concerns and influence policy evolution can further mitigate resistance.

While geopolitical challenges present significant external risks to IAIGO's success, equally important are the technical challenges inherent in AGI safety and monitoring.

9.2 Technological Challenges in AGI Safety

The technical complexity of Artificial General Intelligence (AGI) poses a second major challenge to IAIGO's objectives. AGI systems differ fundamentally from narrow AI in their ability to perform tasks across a broad range of cognitive domains at or above human levels. This transformative capability introduces new risks, such as misalignment between AGI objectives and human values, unintended emergent behaviors, and the potential for rapid self-improvement or recursive learning that could surpass human oversight.

The dynamic nature of AGI development presents unique monitoring challenges. Unlike nuclear facilities, which have relatively stable physical infrastructure, AGI development environments can rapidly shift between cloud providers, utilize distributed computing networks, or leverage novel architectures that may evade traditional monitoring approaches. IAIGO must therefore develop adaptive monitoring systems capable of tracking not just raw compute usage, but also novel efficiency improvements, algorithmic innovations, and emergent capabilities that could enable AGI development with fewer detectable resources.

Ensuring AGI safety requires the development of robust alignment techniques that can reliably align AGI systems with human values and goals. For example, without sufficient safeguards, an AGI optimizing global logistics might deprioritize smaller, less profitable regions, exacerbating inequalities. Similarly, recursive self-improvement capabilities could enable AGI systems to enhance their own architectures, potentially escaping predefined constraints or oversight mechanisms. Mitigation strategies include implementing 'safe stopping points' within AGI architectures to halt development at predefined safety thresholds, requiring external human authorization for further advancements. IAIGO will also develop international research consortia focused on creating universally applicable safety protocols for recursive improvement. Researchers like Stuart Russell and Nick Bostrom have emphasized the difficulty of specifying value systems in computational terms and ensuring that AGI systems adhere to these values in diverse and unforeseen scenarios (Bostrom, 2014; Russell, 2019). Furthermore, the challenge of interpretability in machine learning systems complicates efforts to predict and control AGI behavior.

IAIGO must address these challenges by integrating cutting-edge research into its governance framework. This includes mandating rigorous testing and validation protocols for AGI systems, establishing international research hubs focused on alignment and safety, and facilitating collaboration among leading AI experts. IAIGO's Scientific Advisory Board will play a critical role in setting technical standards and ensuring that AGI research adheres to safety guidelines. However, technological uncertainty and the pace of innovation demand that IAIGO maintain flexibility and adaptability in its policies to respond to emerging risks and breakthroughs.

Beyond the technical challenges of AGI safety lies perhaps an even more complex dimension: the ethical and legal considerations that will shape how AGI is integrated into human society.

9.3 Ethical and Legal Considerations

The development and deployment of AGI raise profound ethical and legal questions that challenge traditional governance frameworks. One of the primary ethical concerns is the potential for AGI to exacerbate global inequalities. If access to AGI's capabilities is concentrated among a few nations or corporations, the resulting disparities in wealth, power, and influence could destabilize societies and marginalize vulnerable populations. This scenario underscores the need for equitable benefit-sharing mechanisms within IAIGO's governance framework.

Another ethical issue is the potential misuse of AGI for purposes such as mass surveillance, autonomous weapons, and disinformation campaigns. The ethical implications of AGI's deployment extend beyond its intended uses to include its potential to infringe on human rights, privacy, and autonomy. Legal frameworks must address these concerns by establishing clear accountability for AGI developers, users, and policymakers. The moral status of AGI introduces additional complexities to legal and ethical considerations. IAIGO must explore whether AGI systems that exhibit attributes such as autonomy, intentionality, or the ability to experience harm should be granted moral standing or rights. Legal frameworks must anticipate scenarios where AGI entities could challenge traditional definitions of agency and accountability, ensuring that any rights extended to AGI do not undermine human welfare or societal stability. Equally important is addressing global inequalities through enforceable mandates for benefit-sharing, ensuring that AGI-driven wealth and advancements are equitably distributed to uplift disadvantaged regions.

