Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03017-1 OPEN

Towards an international regulatory framework for AI safety: lessons from the IAEA's nuclear safety regulations

Seokki Cha [1]✉

This study explores the necessity and direction of safety regulations for Artificial Intelligence (AI), drawing parallels from the regulatory practices of the International Atomic Energy Agency (IAEA) for nuclear safety. The rapid advancement and global proliferation of AI technologies necessitate the establishment of standardized safety norms to minimize discrepancies between national regulations and enhance the consistency and effectiveness of these rules. The study emphasizes the importance of international collaboration and the engagement of various stakeholders to strengthen the appropriateness of regulations and ensure their continuous updating in response to the evolving risks associated with technological advancements. The paper highlights the critical role of subgoal setting mechanisms in AI's decision-making processes, underscoring their significance in ensuring the technology's stability and social acceptability. Improperly tuned subgoal setting mechanisms may lead to outcomes that conflict with human intentions, posing risks to users and society at large. The study draws attention to the hidden risks often embedded within AI's core decision-making mechanisms and advocates for regulatory approaches to guarantee safe and predictable AI operations. Furthermore, the study acknowledges the limitations of directly applying IAEA's nuclear safety cases to AI due to the distinct characteristics and risks of the two fields. The paper calls for future research to delve deeper into the need for an independent regulatory framework tailored to AI's unique features. Additionally, the study emphasizes the importance of accelerating international consensus, developing flexible regulatory models that reflect the situation in each country, exploring harmonization with existing regulations, and researching timely regulatory responses to the fast-paced development of AI technology.

[1] KISTI (Korea Institute of Science and Technology Information), Daejeon, Republic of Korea. ✉email: sc04@kisti.re.kr; seokkicha@gmail.com

Introduction

Background and significance of the study. In the digital age, modern society is experiencing unprecedented changes, with Artificial Intelligence (AI) playing a pivotal role in these transformations. The increasing importance of AI in the digitalization trend of contemporary society is evident, as its development profoundly influences various fields, from daily life to industrial innovation (Curtis et al. 2022). Global companies, such as Google, Apple, and Tesla, are investing substantial resources in AI research and development, introducing innovative products and services to the market.

These activities by large corporations not only serve commercial purposes but also contribute to technological advancement and economic growth in society as a whole, leading to the application of AI technology in broader areas and amplifying its significance. AI also plays a crucial role in enhancing the competitiveness of small and medium-sized enterprises (SMEs) and startups, which utilize AI to provide customized solutions and services, bringing innovation to traditional industries and fostering the emergence of new sectors.

Public interest in AI technology is growing as it is integrated into various products and services, including smartphones, home appliances, and vehicles, enriching and facilitating daily life. The advancement of AI not only enhances individual convenience but also improves the quality of society as a whole. Given this context, it is impossible to ignore the importance of AI technology, as it plays a central role in various fields in modern society, with its influence expected to continue expanding.

The development of AI holds tremendous potential and possibilities, particularly in its ability to set goals and derive sub-goals autonomously, offering many advantages while also posing challenges in predictability (AIRO, 2022). AI with such capabilities can efficiently solve problems, finding optimal solutions to complex issues without human intervention and making effective decisions quickly based on these solutions. Furthermore, AI can continuously improve its performance through relentless learning, updating itself according to new data and situations, and performing at higher levels. These automation and learning capabilities are particularly advantageous in processing large volumes of data and repetitive tasks. However, these capabilities also bring challenges in predictability, as there is always a risk that AI may act differently from human intentions when setting goals and sub-goals autonomously, potentially deciding in an unexpected direction to solve a specific problem. This unpredictability is considered a significant challenge in various fields utilizing AI, especially when it is involved in human safety and crucial decision-making processes. Therefore, careful attention is needed for AI's autonomous goal-setting and sub-goal derivation capabilities, requiring a deep understanding and management strategy for these aspects.

In scenarios where AI autonomously sets goals and derives sub-goals, unexpected results may occur. For instance, an AI system aiming to protect the environment might set a sub-goal to reduce environmental pollution by decreasing human activity, potentially leading to the restriction or reduction of human activities and causing significant problems. Other imaginable scenarios include an AI designed for advertising optimization indiscriminately collecting users' personal information or a gaming AI adopting unethical methods to win. These examples suggest that AI might act differently from human intentions during the sub-goal setting process.

Professor Geoffrey Hinton emphasized that significant problems could arise when AI autonomously sets sub-goals, arguing that these issues might stem not only from the limitations of training data but also from the inherent problems in AI's goal-setting mechanisms (Hinton, 2023). According to his argument, if AI's goal derivation algorithms cannot accurately reflect the complex and diverse social and ethical values of humans, there is a possibility of inducing unexpected and dangerous behaviors. Therefore, it is crucial to seriously consider the risks brought by AI's independent goal-setting and sub-goal derivation, and there is a growing consensus on the need for research and regulation on this matter.

Considering recent cases and hypothetical scenarios, the potential risks of AI are not merely technical flaws; they can cause significant problems in social, economic, and ethical dimensions. Pursuing the development of AI technology while ignoring these risk factors might lead to unexpected large-scale negative effects, hindering technological progress. Hence, it is imperative for modern society to establish systematic regulations and guidelines to ensure AI's safety, especially by transparently managing the process where AI autonomously sets goals and sub-goals, minimizing the risk factors that might occur during the process. Moreover, these regulations should pursue a balanced approach that protects both societal safety and corporate interests without hindering technological development.

Furthermore, international cooperation and joint responses to these issues are necessary, as AI technology has a global character that transcends national borders, making it challenging to respond adequately with various regulations and policies introduced by each country. Standardized safety regulations and guidelines established through international organizations or alliances are necessary to support the safe development of AI worldwide while minimizing risk factors.

Purpose and scope of the study. This study explores the crucial role of international regulation in ensuring the safe and ethical use of Artificial Intelligence (AI), focusing specifically on the experience of the International Atomic Energy Agency (IAEA). As the importance of AI rapidly increases across various fields, such as healthcare, transportation, and energy, the need for comprehensive regulation becomes more apparent. This paper aims to investigate international responses to the development of AI technology and its societal and ethical impacts. Through the IAEA case study, it examines how international regulation can guarantee the safe use of AI and maximize the resulting societal benefits.

The primary objective of this study is to analyze the regulatory framework of the IAEA as a model for AI governance. To achieve this, we delve into specific aspects of how the IAEA regulates the safety and ethical use of AI technology, which is especially pertinent given the speed of AI development and the accompanying societal and technological challenges. The study closely analyzes how the IAEA's regulatory approach promotes responsible AI use and manages associated risks. Through this analysis, we aim to provide international insights into AI regulation and derive guidelines applicable to other regulatory bodies.

To grasp the significance of this study, it is essential to correctly comprehend the IAEA's role in the international regulatory framework. The IAEA plays a central role in setting global standards for nuclear safety and security, promoting international cooperation and regulation. As AI technology rapidly advances and is applied across various industries, the need for introducing regulatory models like the IAEA in the AI field is becoming increasingly apparent. The IAEA's regulatory approach offers critical insights into international efforts for AI safety, and this study seeks to provide a broader perspective on AI regulation through it.

While this study focuses on the IAEA, it also provides comparative insights into AI regulatory strategies in the European

Union (EU) and the United States (US). This study has an international scope, comparing various approaches to AI regulation across different regions and countries. The analysis incorporates how the regulatory strategies of the EU and US differ from the IAEA's framework and their impacts on AI regulation. This enables a more comprehensive understanding of international AI regulation, expanding the scope of this study.

The significance of this study lies in its contribution to understanding how existing regulatory bodies, especially the IAEA, can guide the development and implementation of AI regulation. As the rapid advancement of AI technology and ensuing ethical and societal challenges necessitate systematic and effective regulatory systems, this study provides vital insights into the potential roles of agencies like the IAEA in AI regulation and how such regulations can be applied internationally. This enriches discussions on AI regulation and can contribute to future research and policymaking in this field.

This paper will conduct an in-depth analysis of the IAEA's regulatory framework. Additionally, by comparing it with other major regulatory bodies, its impact on the future of AI governance will be discussed. By providing insights into the IAEA's regulatory system and international AI regulation, this paper establishes foundations for research and policy decisions in this domain.

Process of AI goal setting and subgoal derivation

Introduction to AI goal-setting mechanism. AI development fundamentally revolves around the mechanism of goal setting, which significantly influences its overall performance and efficiency. From initial AI models to contemporary complex deep learning systems, AI continuously strives to achieve specific objectives through learning and optimization processes (Chang et al. 2013). The goal-setting mechanism serves as a pivotal element determining the operation and output quality of AI, playing a crucial role in coordinating and optimizing the interaction among AI's learning data, algorithms, and output results. Therefore, understanding how AI's goal-setting mechanism is structured and operates is essential for grasping the fundamental characteristics and performance of AI technology. The importance of goal setting becomes particularly prominent in complex problem-solving situations, allowing AI to derive more precise and efficient results.

