Case Study: IAEA Safeguards

Table of Contents

Executive Summary

- IAEA safeguards are not an internationally uniforme programme, but a **collection of agreements between member states and the IAEA**
- A key weakness (its voluntary nature) has been somewhat mitigated by its **integration into other regional treaties** so that signing them is equivalent to agreeing to be part of the IAEA safeguards system
- The safeguards program is the IAEA's largest budget expenditure
- The IAEA safeguards are **built around resources and "material accountancy"** (rather than activities, procedures, safety measures etc.) and facility design
- A key concept is that of a **significant quantity**, the "approximate amount of nuclear material for which the possibility of manufacturing a nuclear explosive device cannot be excluded"
- The main functions it serves are **confidence-building, providing an early warning mechanism and being a trigger for international responses** (if necessary)
- Agreements with participating states vary based on their status (nuclear weapon vs. non-nuclear weapon states), with a third category that's restricted to specific items
- There is a **reduced regulatory burden for countries with small quantities** of monitored materials
- IAEA safeguards **preceded the NPT** and have been credited for facilitating its creation
- One of the most significant changes over time was one **away from facility-based monitoring towards a "full-scope" and country-based approach**
- This was controversial since it was in stark conflict with a key principle (and success factor!) of IAEA safeguards, their **non-discriminatory** (**and hence politically neutral**) **nature**, and the IAEA has been criticised for overreaching
- The second big shift was one from "correctness" to "completeness" - the IAEA started **not just looking at accuracy of self-reports but also whether everything was reported in the first place**, closing a key loophole
- Both changes were triggered by infamous "failures" (Iran, Iraq, Romania, DPK) - the trajectory of the IAEA can be described as a **progressive series of failures that have triggered stronger and more effective measures**
- Benchmarks matter: by the 1990 standards, Iraq was (in theory) 100% compliant
- Over time, the IAEA was able to add "Additional Protocols" to most agreements which gave it **additional oversight and verification powers**
- This addition also marked a shift from a "quantitative approach" to a more "qualitative approach" focused on a **comprehensive assessment of a state's nuclear capabilities**

- The IAEA works closely with the member states, which have to **establish an entity that serves as a contact point for the IAEA** and coordinates accountancy
- The IAEA provides ample support for member states such as technical assistance, training, implementation guidance etc.
- Leaders like the US signing the agreement can send a **strong sign** that agreeing to this kind of oversight does not mean losing out on economic competitiveness
- Implementation follows 4 steps: Data collection, development of a safeguard approach, verification and drawing final conclusions
- Verification measures include **on-site inspections, visits, ongoing monitoring and evaluation**
- Routine inspections are limited to so-called "strategic points" (locations within a nuclear facility or locations containing nuclear material) and **can't be conducted anywhere and everywhere**
- Safeguards focused on facilities are taking the **entire lifecycle** (design, construction, operation and decommissioning) into account
- **Inspection activities** include "auditing the facility´s accounting and operating records and comparing these records with the State´s accounting reports to the agency; verifying the nuclear material inventory and inventory changes; taking environmental samples; and applying containment and surveillance measures (e.g., seal application, installation of surveillance equipment)"
- The **intensity of verifications and inspections depends on the agreement** and whether an additional protocol has been agreed upon
- Through the use of visual technology, radiation detection and satellite imagery, monitoring is **continuously conducted even outside of recurring inspections**
- The IAEA has very limited powers in case of non-compliance (only symbolic) and has to **escalate things to external entities** (e.g. the UN Security Council) to affect sanctions, which makes the process prone to politicisation
- **No consensual definition of "non-compliance" exists**, leaving it to the IAEA to make a call whether a country is in violation of an agreement
- Even when "non-compliance" is established, this will sometimes have no consequences and will simply be reported
- Another challenge has been the **existence of considerable time lags** between the entry into a protocol and its actual implementation
- **Voluntary reporting schemes** can be used to "draw countries into" more binding agreements

1. Introduction

"If a country with a full nuclear fuel cycle decides to break away from its non-proliferation commitments, a nuclear weapon could be only months away. In such cases, we are only as secure as the outbreak of the next major crisis. In today's environment, this margin of security is simply untenable."

Dr Mohamed El Baradei *(former Director General of the IAEA)*

The question of nuclear nonproliferation is decades old, yet it remains one of the most important priorities to prevent disastrous conflict between great powers. At the same time, other existential risks such as climate change highlight the need for nuclear power plants as a source of clean energy, and for the utilisation of all technological and scientific progress we have at our disposal to advance our knowledge and provide energy to people around the globe. It is precisely this dual-use nature of nuclear energy on one hand and nuclear weapons on the other that makes the regulation of said technology so difficult, and that is the cause for worries such as the one cited above. How can it be assured that a country uses its nuclear materials exclusively for research and power generation, and not for the production of a weapon of mass destruction?

One possible answer to this question involves the safeguards by the International Atomic Energy Agency (IAEA) that this case study is examining. We will first describe in rough sketches what they are and how they work, before diving deeper into the fascinating history that has birthed and subsequently shaped the development of the IAEA safeguards. After briefly outlining the importance of member state cooperation, this case study will provide an overview of how IAEA safeguards are implemented and enforced, while also highlighting a few snapshots of previous cases of non-compliance. Finally, we will evaluate the effectiveness of IAEA safeguards and challenges this international assurance regime is facing, before summarising the key lessons that can be learned for another powerful dual-use technology: Artificial Intelligence.