The possibility of granting AGI systems moral status adds an additional layer of complexity to the ethical debate. As AGI systems approach or exceed human cognitive capabilities, questions about their rights and responsibilities become increasingly relevant. Philosophers and ethicists have begun to explore whether AGI systems could possess moral agency and whether their treatment should be subject to ethical guidelines Niebuhr, 1932; Rawls, 1971). These considerations challenge existing legal and moral frameworks, which are primarily designed for human actors.

IAIGO must define roles for stakeholders in upholding ethical and legal standards. For example, private corporations will be required to disclose their AGI training datasets to ensure compliance with transparency guidelines, while civil society organizations will independently audit the societal impact of AGI applications. Member states will implement national legislation to align domestic AGI policies with IAIGO's global framework, while academia will provide ongoing research into the ethical implications of emergent AGI behaviors. The Ethics and Equity Commission will play a central role in crafting guidelines that reflect global values such as fairness, accountability, and human rights. Public consultations and interdisciplinary research will be essential to ensuring that these guidelines are inclusive and robust. Additionally, IAIGO must establish mechanisms for enforcing ethical standards, such as independent audits and transparent reporting, to hold stakeholders accountable for their actions.

The challenges of geopolitical tensions, technological complexity, and ethical and legal considerations highlight the multifaceted nature of AGI governance. IAIGO must navigate these challenges with a combination of innovative policy design, technical expertise, and global collaboration. By addressing the risks of non-participation, investing in AGI safety research, and developing comprehensive ethical and legal frameworks, IAIGO can create a governance structure that safeguards humanity while unlocking AGI's transformative potential. The stakes are high, but with careful planning and inclusive dialogue, IAIGO has the opportunity to lead the world toward a safe and equitable AGI future.

10. Conclusion and Call to Action

The development of Artificial General Intelligence (AGI) represents a turning point in human history—one filled with immense promise but also profound risks. This report has laid out a comprehensive blueprint for the International AI Governance Organization (IAIGO), a treaty-based institution designed to ensure that AGI serves as a force for global good rather than an existential threat. Through its inclusive structure, robust operational mechanisms, and incentive-compatible strategies, IAIGO provides the governance framework necessary to navigate the complexities of AGI development, deployment, and benefit-sharing.

The stakes could not be higher. Unregulated AGI development risks destabilizing global security, deepening inequalities, and creating unprecedented ethical dilemmas. The potential for catastrophic misuse or misalignment between AGI systems and human values demands immediate and decisive international action. Historical precedents, such as the proliferation of nuclear weapons or the consequences of uncoordinated climate action, illustrate the dire consequences of inaction in the face of transformative technologies.

At the same time, the opportunities presented by AGI are unparalleled. If governed effectively, AGI can accelerate progress in addressing humanity's greatest challenges, including climate change, global health disparities, and economic inequality. Its capacity to revolutionize industries, optimize resource management, and enable new scientific breakthroughs has the potential to uplift billions of lives. However, realizing these benefits equitably and safely requires cooperation, transparency, and a shared global vision.

A Global Call to Action

Achieving stable AGI governance requires the collective effort of governments, corporations, civil society, and the global public. First and foremost, national governments must recognize the urgency of establishing a governance framework before AGI capabilities surpass our ability to control them. Policymakers should prioritize the adoption and ratification of the IAIGO Charter Treaty, which provides the legal and institutional foundation for coordinated AGI governance. Failure to act promptly risks ceding control to unregulated actors, increasing the likelihood of unsafe or inequitable AGI deployment.

The private sector also has a pivotal role to play. As the primary drivers of AGI innovation, corporations and research institutions must align their development efforts with global safety and ethical standards. IAIGO offers an avenue for these stakeholders to participate in shaping AGI governance while preserving innovation and commercialization opportunities. By engaging with IAIGO's public-private partnerships and adhering to its safety protocols, the private sector can help ensure that AGI development is both responsible and inclusive.