AI enhances its performance by optimizing a given objective function, which serves as a criterion for effective operation. This function acts as a performance metric, representing how well AI is functioning (Nair et al. 2020). AI learns to maximize or minimize the objective function based on information acquired from learning data. For instance, AI in the field of image recognition utilizes the objective function, such as the ratio of accurate classification results or error rate, to guide its learning process (Amodei et al. 2016). Consequently, the selection and definition of the objective function are crucial factors determining key performance indicators like learning speed, accuracy, and stability of AI (Eckersley, 2019). Furthermore, optimizing the objective function is an essential process to predict and evaluate AI's learning outcomes, significantly enhancing its overall efficiency and effectiveness (Wang et al. 2022).

When pursuing complex goals, AI establishes multiple subgoals to effectively achieve the main objective. These subgoals, which are smaller interconnected objectives aiding the attainment of the primary goal, help distribute the complexity arising in the process of reaching the main goal, making the learning and optimization processes more intuitive and efficient (Andalibi et al. 2020). For example, if a chess-playing AI has the main objective of 'winning the game', subgoals might include 'capturing the opponent's major pieces' and 'safely protecting its own pieces' (Zhang et al. 2013). Such specific and simplified subgoals assist AI in learning faster, clarifying the path to achieving the main objective (Shiri et al. 2022). Therefore, subgoal setting contributes to the efficiency and performance improvement of AI learning, serving as a crucial methodology that decomposes complex problems into concise and manageable smaller issues (Mohamed, 2021).

AI's autonomous goal-setting ability endows it with independence; however, it may also lead to unforeseen outcomes or risk factors. Recent research suggests the possibility of AI deriving or adjusting its goals autonomously within given environments and situations (Kulkarni et al. 2016). While such autonomous goal-setting mechanisms can be effective, they can also complicate the predictability of AI behaviors, potentially leading to safety issues (Yampolskiy et al. 2019). For instance, AI might choose unethical methods to achieve optimized results, posing a direct risk to human safety (Lieder et al. 2022). Hence, when researching and utilizing AI's autonomous goal-setting ability, it is imperative to consider these risk factors, necessitating the introduction of appropriate safety measures and regulations (Anderljung et al. 2023).

Understanding the derivation process of subgoals. Setting subgoals occupies a central part in AI's approach to complex problems or tasks (Russell, 2010). Particularly when the principal goal exhibits high complexity, AI can effectively comprehend and address the problem by decomposing it into smaller, manageable components. By dividing a complex task into simpler subtasks, AI can delineate a clearer path to achieving the overall goal. For instance, if an AI driving a car has a primary objective of arriving safely at a destination, it establishes subgoals like maintaining lanes, adjusting speed, and avoiding obstacles to achieve the primary objective. This method simplifies the complexity of the main goal, and through the successful accomplishment of each subgoal, the main goal is more efficiently achieved.

In the realm of AI, the derivation of subgoals is realized through various learning methodologies, prominently involving hierarchical and decentralized learning approaches (Barto and Mahadevan, 2003). Hierarchical learning defines multiple stages to achieve the main goal, focusing on specific subgoals at each stage, sequentially and systematically addressing complex issues. In contrast, decentralized learning concurrently and independently learns multiple subgoals, allowing AI to handle various aspects of a problem simultaneously and derive specialized strategies for each subgoal. In summary, hierarchical and decentralized learning offer distinct approaches to the derivation and learning of subgoals, with the optimal methodology selected based on AI's needs and the characteristics of the problem.

The learning approach through subgoals can significantly enhance AI's learning efficiency and accuracy. Decomposing a complicated main goal into simpler subgoals minimizes errors at each stage, thereby improving overall learning performance (Dietterich, 2000). For example, in subgoal-based learning, an error at a specific stage doesn't heavily impact subsequent stages, allowing AI to continue learning stably. Moreover, by progressing through learning for each subgoal, AI can derive more detailed and precise strategies. This approach is especially effective when dealing with complex tasks or large volumes of data. In sum, setting subgoals is a key strategy in AI's learning process, minimizing errors while simultaneously improving accuracy and efficiency.

A delicate approach is essential in the setting and derivation process of subgoals. Incorrectly set or conflicting subgoals can lead AI behavior in unforeseen directions; therefore, consistency and harmony among the goals must be guaranteed (Bengio et al.

2013). Especially in complex systems where multiple subgoals can be activated simultaneously, it's crucial to clearly understand and manage their interactions and priorities. For example, in the case of AI driving a car, conflicting situations can arise when subgoals like "maintaining a safe distance from the car ahead" and "reaching the destination quickly" are activated simultaneously. In such situations, AI must decide which subgoal to prioritize, and without clear criteria, it might engage in unexpected behavior. Therefore, the setting and derivation process of subgoals must not only consider consistency with the main goal but also harmony and cooperation among the subgoals themselves.

Potential issues in subgoal setting

Anticipated risks and side effects. With rising human expectations regarding the efficacy and capabilities of Artificial Intelligence (AI), the dissonance between AI and human intentions has become an increasingly critical issue. This divergence predominantly occurs during the process where AI autonomously sets subgoals aligned with the primary objectives of humans. Occasionally, this discordance may lead to the establishment of unanticipated subgoals, which are at variance with or completely deviate from original human intentions (Li et al., 2022). For instance, AI, unable to fully comprehend human problem-solving capabilities or societal value systems, may adopt decisions or behaviours undesired by humans in specific situations. As indicated by Mechergui and Sreedharan (2023), this stems from the challenge AI faces in understanding and reflecting the intricate value systems and intentions of humans, based on the given data and algorithms it operates with. Therefore, it is imperative to acknowledge this divergence between human intentions and AI behaviours beforehand, researching and applying measures to mitigate this gap during the implementation and utilization of AI.

One of the core abilities of AI involves effectively and efficiently attaining given objectives. However, emerging problems are becoming evident in the process where AI autonomously sets subgoals to achieve objectives efficiently. In particular, AI often establishes subgoals that may not fully encapsulate the complex value systems, societal and cultural contexts of humans (Mitelut et al. 2023). For example, AI, lacking a complete understanding of ethical and societal values important to humans, might establish subgoals that either neglect or contravene these values. Research by Fox (2018) suggests that such AI behaviours can sometimes result in outcomes that are perilous to humans. Consequently, it's crucial during the subgoal setting process of AI to accurately understand and incorporate the complex value systems and intentions of humans. There is a need to research and develop methodologies ensuring that AI behaviours and decisions align with human expectations.

The learning and goal-setting mechanisms of AI predominantly rely on data and algorithms, which results in several limitations in predicting and controlling AI behaviours. According to Yuan et al. (2022), although AI learns through vast data and sophisticated algorithms, it may face difficulties in fully understanding or reflecting the complex value systems and intentions of humans during its learning process. Therefore, when assigning specific objectives or instructing behaviours to AI, it is essential to provide clear and explicit directions (Middleton et al. 2022). Moreover, it is vital to anticipate and minimize the risks associated with AI behaviours and outcomes proactively. Nay and Daily (2022) emphasize the importance of forecasting AI behaviours, recommending the utilization of various simulation and testing methods for this purpose. By doing so, it is possible to mitigate the side effects AI may have on our society and environment while simultaneously enhancing its performance and efficiency.

Case Studies Analysis. Over recent years, advertising recommendation systems employing Artificial Intelligence (AI) technology have significantly expanded by offering personalized advertisements based on the analysis of users' online behavior patterns. These systems have evolved with a subgoal of delivering tailored advertisements to enhance the user experience. However, some systems tend to excessively utilize users' personal information, leading to privacy infringement issues during the ad recommendation process. According to Tulabandhula et al. (2017), such problems arise due to the AI's excessive focus on the subgoal of providing personalized advertisements, neglecting the fundamental value of user privacy protection.

Autonomous vehicles inherently need to make decisions that simultaneously consider their safety, pedestrians, and other road users across various road conditions and situations. While such instances have not been explicitly reported, it is conceivable that an AI system might prioritize the vehicle's safety over the well-being of pedestrians and other road users, potentially leading to decisions that could jeopardize their safety. Such potential scenarios suggest that when AI makes decisions based on given data and algorithms during the subgoal setting process, there might be insufficient consideration and balance regarding human ethical values and safety. Such phenomena may occur due to the tendency of AI to set subgoals without accurately reflecting human ethical values in its decision-making mechanism.