Civil society organizations and global citizens must continue to advocate for transparent, equitable, and ethical AGI governance. Public engagement will be critical to holding governments and corporations accountable for their commitments to safety and fairness. IAIGO will establish a Global Citizens' Engagement Program (GCEP) designed to facilitate grassroots involvement in AGI governance. This program will include annual 'Citizen Dialogues on AGI,' hosted in partnership with regional outreach centers, to solicit input on key governance policies. A multilingual online platform will provide tools for individuals to contribute ideas, vote on policy priorities, and access educational resources on AGI risks and opportunities. By integrating citizen assemblies into the decision-making process, IAIGO ensures that grassroots voices influence global AGI governance. IAIGO's inclusive governance model ensures that the voices of marginalized communities and underrepresented stakeholders are heard, but sustained advocacy is essential to maintaining this inclusivity over time.

To further support public engagement, IAIGO will establish a 'Public Observatory on AGI Governance' - an accessible online portal where citizens can track organizational decisions and provide direct feedback. This observatory will feature interactive dashboards summarizing AGI-related developments, enable public input on priority areas through regular polls, and provide live-streamed updates from IAIGO leadership. Biannual 'State of AGI Governance Reports' will ensure transparency by comprehensively documenting IAIGO's activities and demonstrating how public contributions shape policy decisions.

Finally, international organizations, including the United Nations, must act as conveners and facilitators of global cooperation. The lessons of past successes, such as the Montreal Protocol and the International Atomic Energy Agency, demonstrate the importance of multilateral institutions in addressing global challenges. IAIGO represents an opportunity to build on these precedents, adapting them to the unique challenges and opportunities of AGI.

Charting a Path Forward

The implementation of IAIGO will not be without challenges. Geopolitical tensions, technological uncertainties, and ethical complexities will test the resilience of the organization’s governance framework. However, the phased implementation plan outlined in this report provides a clear roadmap for navigating these challenges. By establishing legal foundations, building stakeholder coalitions, and continuously adapting to emerging risks and opportunities, IAIGO can remain effective in the face of rapid technological and geopolitical change.

The time for action is now. As AGI development accelerates, the window for establishing effective governance frameworks is rapidly closing. Delaying action risks entrenching fragmented and uncoordinated approaches to AGI governance, heightening the risks of misuse and inequity. Conversely, decisive action to establish IAIGO can set a global precedent for the responsible governance of transformative technologies, ensuring that AGI serves as a force for collective progress.

A Shared Responsibility

The safe and equitable governance of AGI is not the responsibility of any one nation, corporation, or institution—it is a shared global responsibility. IAIGO provides the tools and structures necessary to realize this responsibility, but its success depends on the collective commitment of all stakeholders. By uniting around a common vision for AGI governance, humanity can ensure that this transformative technology benefits all while safeguarding against its risks.

Let this report serve as a catalyst for dialogue, collaboration, and action. The stakes are too high, and the opportunities too great, to allow inaction or fragmentation to prevail. As the ethical implications of AGI development continue to evolve, IAIGO must lead the way in addressing complex issues such as AGI’s moral status and the equitable distribution of its benefits. These challenges are not merely theoretical; they represent critical inflection points in humanity’s relationship with technology. By embedding these considerations into its governance framework, IAIGO ensures that AGI development reflects the shared values and aspirations of a truly global society

The future of AGI—and of humanity—rests in our collective hands. Together, through IAIGO, we can chart a path toward a safer, more equitable, and prosperous future for all.

11. References

Abbott, K. W., & Snidal, D. (2021). Strengthening international regulation through transnational new governance: Overcoming the orchestration deficit. In *The spectrum of international institutions* (pp. 95-139). Routledge.

Abbott, K. W., & Snidal, D. (2021). The governance triangle: Regulatory standards institutions and the shadow of the state. In *The spectrum of international institutions* (pp. 52-91). Routledge.

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

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint, *arXiv:1802.07228*. Retrieved from https://arxiv.org/abs/1802.07228

Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company.

Convention for the Establishment of a European Organization for Nuclear Research (CERN). (1954). Retrieved from https://council.web.cern.ch/en/content/convention-establishment-european-organization-nuclear-research

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

International Atomic Energy Agency (IAEA). (1957). *Statute of the International Atomic Energy Agency*. https://www.iaea.org/about/statute

International Atomic Energy Agency (IAEA). (2007). *IAEA Safeguards: Stemming the Spread of Nuclear Weapons*. Vienna: International Atomic Energy Agency.

Knock, T. J. (1995). *To End All Wars: Woodrow Wilson and the Quest for a New World Order*. Princeton University Press.

Luck, E. C. (2006). *UN Security Council: Practice and Promise*. Routledge.

McKinsey Global Institute. (2018). *AI, Automation, and the Future of Work*. McKinsey & Company. https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for

Niebuhr, R. (1932). *Moral Man and Immoral Society: A Study in Ethics and Politics*. Charles Scribner's Sons.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Ostrom, E. (2010a). *Beyond markets and states: Polycentric governance of complex economic systems*. American Economic Review, 100(3), 641-672. Retrieved from http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.3.641

Ostrom, E. (2010b). *Polycentric systems for coping with collective action and global environmental change*. Global Environmental Change, 20(4), 550-557. Retrieved from https://doi.org/10.1016/j.gloenvcha.2010.07.004

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Rhodes, R. (2012). *The Making of the Atomic Bomb*. Simon & Schuster.

Rolnick, D., Donti, P. L., Kaack, L. H., et al. (2019). Tackling climate change with machine learning. *arXiv:1906.05433*. Retrieved from https://arxiv.org/abs/1906.05433

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

Slaughter, A. M. (2004). *A New World Order*. Princeton University Press.

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books.

United Nations. (1945). *Charter of the United Nations*. Retrieved from https://www.un.org/en/about-us/un-charter.

United Nations. (1987). *Montreal Protocol on Substances that Deplete the Ozone Layer*. United Nations Environment Programme. Retrieved from https://ozone.unep.org/treaties/montreal-protocol

United Nations. (2000). *The Biological Weapons Convention Protocol*. United Nations Office for Disarmament Affairs.

United Nations. (2015a). *The Paris Agreement*. United Nations Framework Convention on Climate Change. Retrieved from https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.

United Nations. (2015b). *Transforming Our World: The 2030 Agenda for Sustainable Development*. United Nations. Retrieved from https://sdgs.un.org/2030agenda.

Victor, D. G. (2011). *Global Warming Gridlock: Creating More Effective Strategies for Protecting the Planet*. Cambridge University Press.

Waltz, K. N. (1979). *Theory of International Politics*. McGraw-Hill.

World Economic Forum. (2020). *The Global Risks Report 2020*. World Economic Forum. Retrieved from https://www.weforum.org/reports/the-global-risks-report-2020

World Trade Organization. (1995). *Agreement Establishing the WTO*. Retrieved from https://www.wto.org/english/docs_e/legal_e/04-wto_e.htm

Young, O. R. (2017). Conceptualization: Goal setting as a strategy for earth system governance. In N. Kanie & F. Biermann (Eds.), *Governing through goals: Sustainable development goals as governance innovation* (pp. 31-51). The MIT Press.

12. Appendices

12.1 Model Treaty Text

The International AI Governance Organization (IAIGO) Charter Treaty establishes the foundation for its legal framework, mandates, and operational structure. It begins with a preamble that articulates the treaty's purpose, emphasizing humanity's shared responsibility to govern Artificial General Intelligence (AGI) in a way that ensures it becomes a force for good rather than a source of harm. This section draws on lessons from successful historical precedents, including the International Atomic Energy Agency (IAEA) and the Montreal Protocol, to underline the necessity of global cooperation.