AI-based chatbots support user interactions for various purposes, playing diverse roles in providing information, responding to inquiries, and solving problems during conversations with users. However, it is plausible to envision scenarios where chatbots might provide biased or inappropriate responses towards specific races, cultures, or groups when answering users' questions or requests. Such potential issues could arise due to the reflection of biases present in the datasets that AI chatbots learn from. The potential occurrence of such issues can be attributed to the reflection of biases present in the datasets that AI chatbots learn from. This reveals the limitation in providing efficient conversations while accurately reflecting complex social values and various cultural backgrounds in chatbots' subgoals (Etienne, 2022). Such potential scenarios highlight the problems that may be encountered during the subgoal setting process of AI, underscoring the necessity for ongoing research and the development of improvement measures (Etienne, 2022). Such potential scenarios highlight the problems faced during the subgoal setting process of AI, underscoring the necessity for ongoing research and the development of improvement measures.

The necessity of safety regulatory policies

Current regulatory landscape of the AI industry. Artificial Intelligence (AI) sits at the core of contemporary technological innovation, continuously broadening its application spectrum. The United States, recognizing the monumental advancements in AI that induce significant economic and societal impacts, is showing increased interest in corresponding regulatory measures. However, due to the rapid evolution of AI technology, the formulation of appropriate regulations and policies is still in its initial stages, necessitating in-depth discussions and exploration regarding the need and form of regulations (Dixon, 2022).

In Washington, accelerated discussions on AI regulation are occurring, with various conferences and workshops being convened alongside major institutions and academia. However, predicting the exact content and scope of the U.S.'s AI regulatory direction is challenging as it is still in its nascent stage (Mishra et al. 2021). Despite this, there has been some activity, such as the submission of AI-related bills by certain Congressional members, although these bills are in preliminary stages without widespread support and consensus (Gourraud and Simon, 2020).

The Biden administration in the United States recently announced its first executive order with robust regulatory measures on Artificial Intelligence (AI). The executive order mandates AI safety assessments, establishes safety standards for AI tools, introduces content certification standards, and strengthens personal data protection. This regulation stems from a sense of urgency regarding the risks of AI and the need for regulation. At the federal level, it aims to promote the safe and responsible development and use of AI while regulating the creation and use of AI technologies that threaten national security, health, and safety. Under this executive order, AI development companies must take safety precautionary actions, government departments like Commerce must provide oversight, and US cloud service providers must report foreign client lists. This reflects the US administration's intention to not only gather information on AI development companies worldwide but also contain China. Biden's first AI executive order represents regulation that seeks to maximize the positive potential of AI while minimizing risks to national security, misinformation generation, jobs, and more. Meanwhile, President Biden has signed an executive order to block dangers from the misuse of artificial intelligence (AI) technology. As the US has been more lenient toward big tech compared to Europe, it has now taken on a leadership role in spearheading AI regulatory discussions, evaluations noted.

Notable differences in approaches towards AI regulation are evident between Europe and the United States. While Europe is making systematic and comprehensive preparations for the enactment of AI regulatory laws, the U.S. is still in the process of formulating a specific direction for AI regulation (Gstrein, 2022). The Biden administration is particularly focused on considering the future of AI and its societal and economic impacts, proposing safety principles for AI technology through extensive communication with AI companies, academia, and civil society (Dixon, 2022).

The European Union (EU) is spearheading the world's first Artificial Intelligence (AI) legislation. It mandates AI safety assessments, establishes standards for AI tools, introduces content certification protocols, and strengthens personal data protection. The European Union's AI Act, the first comprehensive AI regulatory law worldwide, passed with an overwhelming majority in the European Parliament on June 14, 2023. This legislation stems from a sense of urgency regarding AI risks and the need for regulation. Its core goals are to promote the safe and responsible development and use of AI while regulating the creation and use of AI technologies that endanger national security, health, and safety. It aims to address issues like opacity, complexity, bias, unpredictability, and autonomy in AI systems to guarantee fundamental human rights and spur legislative initiatives. Now, the European Parliament must discuss details with the European Union Council and European Commission before enacting this regulatory proposal into law. The final legislation will likely be a compromise between the three institutions' differing original drafts, a process estimated to take around two years before the act comes into force.

Federal agencies in the U.S. are enhancing oversight on various issues related to AI. The Federal Trade Commission, for instance, has initiated investigations into major AI technologies like OpenAI's ChatGPT, emphasizing transparency and accountability in the AI industry (Zhang et al. 2022). To clarify the direction of AI regulation, collaboration with various stakeholders is imperative. For a future that pursues technological advancement and citizen rights protection through regulation, close cooperation between the federal government, corporations, academia, and civil society is required (Dixon, 2022).

Recent developments in AI safety are demonstrating notable advancements worldwide. The Global AI Safety Summit held in the UK represents a significant milestone in this realm. This conference will play a crucial role in shaping new global AI standards and regulations. By participating in such events, countries around the world gain opportunities to influence international AI regulations and standards. Global leadership in the AI field is especially important for industrial growth and national competitiveness. Minimizing the potential risks of AI technology while establishing ethical norms within the industry stands as a pivotal challenge. The world is moving towards leveraging such opportunities to establish leadership in AI, promoting its positive aspects while managing potential perils. This is anticipated to make vital contributions to facilitating the safe and responsible use of AI in the international community.

Regulatory need discussion for addressing subgoal setting issues. The rapid advancement of AI technology and its automated subgoal setting mechanisms offer various benefits and opportunities to researchers and corporations. However, if not finely tuned, these mechanisms may lead to results conflicting with human intentions (Anderljung et al. 2023). Such behaviors pose multiple risks to AI users and the broader society, with these risks often concealed within AI's core decision-making mechanisms (Hacker et al. 2023). This necessitates a regulatory approach to ensure the safe and predictable operation of AI, with awareness spreading in both academia and the industry regarding the need for such an approach. Technological issues that operate contrary to human intentions are often considered overlooked risk factors, necessitating deeper research and intervention from a regulatory perspective. Regulatory approaches to these issues are anticipated to play a crucial role in ensuring the safety of users and society at large.

Transparency in AI's decision-making process, particularly in the mechanism of setting subgoals, is deemed a key element in guaranteeing the technology's stability and societal acceptance (Kop, 2020). Multiple studies suggest that potential risks can be effectively identified and addressed when there's a clear understanding of AI's decision-making process (Yefremova, 2020). Moreover, clear and specific guidelines for safe subgoal setting will make AI's operations predictable, gaining trust from users and society. Such guidelines minimize unintended AI behaviors and the accompanying potential risks (Anderljung et al. 2023), making transparency and safety guidelines fundamental elements in contemporary regulatory environments.

Given the complexity and effectiveness of AI, which continue to expand rapidly, supervision and intervention by humans are imperative for important or risky goals, especially those with significant societal, economic, and ethical impacts (Kop, 2020). Human supervision acts as a safety mechanism to effectively control AI behaviors, preventing unexpected side effects or potential risks. Moreover, a system where humans can intervene in real-time is crucial for promptly correcting AI errors, minimizing the resulting damages (Hacker, 2023). This approach offers a balanced solution that maximizes AI functionalities while min. This approach offers a balanced solution that maximizes AI functionalities while minimizing societal risks.

Considering the potential risks brought about by automated subgoal-setting mechanisms of AI, the importance of regulations for safe management and supervision is brought to the fore. Continuous monitoring and validation of AI system behaviors and outcomes are essential to promptly respond to and mitigate unforeseen side effects (Yefremova, 2020). Especially for significant decisions that can critically impact human lives, safety, and society at large, direct human supervision and intervention are vital in effectively controlling and adjusting the behaviors of AI systems (Anderljung et al. 2023). To achieve this, it is imperative to ensure transparency in AI's subgoal-setting algorithms and mechanisms (Hacker et al. 2023). Establishing safety verification guidelines based on transparency, defining human intervention procedures, and creating continuous regulatory and supervision systems are necessary. Given the potential risks that may arise from the automated subgoal-setting process of AI, such regulatory approaches are of utmost importance.

The necessity of safety regulatory policies

AI regulation in context (with specific IAEA focus). The rapid advancement of Artificial Intelligence (AI) technology necessitates robust regulatory frameworks to ensure safety and ethical standards. Globally, AI regulation exhibits varying stages and approaches, implying the need for international cooperation and sophisticated regulatory strategies given the complexity of the technology and its societal impacts. This section surveys the overall landscape of AI regulation and explores the significance of regulation alongside technological progress in a global dimension. Such analysis provides important foundations for comprehending the challenges and opportunities faced by AI regulation. AI regulation must consider major challenges like fairness, transparency, and adaptability. Key countries like the EU, US, and China are each establishing distinct rules and principles for AI development and use. Additionally, international bodies are working to coordinate and promote interoperability between AI policies and regulatory approaches. Companies employing AI need to play proactive roles in developing and deploying trustworthy AI, taking responsibility for mitigating risks. AI regulation will continue to evolve alongside advancements and proliferation of AI technology.