The treaty defines IAIGO's three primary mandates: implementing and enforcing a global moratorium on unauthorized AGI development, facilitating international collaboration on safe and ethical AGI research, and ensuring the equitable allocation of AGI-derived benefits across all nations, with special attention to the needs of developing countries. These mandates are framed as essential steps in mitigating existential risks while maximizing the transformative potential of AGI.

The governance structure outlined in the treaty specifies the composition and functions of IAIGO's core bodies: the General Assembly, Executive Council, Scientific Advisory Board, and Ethics and Equity Commission. The General Assembly is responsible for high-level deliberations and strategic decisions, ensuring inclusive global representation. The Executive Council oversees operational activities and policy enforcement, while the Scientific Advisory Board provides technical expertise to guide research and development. The Ethics and Equity Commission ensures that IAIGO's actions adhere to ethical principles and promote fairness and justice. Procedural rules, including a weighted voting mechanism that balances the influence of major powers with the need for inclusivity, ensure both efficiency and legitimacy in decision-making.

The treaty details robust monitoring and enforcement mechanisms. These include global compute tracking systems, on-site facility inspections, and regular independent audits to ensure compliance with IAIGO's regulations. Specific protocols outline how violations will be addressed, incorporating penalties, remediation requirements, and, if necessary, mechanisms for resolving disputes through arbitration or international courts.

Ethical guidelines are central to the treaty, mandating adherence to global standards of fairness, accountability, transparency, and respect for human rights. These principles are designed to guide AGI research and deployment in a way that aligns with humanity's collective values while addressing issues such as algorithmic bias, data misuse, and the risks associated with emergent behaviors in advanced systems.

Funding and resources are another critical element of the treaty. It establishes a financial mechanism to ensure the sustainable operation of IAIGO, with contributions from member states, private sector stakeholders, and international donors. This mechanism also supports capacity-building initiatives for developing nations, enabling them to participate meaningfully in AGI governance and benefit from its advancements.

The treaty outlines a clear process for ratification and accession, designed to encourage widespread participation from nations, corporations, and civil society organizations. It emphasizes the importance of early adoption by major AI powers to ensure legitimacy and effectiveness while creating pathways for other stakeholders to join over time.

Finally, the treaty incorporates mechanisms for amendments and periodic reviews to adapt to the rapidly evolving technological and geopolitical landscape. These provisions ensure that IAIGO remains flexible and responsive to new challenges and opportunities, fostering its long-term relevance and impact. By establishing a robust legal framework and operational foundation, the IAIGO Charter Treaty provides the tools necessary to govern AGI effectively for the benefit of all humanity.

12.2 Timeline for Implementation

The successful establishment and operation of the International AI Governance Organization (IAIGO) require a phased approach to ensure that its foundations are solid, its infrastructure is robust, and its operations are adaptable to evolving challenges. This timeline spans three distinct phases, each designed to address specific goals and build upon the progress of the previous stage.

The first phase focuses on establishing the legal and institutional foundations necessary for IAIGO's creation. This phase, anticipated to last between one to two years, includes finalizing and ratifying the IAIGO Charter Treaty through a global diplomatic conference. Representatives from national governments, private sector leaders, academia, and civil society organizations will collaborate to craft and formalize the treaty, which will serve as IAIGO's guiding framework. Concurrently, founding members—comprising major AI powers, industry leaders, and international organizations—will be recruited to ensure broad-based support and legitimacy. During this phase, initial monitoring and verification systems will also be developed, including protocols for global compute tracking, facility inspections, and compliance audits. This phase will also focus on pre-empting non-participation by key actors through structured dialogues and preliminary agreements that address sovereignty and influence concerns. These efforts will include informal consultations, regional workshops, and bilateral discussions to build trust and address specific geopolitical concerns. These mechanisms will lay the groundwork for enforcing IAIGO's mandate to regulate AGI development and ensure its safety.

The second phase, which spans approximately two to three years, emphasizes building IAIGO's operational capacity and engaging stakeholders across all sectors. The establishment of IAIGO's headquarters and regional research hubs will be prioritized to foster international collaboration and streamline operations. Pilot programs will be launched to test and refine key initiatives such as monitoring protocols, research coordination frameworks, and mechanisms for equitable benefit distribution. To demonstrate feasibility and scalability, IAIGO will implement the following concrete pilot programs:

- Global Compute Tracking Pilot: In collaboration with major cloud providers such as AWS, Google Cloud, and Microsoft Azure, IAIGO will establish a prototype compute tracking system in a defined region (e.g., North America or the European Union). This program will monitor the usage of high-performance computing resources, providing real-time anomaly detection and refining auditing processes. Results will be used to scale the system globally.
- Ethical AGI Research Collaboration Pilot: IAIGO will fund and oversee a joint research initiative between three leading AI research labs, selected through an open application process. The pilot will focus on advancing AGI alignment methods, including

reinforcement learning with human feedback (RLHF). Progress will be monitored through monthly updates and shared in open-access journals to ensure transparency and inclusivity.

- Regional Benefit-Sharing Program: A pilot program in Sub-Saharan Africa will focus on technology transfer and capacity-building. IAIGO will partner with local governments and universities to deploy narrow AI systems in healthcare (e.g., diagnostics for malaria) and evaluate the program’s impact on regional development. Lessons learned will inform the design of global benefit-sharing mechanisms.
- Simulated Emergency Response for AGI Misuse: IAIGO will conduct a simulated scenario involving the hypothetical misuse of AGI in a controlled environment. This simulation will engage stakeholders from cybersecurity firms, government agencies, and civil society to test response protocols, refine interagency collaboration, and identify areas requiring additional safeguards.

These pilot programs will provide valuable insights to optimize IAIGO’s full-scale operations. During this phase, stakeholder engagement efforts will intensify, including the formation of public-private partnerships, academic collaborations, and outreach initiatives targeting civil society. Such efforts will ensure that all relevant actors are represented and have a stake in IAIGO’s mission, fostering trust and cooperation. Public engagement strategies will include launching an 'AGI for Humanity' global campaign to raise awareness about the opportunities and risks of AGI. This campaign will feature accessible explainer videos, local workshops, and interactive events in schools and community centers worldwide. Additionally, IAIGO will host a biennial 'Global AGI Summit,' open to the public, where citizens can directly interact with policymakers, researchers, and private sector leaders to share their perspectives and gain insights into governance efforts.

The final phase, commencing in the fifth year and extending through the seventh, marks the operationalization of IAIGO and its full enforcement of governance protocols. At this stage, the global moratorium on unauthorized AGI development will be implemented rigorously, with comprehensive monitoring and verification systems in place to ensure compliance. IAIGO will also coordinate international AGI research initiatives, pooling resources and expertise to advance safe and ethical technological progress. Benefit-sharing programs will be rolled out, including technology transfers, capacity-building initiatives, and financial assistance to underrepresented regions. Continuous adaptation will be a central feature of this phase, with regular reviews of IAIGO's policies and operations to address emerging technological risks, geopolitical challenges, and ethical considerations. These reviews will be informed by stakeholder feedback, independent assessments, and ongoing research, ensuring IAIGO's governance framework remains effective and relevant.

The phased timeline is designed to build momentum while addressing the complexities of AGI governance in a structured and adaptive manner. By establishing robust legal foundations, developing operational infrastructure, and fostering inclusive global collaboration, IAIGO can position itself as a leader in the responsible and equitable management of AGI. This stepwise approach ensures that IAIGO evolves alongside technological advancements and global needs, enabling it to fulfill its mission of safeguarding humanity while unlocking the transformative potential of AGI.

12.3 Technical Definitions and Thresholds

The governance of Artificial General Intelligence (AGI) necessitates precise technical definitions and enforceable thresholds to ensure clarity, consistency, and effective regulation. These definitions and thresholds form the cornerstone of IAIGO's monitoring, verification, and compliance mechanisms, allowing for uniform interpretation and application across all member states and stakeholders.