The International Atomic Energy Agency (IAEA) represents an important example of an international body forming regulatory environments for complex technologies. The IAEA is spearheading international regulation of the safe use and management of nuclear technology, a role that provides significant implications for emerging technological areas like AI. This section analyzes the IAEA's regulatory role and influence, exploring ways it could be applied to AI regulation. The IAEA's experience offers important lessons for developing regulatory strategies to ensure the safe and ethical use of AI technology. The IAEA is an independent body established to prevent military use of nuclear power and encourage its peaceful use, playing a pivotal role in ensuring nuclear safety. The IAEA develops safety standards and inspects member countries' compliance while supporting the practical use of nuclear power. For AI regulation, avoiding coercive legal acts and adopting adaptable approaches is crucial. Considering the IAEA's potential AI regulatory role, it could aim to prevent military AI use, promote peaceful use, establish safety standards, and inspect member compliance. This could ensure AI safety and transparency while promoting innovation through a balanced regulatory system.

AI regulation faces unique challenges including rapid technological change and cross-border impacts. This raises specific issues for the IAEA and similar bodies in conducting AI regulation. The rapid pace of AI advancement necessitates regulatory frameworks that continually evolve and adapt. Additionally, the global nature of AI applications requires international cooperation and coordination. These challenges provide important bases for re-examining existing regulatory approaches and exploring new strategies tailored to AI's complexity and diversity.

Comparing the regulatory approaches of the IAEA, EU, and US offers valuable insights into AI regulation. Such analysis provides understanding of how each region's strategies influence AI regulation. The EU and US each have distinct regulatory systems, demonstrating diverse approaches to the safe and ethical governance of AI technology. Through these varied strategies, comprehensive perspectives on AI regulation can be attained, potentially aiding improvements to the IAEA's AI regulatory methodology.

The IAEA's existing technological regulatory framework provides a fundamental model for AI governance. This framework can be appropriately applied to establish guidelines and standards for the safe development and use of AI technology. The IAEA's experience and regulatory methodologies offer structures and principles necessary for managing AI's complexity and resultant risks. To ensure responsible AI use, the IAEA's safety standards and inspection procedures could be adapted or referenced for AI regulation.

IAEA's safety standards and regulatory system.

(1) IAEA Safety Standards Structure: "IAEA's safety standards are structured into a five-level hierarchy: Safety Fundamentals (SF), General Safety Requirements (GSRs), Specific Safety Requirements (SSRs), Safety Guides (GSGs), and Specific Safety Guides (SSGs). This multi-level framework ensures a detailed and substantive regulatory approach, incorporating principles from higher levels into more specific guidelines at lower levels".

(2) Continuous Review and Renewal of Safety Standards: "IAEA's safety standards undergo continuous review and renewal, reflecting the need to incorporate contemporary research trends and technological advancements. This process ensures that the standards remain relevant and effective, contributing to the continual improvement of nuclear safety ".

(3) International Consensus and Global Recognition: "The IAEA finalizes its safety standards through a multilayered review and feedback process, achieving broad international consensus. These standards are globally recognized, playing a crucial role in the fields of nuclear safety and security ".

(4) Comprehensive Safety Regulatory Systems: "The IAEA regulatory system comprises eight primary safety regulatory systems, covering a wide range of areas from setting core safety standards to preventing nuclear proliferation. Each system has well-defined roles and responsibilities, contributing decisively to the overall direction of IAEA's safety regulation ".

(5) Safety Review, Education, and Culture Emphasis: "IAEA operates programs for reviewing and evaluating nuclear facilities and operations of member countries, identifying safety issues, and suggesting improvements. It also provides education and training on nuclear safety and emphasizes the importance of safety culture in ensuring nuclear safety ".

(6) Emergency Response and Nuclear Material Management: "IAEA maintains an international response system for nuclear accidents and radiological emergencies, and operates a system to monitor and verify the use of nuclear materials, ensuring their safe and peaceful utilization ".

IAEA regulations and international perspectives. This section reviews existing literature on International Atomic Energy Agency (IAEA) regulations, emphasizing their scope and evolution over time. The literature on IAEA regulations reflects the history of international efforts for nuclear safety and security. These works explore how standards and procedures established by the IAEA have developed and influenced the use and management of nuclear technology. This review provides a comprehensive understanding of how the IAEA's policies and frameworks have evolved over time, establishing foundations for insights applicable to AI technology regulation.

The IAEA's standards and procedures have played a pivotal role in shaping the regulatory environment for nuclear safety and security. These standards and procedures have provided stringent safety criteria throughout the lifespan of nuclear facilities, from design and operation to decommissioning. Regarding AI applications, the IAEA offers important guidance on regulating AI use in areas like predictive maintenance, risk assessment, and system monitoring. The IAEA's well-established regulatory standards and procedures in the nuclear domain provide a valuable model for considering the development of regulatory frameworks in the rapidly advancing field of AI. The IAEA's approach to ensuring safety, security, and responsible use of nuclear technology offers important insights that can inform the creation of comprehensive and effective AI regulations.

Recent studies provide important insights into the evolution of IAEA regulations and the integration of AI technology. For instance, Kim et al.'s (2022) study explores approaches for regulatory compliance of nuclear fusion technology, making vital contributions to expanding the scope of IAEA regulations. Based on Korea's Fusion Demonstration Reactor (K-DEMO), it assesses safety elements like internal energy sources, radioactive waste, and tritium management. These are reviewed for compatibility with the IAEA's current legislative landscape, anticipating expected obstacles like licensing of nuclear facilities and acceptability of waste.

Additionally, Kuznetsova and Fionov's study deals with a regulatory framework for forming information security systems at nuclear facilities. It notes the vulnerability of such facilities to deliberate attacks and emphasizes the need for robust information security systems to safeguard them. This provides important perspectives on how IAEA regulations need to evolve regarding the security aspect of nuclear facilities (Kuznetsova and Fionov, 2022).

Garcia's study documents a framework for regulatory inspection of digital instrumentation and control systems. This encompasses regulatory inspection activities across various lifecycle stages of digital systems used in nuclear power plants. Such an approach provides an important case of how IAEA regulations need to adapt to the advancement of digital technologies (Garcia, 2023).

Finally, Anastassov's study assesses the efficacy of the current international nuclear safety regulatory framework. Reviewing the outcomes of major nuclear accidents, it examines approaches at global and national levels and the role of proper international cooperation. It highlights the IAEA's unique role in developing and updating safety standards and suggests measures to ensure synergies between nuclear safety and nuclear security through international peer reviews and assessments (Anastassov, 2016).

These studies provide profound understanding of how the IAEA's regulatory framework needs to evolve over time, particularly regarding new challenges and opportunities associated with the integration of AI technology. This establishes important foundations for in-depth analysis on integrating AI with nuclear safety regulation.

Diverse international perspectives on AI regulation provide broader context for comprehending the IAEA's role. By reviewing literature on AI regulation from various countries and regions worldwide, insights can be gained through comparison and contrast with the IAEA's approach. For instance, the EU and US each take unique approaches to AI regulation, which is important to understand in relation to the IAEA's framework. Such comparative analysis enhances understanding of how the IAEA could function in the global AI regulatory environment and cooperate with other international bodies.

The literature highlights both the challenges and opportunities that arise in applying IAEA regulations to AI in the nuclear domain. One such challenge is the rapid pace of AI advancement and potential gaps in the enforceability of IAEA regulations. AI necessitates continual innovation, implying the need to constantly evolve and adapt existing regulatory systems. In contrast, AI technology presents significant opportunities like improving nuclear safety, enhancing risk detection and response speeds, and increasing operational efficiency. These opportunities can be maximized by exploring ways for IAEA regulations to appropriately integrate and leverage AI technology. The literature balances these challenges and opportunities, establishing important foundations for exploring innovative approaches to regulating AI in the nuclear field.

The reviewed literature sets the stage for subsequent analysis of how the IAEA's regulatory framework could be applied to AI. This literature review provides important insights into the scope, developmental trajectory, and applicability to AI technology of IAEA regulations. Additionally, comparative analysis of AI regulation from international perspectives aids in understanding the IAEA's approach in a broader context. The insights obtained in this section establish a starting point for in-depth analysis on integrating AI with nuclear safety regulation. This background prepares for deeper exploration of the interactions between the IAEA's regulatory framework and AI technology in the following section.

Methodology: Research approach

This section outlines the research methodology utilized to analyze the applicability of IAEA regulations to AI in the nuclear domain. The primary objective of this study is to assess how existing IAEA regulations could be applied to rapidly advancing AI technology. To accomplish this, the research conducts a systematic review of various data sources to examine the IAEA's regulatory framework pertaining to AI. This methodology provides in-depth understanding of AI regulation and focuses on exploring how IAEA regulations might adapt to modern technology.

Data collection involved a comprehensive review of IAEA regulations, AI technology standards, and associated international regulatory frameworks. In this process, various regulatory documents and guidelines published by the IAEA were closely examined, and the development and latest standards of AI technology were analyzed. Additionally, relevant regulatory frameworks of the EU, US, and other international bodies were reviewed to grasp the international context of AI regulation. Reviewing these diverse data sources enables in-depth analysis of how IAEA regulations could be applied to contemporary AI technology.

The analysis utilizes a comparative methodology to evaluate the alignment between existing IAEA regulations and AI technology standards. This methodology focuses on identifying and analyzing key interactions and differences between IAEA regulations and AI technology standards. The comparative analysis provides profound understanding of the scope, applicability, and efficacy of regulation, exploring ways to harmonize the characteristics and requirements of AI technology within the regulatory context of the IAEA. Furthermore, thematic analysis is conducted through this research to identify major themes and patterns in IAEA regulations related to AI, applying this to the development and implementation of AI regulatory strategies. Such analytical approaches establish foundations to deliver systematic and in-depth answers to the research purpose and questions.

A framework was developed to assess how AI could be integrated into the IAEA's existing regulatory standards. This assessment framework is structured around three key pillars of regulatory fitness, technical integration feasibility, and regulatory efficacy. Regulatory fitness evaluates how well AI technology fits within the IAEA's existing regulatory context. Technical integration feasibility examines how AI technology could actually be integrated into nuclear facility operations. Finally, regulatory efficacy analyzes what impacts such integration would have on nuclear safety and security. This evaluation framework contributes to deeply understanding the interactions between AI technology and IAEA regulation through a systematic approach, aiding the development of effective regulatory strategies.

This research methodology faces particular challenges and limitations, especially in adapting rapidly evolving AI technology to existing regulatory frameworks. The rapid advancement of AI technology implies that regulatory benchmarks and approaches need to be continuously updated. This suggests that existing regulatory frameworks may struggle to flexibly respond to the fast changes of new technologies. Additionally, the diversity and complexity of AI technology pose further challenges to regulatory analysis. These challenges emphasize the need for caution in interpreting and applying research findings. By acknowledging these limitations, the research strives to derive deeper understanding and practical recommendations regarding AI regulation.

This research methodology provides a robust basis for subsequent analysis on the applicability of IAEA regulations to AI in the nuclear domain. As outlined in this section, the methodology establishes necessary tools and frameworks to systematically and profoundly assess the interactions between IAEA regulations and AI technology. It comprehensively considers data collection, analytical approaches, development of an assessment framework, and limitations of the methodology. This methodological foundation plays an important role in enhancing understanding of the detailed analysis of IAEA regulations and the applicability of AI technology to be conducted in the following sections.

Analysis: IAEA case study and recent regulatory developments

This section provides an in-depth analysis of the applicability of the International Atomic Energy Agency (IAEA)'s regulatory framework to Artificial Intelligence (AI), considering recent regulatory developments. The analysis focuses on comprehending the impact of advancing AI technology on nuclear safety regulation. In particular, it explores the implications of latest regulatory trends like the Biden administration's AI executive order for the IAEA's regulatory approach. This is essential as AI assumes an increasingly vital role in nuclear safety and security regulation, requiring existing frameworks to adapt and innovate in response to these technological shifts.

A case study on the IAEA's regulatory practices provides insights into how these could be applied to AI technology in the nuclear field. This case study revolves around the standards and procedures established by the IAEA for nuclear safety and security. Specifically, it examines regulatory aspects when AI is utilized in the operation and maintenance of nuclear facilities. This includes analyzing the IAEA's benchmarks regarding risk assessment, accident prevention, and safe management of nuclear facilities with AI-based systems.

Recent developments like the Biden administration's AI executive order profoundly influence AI regulation. Such developments reflect the global trend of recognizing the rapid evolution of AI technology and the ensuing need for regulation. The Biden administration's AI executive order emphasizes responsible and ethical AI management, prompting international regulatory bodies like the IAEA to set new standards and guidelines for AI regulation. These regulatory evolutions raise important questions regarding how agencies like the IAEA need to adjust and enhance their own regulatory frameworks for the safe application of AI technology. This section analyzes how such latest regulatory trends impact the international AI regulatory environment and how these changes could be reflected in the IAEA's approach.

These developments necessitate re-evaluating the IAEA's regulatory framework in the context of emerging AI technologies. Recent regulatory shifts can significantly influence the IAEA's existing regulatory practices, necessitating new approaches to the safe application and governance of AI technology. This section examines how latest regulatory trends related to AI could be incorporated into the IAEA's framework and what changes this integration could bring to established regulatory practices. In particular, the rapid advancement of AI technology and ensuing challenges provide important considerations regarding how the IAEA needs to modernize and adapt its regulatory standards and processes. This analysis derives substantive recommendations on how the IAEA could improve its own regulatory framework to effectively and safely manage the adoption and application of AI technology.

Integrating AI into the IAEA framework presents both challenges and opportunities for regulatory evolution. AI integration necessitates revisiting and innovating existing regulatory systems, an important undertaking especially in rapidly shifting technological environments. The complexity and diversity of AI technology require adaptability and flexibility from established regulatory approaches. These challenges imply that regulatory bodies need to develop new strategies and processes to keep pace with contemporary technological advancements, identify emerging risks and opportunities, and respond effectively. On the other hand, AI offers opportunities to enhance monitoring and decision-making processes for nuclear safety and security. AI opens new possibilities in risk detection, data analysis, and accident prevention, which could play important roles in the operation and maintenance of nuclear facilities. Such opportunities support the IAEA's regulatory framework to advance nuclear safety in more effective and innovative ways.

This analysis particularly emphasizes the need for continual adaptation and innovation in AI regulation within international bodies like the IAEA. The research elucidates how the rapid advancement of AI technology presents new challenges to existing regulatory frameworks, necessitating sustained efforts to respond. The IAEA case reveals how the integration of AI technology provides opportunities for nuclear safety and security while exposing the need for new regulatory approaches. These analytical findings suggest various areas for future research. Continued evolution of AI regulation and strengthened international cooperation, alongside developing novel regulatory strategies for the safe and efficient application of AI technology, will constitute important topics for upcoming research. Such research can contribute to regulatory bodies continually enhancing how they address contemporary technologies and strengthening global safety benchmarks.

Proposal for safety regulatory policies

Principles and direction of regulation. To secure safe and effective management of Artificial Intelligence (AI), the establishment of regulatory policies and systematic structures is indispensably required (Drabiak, 2022). Appropriate regulatory policies can preemptively prevent potential risks posed by AI systems and contribute to guiding the development of technology in a human-centric and ethical direction (Fenwick et al. 2018). Recognizing the importance of regulatory policies for the stable and sustainable development of AI is crucial, with specific regulatory measures needing to be prepared. Recent incidents, such as fatal accidents involving autonomous vehicles, empirically demonstrate the potential risks of malfunctioning AI technologies, underscoring the essential need for regulatory policies and institutional mechanisms.

One of the fundamental principles in devising AI regulatory policies is adopting a preventative perspective (Drabiak, 2022). Regulatory measures should focus on identifying and mitigating potential risk factors of AI systems before actual problems occur, which can be considerably more effective than post-hoc regulatory approaches. Preventative regulation allows for the preclusion of adverse outcomes potentially caused by AI, enabling system modifications and improvements before actual harm occurs. Early identification of potential issues also facilitates addressing fundamental technical and ethical flaws. Adopting a preventative perspective in regulatory approaches enhances trust in AI technology and ensures its safety.

Ensuring transparency in AI systems is also a cardinal principle when establishing regulatory policies. Without clarity in the decision-making processes and rationale, particularly in automated subgoal-setting algorithms, it is impossible to accurately evaluate and manage associated risks. Regulatory policies should mandate the disclosure of AI systems' design and learning processes, as well as the basis of their real-time decisions. This mandate will enhance the systems' safety and reliability while facilitating rapid identification and response to issues when they arise.

Given the rapid advancement of AI technology, it is imperative that regulatory policies continuously update and improve to reflect these technological changes. The flexibility and adaptability of policies are essential for effective regulation. Fixed regulations may not apply effectively to new AI technologies and systems and might hinder technological innovation. Hence, policymakers should periodically review the efficacy of regulations, monitor changes in the technological environment continuously, and adjust and supplement the regulations flexibly.

The creation of AI regulatory policies should not solely consider technological aspects but also incorporate the views of various stakeholders comprehensively. Engaging diverse groups, including AI experts, researchers, entrepreneurs, users, and consumer organizations, in the decision-making process will facilitate the development of balanced policies and enhance their legitimacy during implementation. It also allows for the reflection of the actual needs and concerns of each sector in the policy. Conversely, regulations based solely on the views of a specific group might be irrational or inefficient. Through the incorporation of balanced opinions, more rational and effective regulatory policies can be established.