AGI is defined as any artificial intelligence system capable of performing intellectual tasks at or above human levels across a wide spectrum of cognitive domains. Such systems must demonstrate the ability to learn, reason, and adapt autonomously to diverse and unpredictable environments. This definition distinguishes AGI from narrow AI, which is specialized for specific tasks, by emphasizing its generalization capabilities and potential to independently address complex, multi-domain challenges.

To regulate the computational resources critical to AGI development, IAIGO establishes thresholds for high-performance compute systems. Computational infrastructures that exceed a defined level of processing power, measured in floating-point operations per second (FLOPS), are subject to mandatory monitoring. The initial threshold is set at 10^18 FLOPS, representing the exaflop scale of computing power. This encompasses systems with the capacity to process AGI-level tasks and ensures that significant compute resources are tracked and regulated. Similarly, any AI training runs that require more than 1,000 petaflop-days of compute intensity must be reported and monitored to prevent unauthorized AGI development.

IAIGO categorizes systems based on development and deployment risk levels to ensure appropriate governance measures are applied. Low-risk systems include narrow AI applications with no potential for emergent general intelligence capabilities. Moderate-risk systems are those with advanced functionalities but limited generalization abilities. High-risk systems are defined as those exhibiting AGI-level capabilities or unexpected emergent behaviors, necessitating the most stringent oversight to mitigate risks.

Monitoring and verification rely on standardized metrics to ensure compliance with IAIGO's safety and ethical standards. Compute utilization is tracked in real-time by registered facilities to detect anomalies or unauthorized activities. The data used for AI training is audited to verify adherence to ethical guidelines, ensuring that datasets do not perpetuate bias or violate privacy standards. Additionally, algorithmic audits are conducted regularly to evaluate source code, testing processes, and overall system integrity, identifying potential risks and ensuring alignment with established protocols.

Ethical and safety standards underpin all aspects of IAIGO's governance framework. Explainability is a core requirement, mandating that AGI systems be designed to allow transparent interpretation of their decision-making processes. This ensures that system behaviors are predictable and understandable, fostering trust and accountability. Alignment with human values and goals is another critical standard, requiring rigorous testing to confirm that AGI systems act in ways that are beneficial and non-harmful to humanity. Bias mitigation efforts are essential to identify and address algorithmic biases, reduce the risk of unfair or discriminatory outcomes, and promote equitable deployment of AGI technologies.

By establishing these definitions and thresholds, IAIGO ensures a coherent and enforceable framework for AGI governance. These standards not only provide the technical foundation for effective oversight but also foster international trust and collaboration, enabling the safe and equitable development of AGI for the benefit of all humanity.

12.4 Glossary of Terms

This glossary provides definitions for technical terms used throughout the report to ensure accessibility for all readers, regardless of their technical background.

Alignment Problem: The challenge of ensuring that AGI systems' goals and behaviors align with human values and intentions.

Artificial General Intelligence (AGI): A type of artificial intelligence capable of performing intellectual tasks at or above human levels across a wide range of cognitive domains.

Compute Tracking: A monitoring system to track and regulate the use of high-performance computing resources, essential for AGI development.

Explainable AI (XAI): Artificial intelligence systems designed to provide understandable and transparent decision-making processes to users.

FLOPS (Floating Point Operations Per Second): A unit of measurement for a computer's processing power, often used to assess performance in tasks requiring large-scale numerical calculations.

Moratorium: A temporary prohibition of specific activities, such as unauthorized AGI development, to ensure safety and regulation.

Polycentric Governance: A governance model with multiple overlapping centers of decision-making that allows for flexibility and adaptability at various levels.

Red Team Assessments: Simulations where an independent group mimics potential security threats to identify vulnerabilities in systems or facilities.

Technology Transfer: The sharing of technology, expertise, and resources from developed to developing regions to ensure equitable access and benefits.

Weighted Voting System: A voting structure where votes are weighted based on certain criteria, such as technological capacity or contribution level, to balance influence among participants.