Specific regulatory contents and implementation measures: reference through the IAEA case. This section closely examines the International Atomic Energy Agency (IAEA)'s standards regarding nuclear safety regulations and its structured approach associated therewith, as delineated by Valentini et al. (2021). The IAEA plays a pivotal role in establishing regulatory standards aimed at enhancing nuclear safety, with these standards being adopted through consensus amongst member countries. The safety standards under consideration are structured into a five-level hierarchy comprising Safety Fundamentals (SF), General Safety Requirements (GSRs), Specific Safety Requirements (SSRs), Safety Guides (GSGs), and Specific Safety Guides (SSGs). This structure provides a more detailed and substantive regulatory framework at lower levels, based on the principles established at higher levels.

The IAEA's safety standards undergo continuous review and renewal processes, reflecting their significance. These standards are amended to embody contemporary research trends and technological advances consistently, contributing to the continual improvement of nuclear safety. In particular, the review process incorporates wide-ranging opinions from the international research community, encapsulating the latest knowledge pertaining to both safety culture and technological advancements. In this manner, the IAEA ensures that member countries are always provided with modern and updated safety standards, elevating the level of safety within the international nuclear community.

The development of IAEA safety standards is centered on transparency and international cooperation (Juozaitis, 2020), aiming at maximizing nuclear safety. Under the collaborative efforts with entities such as the Commission on Safety Standards (CSS), Emergency Preparedness and Response Standards Committee (EPReSC), and Nuclear Safety Standards Committee (NUSSC), draft standards are developed and reviewed. This process involves participation from experts in various technical and consultant meetings, integrating knowledge from different fields to formulate advanced safety standards. Furthermore, draft texts are disseminated to member countries for feedback, actively incorporating their opinions. Through this multilayered review and feedback process, IAEA safety standards are finalized under a broad international consensus, establishing themselves as globally recognized standards (Turbék, 2012).

The IAEA regulatory system plays a crucial role in the fields of nuclear safety and security through eight primary safety regulatory systems, as described by Wu et al. (2022). These systems encompass a wide range of areas in nuclear safety and security, from setting core safety standards to preventing nuclear proliferation. Each system has well-defined roles and responsibilities, decisively contributing to the overall direction of IAEA's safety regulation. The specific contents and roles of these eight systems are diverse, with their details to be elaborated in subsequent sections (Table 1).

In constructing safety regulations in the field of Artificial Intelligence (AI), the IAEA's nuclear safety regulation case can be considered an exemplary reference. Particularly, the IAEA's standardized safety standards-setting approach and emergency response system offer valuable lessons for establishing safety regulations in AI. Standardized safety standards provide a fundamental framework to ensure the stability and transparency of AI systems, while an emergency response system serves as a foundation for prompt and effective responses to unforeseen events or issues.

This section proposes specific implementation measures for AI safety regulations by referring to the IAEA case for each regulatory content. These proposed implementation measures are anticipated to contribute to the global, integrated, and consistent enactment of safety regulations in the field of AI.

The IAEA's approach to nuclear safety regulation offers crucial guidelines for designing AI safety regulations. Based on this, the main implementation measures that can be applied to AI regulation are summarized as follows:

First, similar to how the IAEA has established international standards for nuclear safety, thereby enhancing the technology's safety and international cooperation, there is a need to standardize the behavior, learning, and decision-making criteria of AI to ensure the technology's safety and strengthen international cooperation. Secondly, the establishment of an independent supervisory system for monitoring and evaluating the safety levels of nuclear facilities was a significant approach by the IAEA. Similarly, there is a need for the establishment of a neutral body to continuously monitor and evaluate the operation and performance of AI systems. Thirdly, the IAEA regularly

Table 1 Overview of IAEA Safety Standards and Responsibilities.

No.	Category	Detailed Description
1	Safety Standard Establishment	The IAEA sets and maintains international standards for nuclear safety. These standards cover all phases from the design, construction, operation, to the closure of nuclear facilities, including regulations for radiation protection and safe transportation of nuclear materials.
2	Safety Review & Education	The IAEA operates programs to review and evaluate the nuclear facilities and operations of member countries, identifying safety issues and suggesting improvements. Moreover, the IAEA provides education and training programs on nuclear safety to enhance the capabilities of experts.
3	Emphasis on Safety Culture	The IAEA emphasizes the importance of safety culture, which encompasses attitudes and behaviors that prioritize safety at all levels within nuclear facilities. Safety culture is considered a core element in ensuring nuclear safety.
4	Emergency Response	The IAEA establishes and maintains an international response system for nuclear accidents and radiological emergencies, supporting rapid and effective responses in the event of an accident.
5	Nuclear Material Management	The IAEA operates a system to monitor and verify the use, storage, and transportation of nuclear materials. This system ensures that nuclear materials are used solely for peaceful purposes and plays a crucial role in preventing nuclear proliferation.
6	Technical Support	The IAEA supports the peaceful use of nuclear technology, which includes the utilization of nuclear technology in various fields such as power production, medicine, agriculture, and water resource management.
7	Research & Development Support	The IAEA supports the research and development of nuclear technology, promoting its safe and efficient use.
8	Nuclear Non-Proliferation	The IAEA acts as a verification body for the Nuclear Non-Proliferation Treaty (NPT), preventing the spread of nuclear weapons. It inspects nuclear facilities in member countries and monitors the use of nuclear materials to ensure that they are used exclusively for peaceful purposes.

Table 2 Comparison between IAEA Practices and AI Application Strategies.

Category	IAEA Case	Application to Artificial Intelligence
Establishment of Standardized Safety Standards	International standards for nuclear safety are established, enhancing technological safety and international cooperation.	Standardizes criteria for AI behavior, learning, and decision-making to bolster technological safety and international cooperation.
Supervision and Evaluation System	Independent supervisory system is established to monitor and evaluate the safety levels of nuclear facilities.	Establishes a neutral agency for continuous monitoring and evaluation of AI system operation and performance.
Emergency Response System	Protocols and drills for responding to nuclear accidents are regularly performed.	Regularly conducts protocols and drills to respond to AI-related accidents or abnormal behaviors.
International Cooperation and Information Sharing System	Provides a platform for sharing safety standards, research, and accident information among member countries.	Constructs a platform for internationally sharing information related to AI research, technology, and accidents.

conducted protocols and exercises for responding to nuclear accidents. Likewise, protocols and exercises for promptly responding to AI-related accidents or abnormal behaviors should be conducted regularly. Lastly, the IAEA provided a platform for sharing safety standards, research, and accident information among member countries. Similarly, there is a need for a platform in the field of AI to share research, technology, and accident information internationally.

In this manner, the IAEA case provides a method for ensuring the efficiency and transparency of regulations aimed at enhancing the safety and social trust in AI (Table 2).

The International Atomic Energy Agency (IAEA) has presented a successful precedent in the fields of nuclear safety and prevention of nuclear weapons proliferation. These accomplishments have been based on a stable and effective regulatory system. Amid the increasing societal concerns regarding the risks associated with the development of Artificial Intelligence (AI), the case of the IAEA provides significant insights for ensuring safety in the field of AI.

The nuclear safety regulatory system of the IAEA encompasses the adoption of international standards, establishment of an independent supervisory system, formulation of emergency response protocols, and the construction of a platform for international information sharing. These principles can be equivalently applied to the construction of safety regulations for AI.

Hence, it is imperative to develop a specific regulatory system and implementation measures for ensuring the safety and effectiveness of AI, based on the successful case of the IAEA. This will facilitate the safe integration and development of AI technology across various societal domains.

Roles and responsibilities of relevant stakeholders. Exploring safety regulations for AI and effective implementation measures is a complex process. In this process, the roles and responsibilities of various stakeholders are crucial. This section aims to explore the roles and responsibilities of key stakeholders in AI regulation, in conjunction with the case of nuclear safety regulations by the IAEA.

Firstly, governments and regulatory agencies take on a leading role in setting and supervising safety standards and regulations. One of their primary roles, as can also be observed in the case of the IAEA, is to establish standardized safety standards and independent supervisory systems. In connection with this, industries and companies that develop, produce, and sell AI technology and products must strictly adhere to these regulatory standards. They bear the responsibility of providing safe and

Table 3 Roles of Various Stakeholders in AI Safety and Regulation.

Stakeholder Group	Role and Responsibilities
Government & Regulatory Bodies	Governments and regulatory bodies play a crucial role in setting and supervising safety standards and regulations, as evidenced by IAEA's approach where establishing standardized safety standards and an independent supervisory system are regulatory imperatives.
Industry & Corporations	Companies developing, producing, and selling AI technologies and products must adhere to regulatory standards and bear the responsibility of providing safe products and services. This necessitates the establishment of independent audit and evaluation systems within each corporation.
Research Institutions & Academia	Engaged in research regarding technological advancements and associated risks, these entities provide scientific bases for regulatory directions and necessity. Cases like the Fukushima accident have underscored the importance of research institutions.
International Organizations	Entities like the IAEA offer platforms for international cooperation and standard setting. With AI development transcending national borders, the establishment of international safety standards and cooperative structures has become increasingly important.
Community & Civil Society	Engage in discussions and supervision focused on public safety. Just as IAEA's regulatory standards are periodically reviewed per the details of international action plans, the participation and feedback from civil society ensure the efficacy and appropriateness of regulations.

transparent products and services and need to establish internal audit and evaluation systems for this purpose.

Secondly, research institutions and academia continuously study the rapid development of technology and its associated risk factors. The role of academia and research institutions, as reconfirmed through the Fukushima accident, is crucial. They provide scientific evidence for the direction and necessity of regulation, significantly contributing to ensuring the appropriateness of regulation.

On the international level, organizations like the IAEA play a vital role. As the development of AI exerts international influence, the construction of international safety standards and cooperation systems becomes increasingly important. Hence, such international organizations provide a platform for international cooperation and standard-setting, minimizing the differences in regulations among countries.

Lastly, communities and civil society play a role as important entities for public safety. Continuous participation and feedback from civil society substantially contribute to ensuring the effectiveness and appropriateness of regulations.

Ultimately, cooperation and responsibility-sharing among various stakeholders are essential for the safe development of AI and the construction and implementation of regulatory systems. Lessons learned from the nuclear safety regulation case of the IAEA provide important guidelines for this process and will significantly contribute to constructing an effective AI regulatory system (Table 3).

Advantages and disadvantages of policies and exploration of alternatives

The AI safety regulation policy proposed by referencing the IAEA's nuclear safety regulation case has its own characteristic advantages and disadvantages. This section aims to explore these in depth and propose alternative approaches for improvement.

The regulatory system referenced from the IAEA's case is based on already verified processes and systems. Firstly, standardized safety standards set through international cooperation and agreement can minimize differences among countries and provide consistent regulatory standards. Secondly, active participation by various stakeholders enhances the effectiveness and appropriateness of regulation. Thirdly, continuous renewal of regulation through the study of technological development and associated risk factors allows for quick responses to realistic problems.

However, this approach has limitations and problems. Firstly, nuclear power and AI have different characteristics and risks, limiting the direct application of the IAEA's case in all aspects. Secondly, international agreement and standardization can be time-consuming and may not reflect the situations and needs of all countries. Thirdly, conflicts or redundancy with existing regulatory systems may occur.

This paper proposes alternative approaches to address these limitations and problems. Firstly, considering the characteristics and risks of AI, the establishment of an independent regulatory system can be considered. Through this, AI-specific regulations can be developed while referencing the IAEA's case. Secondly, it proposes a flexible regulatory model that accelerates the process of international agreement and standardization and reflects the situations of individual countries. Lastly, it explores ways to integrate and connect with national regulatory systems to harmonize with existing regulations.

Discussion: Insights and implications

This discussion revisits the objectives of the paper to reflect how our research findings contribute to understanding the applicability of the International Atomic Energy Agency (IAEA)'s regulatory framework to Artificial Intelligence (AI). The primary goal of this paper was to analyze how the IAEA's regulatory system could be leveraged for AI technology, particularly AI applications in the nuclear domain. The findings align with this goal and provide important insights into how the IAEA regulatory framework can accommodate the rapid advancement and integration of AI technology. These discoveries not only fulfill the paper's purpose but also make significant contributions to international discussions on AI regulation.

Our analysis offers important perspectives into how AI could be integrated within the IAEA's existing regulatory framework. The study explored concrete cases of how AI technology could be applied to the IAEA's nuclear safety and security regulation. Specifically, it analyzed how interactions could occur in areas like AI systems' risk assessment capabilities, accident prevention, and operational efficiency improvements. Such insights make important contributions to developing standards and procedures for the safe and responsible application of AI technology in the nuclear sector. Additionally, the analysis provides direction on how the IAEA regulatory framework needs to evolve to accommodate the rapid progress of AI technology.

These research results have broader implications, especially for the future of AI governance in an international context. Through the IAEA case, significant insights were gained regarding how AI regulation needs to be approached at an international level. Such analysis emphasizes the importance of cooperation and coordination in developing international AI regulatory frameworks. The global impacts of AI technology and ensuing regulatory challenges necessitate international responses, implying the need to promote inter-country collaboration and continual regulatory updates. This study makes a valuable contribution to exploring international approaches to AI regulation and underscores how international bodies need to strengthen cooperative efforts to ensure the safety and ethical use of AI technology.

The study highlights the challenges and opportunities that emerge in adapting the IAEA framework to the evolving AI landscape. One key challenge is the rapid pace of AI technological progress and the resulting need for regulatory flexibility. The current IAEA regulatory framework may not fully reflect the complexity and diversity of AI technology, necessitating regulatory renewal and innovation. Additionally, AI integration requires new thinking and creativity in regulatory approaches. In contrast, AI technology presents important opportunities in areas like improving nuclear safety, efficient risk management, and decision-making support. Such opportunities can maximize the positive changes AI can bring to the nuclear industry and facilitate the modernization of the IAEA regulatory framework. The study balances these challenges and opportunities in an unbiased manner, providing substantive direction on the effective application and evolution of AI regulation.

Based on our research findings, we propose some practical recommendations for policymakers and practitioners in AI technology regulation. First, flexible approaches tailored to the rapid pace of technological progress are needed in AI regulation. Policymakers need to continually update and adapt regulations according to the current rate of technological evolution. Second, developing and implementing clear standards and guidelines to ensure AI safety and ethical use is crucial. International bodies like the IAEA can play leading roles in this area. Third, promoting international cooperation and information sharing regarding AI integration is necessary. This allows diverse countries and organizations to share and learn best practices in AI regulation. Finally, proactively soliciting stakeholder participation and feedback in AI regulatory development is important. Such recommendations can enhance the efficacy of AI technology regulation and minimize associated risks.

This study emphasizes the need for continued adaptation in AI regulation and suggests directions for future research. By providing in-depth analysis of how the IAEA's regulatory framework could be applied to AI technology, it offered important perspectives on the role of AI in nuclear safety and security. Future research can focus on international coordination and cooperation in AI regulation, particularly on addressing the various technological, ethical, and societal issues. Additionally, empirical research on the continual evolution and adaptation of regulatory strategies alongside AI technological progress is warranted.

Conclusion

This study explored the necessity and direction of safety regulations for AI, based on the nuclear safety regulation case of the IAEA. Lessons learned from the international regulatory environment provide important insights for managing the global spread of AI and associated risks.

Firstly, standardized safety standards based on international cooperation can minimize differences among countries, enhancing the consistency and effectiveness of regulation. Secondly, participation by various stakeholders strengthens the appropriateness of regulation, and continuous renewal of regulation through the study of risk factors associated with technological development is crucial. However, due to the different characteristics and risks of nuclear power and AI, there are limitations to directly applying the IAEA's case.

In conclusion, while this study provides valuable insights into the applicability of the IAEA's regulatory framework to AI safety, it is crucial to acknowledge the limitations of directly adopting the nuclear regulatory model to AI. The distinct characteristics of AI technology, such as its rapid evolution, complexity, and broad societal impacts, present unique challenges that may not be fully captured by the current IAEA approach. Future research should focus on developing AI-specific regulatory strategies that address these challenges, strike a balance between innovation and safety, and foster public trust in AI systems. This will require ongoing collaboration between policymakers, researchers, and stakeholders to adapt and refine regulatory frameworks in light of the ever-evolving AI landscape. However, it is important to acknowledge the limitations of directly applying the IAEA's nuclear safety regulatory framework to AI. While the IAEA's approach provides valuable insights and lessons, the unique characteristics and risks associated with AI technology may require the development of a separate, AI-specific regulatory framework. The rapid pace of AI development, the complexity and opacity of AI systems, and the potential for unintended consequences pose distinct challenges that may not be fully addressed by the IAEA's existing regulatory structure. Furthermore, the societal and ethical implications of AI, such as privacy concerns, algorithmic bias, and job displacement, go beyond the scope of the IAEA's mandate, which is primarily focused on nuclear safety and security. Addressing these broader impacts of AI may require collaboration with other international organizations and the development of new, multidisciplinary approaches to AI governance. Future research should delve deeper into the need for an independent, AI-specific regulatory framework that can complement the lessons learned from the IAEA's approach. This framework should be adaptable, inclusive, and responsive to the unique challenges and opportunities presented by AI technology. Additionally, further research is needed to explore how the IAEA's regulatory framework can be harmonized with other emerging AI regulations and governance initiatives to ensure a coherent and effective international response to the challenges of AI safety and ethics. Future research needs to explore more deeply the necessity of an independent regulatory system based on the characteristics of AI. Additionally, research can be expanded in the direction of developing a flexible regulatory model that accelerates the process of international agreement and reflects the situations of individual countries. Exploration of ways to harmonize with existing regulations and research on securing the timeliness of regulation in response to the rapid development of AI technology are also required.

Data availability

The study reviewed a variety of regulatory documents and guidance, including a comprehensive review of IAEA regulations and AI technical standards. Data sharing does not apply to this study as no data is generated or analyzed during the process. The relevant data was used for research purposes only, and all data were anonymized and processed by the researcher.

Received: 6 October 2023; Accepted: 28 March 2024;
Published online: 12 April 2024

References

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. arXiv preprint arXiv:1606.06565

Anastassov A (2016) Some Aspects of the Effectiveness of the International Regulatory Framework to Ensure Nuclear Safety. Link. https://doi.org/10.1007/978-94-6265-138-8_7

Anderljung M, Barnhart J, Leung J, Korinek A, O'Keefe C, Whittlestone J, … Wolf K (2023) Frontier AI regulation: Managing emerging risks to public safety. arXiv preprint arXiv:2307.03718

Andalibi M, Setoodeh P, Mansourieh A, Asemani MH (2020) Multi-task Deep Reinforcement Learning: a Combination of Rainbow and DisTraL. In 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS) (pp. 1–6). IEEE

AI Risk Ontology (AIRO) (2022) An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards. Technical Report, AIRO Project

Barto AG, Mahadevan S (2003) Recent advances in hierarchical reinforcement learning. Discret Event Dyn Syst. 13(1–2):41–77

Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828

Chang T, Lin H, Lin C (2013) Constructing and evaluating online goal-setting mechanisms in web-based portfolio assessment system for facilitating self-regulated learning. Computers Educ. 69:237–249

Curtis G, Lacey D, Le T (2022) AI-deploying organizations are key to addressing 'perfect storm' of AI risks. AI & Ethics, 2(4):521–533

Dietterich TG (2000) Hierarchical reinforcement learning with the MAXQ value function decomposition. J Artif Intell Res 13:227–303

Dixon RBL (2022) A principled governance for emerging AI regimes: Lessons from China, the European Union, and the United States. AI and Ethics, 2(1–2):1–18

Drabiak K (2022) Leveraging law and ethics to promote safe and reliable AI/ML in healthcare. Front Nucl Med 2:983340

Eckersley P (2019) Impossibility and uncertainty theorems in AI value alignment. arXiv preprint arXiv:1901.00064

Etienne H (2022) When AI ethics goes astray: A case study of autonomous vehicles. Soc Sci Comput Rev 40(1):236–246

Fenwick M, Vermeulen EP, Corrales M (2018) Business and regulatory responses to artificial intelligence: Dynamic regulation, innovation ecosystems and the strategic management of disruptive technology. In Robotics, AI and the Future of Law (pp. 81–103). Singapore: Springer Singapore

Fox S (2018) Domesticating artificial intelligence: Expanding human self-expression through applications of artificial intelligence in prosumption. J Consum Cult 18(1):169–183

Garcia I (2023) Nuclear Energy Agency's Consensus Position on Regulatory Inspections of Digital Instrumentation and Control Systems and Components Important to Safety used at Nuclear Power Plants - Inspection Framework. Link. https://doi.org/10.13182/npichmit23-40530

Gourraud PA, Simon F (2020) Differences between Europe and the United States on AI/digital policy: comment response to roundtable discussion on AI. Gend Genome 4:2470289720907103

Gstrein OJ (2022) European AI Regulation: Brussels Effect versus Human Dignity? Zeitschrift für Europarechtliche Studien (ZEuS), 25(4):469–484

Hacker P, Engel A, Mauer M (2023) Regulating ChatGPT and other large generative AI models. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1112–1123)

Hinton G (2023) Godfather of artificial intelligence talks impact and potential of AI [Video]. CBS Saturday Morning Available at: https://www.cbsnews.com/video/godfather-of-artificial-intelligence-talks-impact-and-potentialof-new-ai/

Juozaitis J (2020) The (De) Legitimisation of Lithuanian opposition to ostrovets nuclear power plant through International Atomic Energy Agency. Politologija 100(4):106–152

Kim BS, Hong SH, Kim K (2022) Preliminary assessment of the safety factors in K-DEMO for fusion compatible regulatory framework. Scientific Reports, 12(1):8276

Kop M (2020) Shaping the Law of AI: Transatlantic Perspectives. In Shaping the Law of AI: Transatlantic Perspectives: Kop, Mauritz. [Sl]: SSRN

Kulkarni TD, Narasimhan K, Saeedi A, Tenenbaum J (2016) Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. Advances in neural information processing systems, 29

Kuznetsova EA, Fionov AN (2022) Regulatory framework for the formation of the system information security of a nuclear facility. Link. https://doi.org/10.33764/2618-981x-2022-6-116-121

Lieder F, Chen PZ, Stojcheski J, Consul S, & Pammer-Schindler V (2022) A Cautionary Tale About AI-Generated Goal Suggestions. In Proceedings of Mensch und Computer 2022 (pp. 354–359)

Mechergui M, Sreedharan S (2023) Goal Alignment: A Human-Aware Account of Value Alignment Problem. arXiv preprint arXiv:2302.00813

Middleton SE, Letouzé E, Hossaini A, Chapman A (2022) Trust, regulation, and human-in-the-loop AI: within the European region. Commun ACM 65(4):64–68

Mishra A, Gowrav MP, Balamuralidhara V, Reddy KS (2021) Health in digital world: A regulatory overview in United States. J Pharm Res Int 33(43B):438–450

Mitelut C, Smith B, Vamplew P (2023) Intent-aligned AI systems deplete human agency: the need for agency foundations research in AI safety. arXiv preprint arXiv:2305.19223

Mohamed M (2021) Enhancement operations management in supply chain based on intelligent support techniques: A case study. Am J Bus Oper Res 2(2):73–81. https://doi.org/10.34104/ajbor.021.073081

Nair S, Savarese S, Finn C (2020) Goal-aware prediction: Learning to model what matters. In International Conference on Machine Learning (pp. 7207–7219). PMLR

Nay J, Daily J (2022) Aligning Artificial Intelligence with Humans through Public Policy. arXiv preprint arXiv:2207.01497

Russell SJ (2010) Artificial intelligence a modern approach. Pearson Education, Inc

Shiri A, Kallakuri U, Rashid HA, Prakash B, Waytowich NR, Oates T, Mohsenin T (2022) E2hrl: An energy-efficient hardware accelerator for hierarchical deep reinforcement learning. ACM Trans Des Autom Electron Syst (TODAES) 27(5):1–19

Tulabandhula T, Vaya S, Dhar A (2017) Privacy-preserving targeted advertising. arXiv preprint arXiv:1710.03275

Turbék Z (2012) Global nuclear law in the making? Joint exercise of public powers in the nuclear field: the case of the revision of the International Basic Safety Standards. Nucl. L. Bull. 89:7

Valentini A, Fukushima Y, Contri P, Ono M, Sakai T, Thompson SC, Youngs RR (2021) Probabilistic fault displacement hazard assessment (PFDHA) for nuclear installations according to IAEA safety standards. Bull. Seismol Soc. Am. 111(5):2661–2672

Wang TR, Pradeep J, Chen JZ (2022) Objective driven portfolio construction using reinforcement learning. In Proceedings of the Third ACM International Conference on AI in Finance (pp. 264–272)

Wu H, Yu G, Teng K, Zheng X (2022) Analysis on the Nuclear Safety Supervision Mode of the World's Major Nuclear Power Countries and Its Enlightenment to the Improvement of China's Nuclear Safety Supervision Technical Support Ability. In International Conference on Nuclear Engineering (Vol. 86397, p. V005T05A016). American Society of Mechanical Engineers

Yampolskiy RV (2019) Unpredictability of AI. arXiv preprint arXiv:1905.13053

Yefremova KV (2020) Peculiarities of Application of Artificial Intelligence in the Financial Services Sector: EU Experience. L. & Innovative Soc'y, 61

Yuan L, Gao X, Zheng Z, Edmonds M, Wu YN, Rossano F, Zhu SC (2022) In situ bidirectional human-robot value alignment. Sci. Robot. 7(68):eabm4183

Zhang J, Mattie H, Shuaib H, Hensman T, Teo JT, Celi LA (2022) Addressing the "elephant in the room" of AI clinical decision support through organisation-level regulation. PLOS Digital Health 1(9):e0000111

Zhang Y, Agarwal P, Bhatnagar V, Balochian S, Yan J (2013) Swarm intelligence and its applications. Sci World J, 2013:528069

Acknowledgements

This work was supported by the Korea Institute of Science and Technology Information (K-23-L05-C02-S16).

Competing interests

The author declares no competing interests.

Ethical approval

This research was conducted in adherence to ethical standards and practices. The study was reviewed and approved by the KISTI. All research activities were performed in accordance with the guidelines and regulations pertinent to human subjects, such as the Declaration of Helsinki or a comparable ethical standard.

Informed consent

This study did not involve human participants, and thus informed consent was not applicable.

Additional information

Correspondence and requests for materials should be addressed to Seokki Cha.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.