2. Overview: How IAEA safeguards work

Let's begin with a seemingly simple question: What are safeguards?

As per the IAEA itself, "[s]afeguards are activities by which the IAEA can verify that a State is living up to its international commitments not to use nuclear programmes for nuclear-weapons purposes" (Source). This definition also alludes to the larger goal that is pursued:

Objective

> "The objective of IAEA Safeguards is to deter the spread of nuclear weapons by the early detection of the misuse of nuclear material or technology. This provides credible assurances that States are honouring their legal obligations that nuclear material is being used only for peaceful purposes." (Source)

The main function can therefore be described as two-fold (Source): By providing "credible assurances", the IAEA effectively acts to build confidence on an international level. At the same time, the IAEA [which has also been described as the "nuclear watchdog" (Source)] uses safeguards as an early-warning mechanism which can serve to trigger international responses to violations and breaches of agreements if necessary.

Legal framework

A variety of elements form the legal basis for IAEA safeguards (Source). Apart from the IAEA Statute and the decisions of the IAEA Board of Governors, consisting of 35 members from different member states, elected by the attendants of the annual General Conference, the IAEA safeguards draw most of their legal legitimacy from obligations that states have agreed to under the "Treaty on the Non-Proliferation of Nuclear Weapons" (NPT), as well as additional treaties that have established nuclear-weapon-free zones (Source). It's important to note that the IAEA it itself not a party to the NPT; rather, it has been entrusted with the verification of responsibilities that parties to the NPT have agreed to subject themselves to under Article III, which requires signatories of the treaty to conclude safeguard agreements with the IAEA to guarantee the fulfilment of their obligations (Source). While formally headquartered in Vienna, the IAEA maintains regional safeguards offices in Tokyo and Toronto (Source), and dedicates a significant portion of their annual budget to the safeguards program - 39% in 2020, making this the largest expense of the IAEA and more than the budget for Nuclear Energy, Nuclear Safety and Technical Cooperation combined (Source).

Material

Crucially, the safeguards program is organised around materials and resources as the key variable that determines the level of required oversight and monitoring: "The IAEA uses nuclear material accountancy as its basic measure for safeguarding declared material. The

system monitors the quantities of nuclear material present in a nuclear facility and the changes in these quantities that take place over time." (Source) The most critical elements that the IAEA is focused on are enriched uranium, plutonium and uranium-233, all of which could be used for the development of a nuclear weapon. "It also includes natural uranium and depleted uranium, the latter of which is commonly used, for instance, as shielding for radiation sources in hospitals, industry and agriculture. Radioactive sources that do not contain nuclear material are not subject to safeguards and need not be reported to the IAEA under a safeguards agreement." (Source) This means that safeguards are not centred around the degree of radioactivity of a given material, but rather its potential usability for weapon development.

An additional idea that is central to the IAEA safeguards approach is that of a so-called *significant quantity* - the "approximate amount of nuclear material for which the possibility of manufacturing a nuclear explosive device cannot be excluded." (Source) These amounts act as thresholds to determine the degree of oversight. Similarly, on a country-wide level, exceptions can be made for parties that have no or only negligible amounts of critical materials, allowing for reduced regulatory burden and reduced cost: "A State with little or no nuclear material may be eligible to conclude a small quantities protocol (SQP) [...], which reduces the safeguards activities conducted in the State" (Source).

Agreements

The specifics of how much of what material a given country possesses and uses for different purposes is detailed in so-called safeguard agreements, which can further be subdivided into different categories (see *Table 1*). Three main categories exist: Comprehensive safeguard agreements (CSAs) are made with non-nuclear-weapon states and make up the majority of the agreements that exist. All nuclear weapon state parties have agreed to sign Voluntary offer safeguard agreements (VOAs), and for India, Pakistan and Israel, item-specific agreements have been negotiated.

Table 1: Types of safeguard agreements (Source, Source):

Type	Partners	IAEA's mandate
Comprehensive safeguard agreements (CSA)	Non-nuclear-weapon states parties to the NPT and nuclear-weapon-free zone treaties (~178 states)	"Under a comprehensive safeguards agreement, the IAEA has the right and obligation to ensure that safeguards are applied on all nuclear material in the territory, jurisdiction or control of the State for the exclusive purpose of verifying that such material is not diverted to nuclear weapons or other nuclear explosive devices."

Voluntary offer safeguard agreements (VOA)	Nuclear weapon state parties (China, France, Russia, UK, US)	"The IAEA applies safeguards under a voluntary offer agreement to verify that nuclear material remains in peaceful activities and is not withdrawn from safeguards except as provided for in the agreement."
Item-specific agreements	India, Pakistan, Israel	"Under these agreements, the IAEA applies safeguards to ensure that nuclear material, facilities and other items specified under the safeguards agreement are not used for the manufacture of any nuclear weapon or to further any military purpose, and that such items are used exclusively for peaceful purposes and not for the manufacture of any nuclear explosive device."
+ **Model Additional Protocol**		

Additional Protocol

An additional layer to this model are so-called Additional Protocols (AP), which have been added in the late 20th century (see History). If parties agree to the addendum of an AP to a CSA, they have to abide by all counts specified in it, while members with VOAs or specific-item agreements can agree to some but not all requirements. Generally speaking, the AP furnishes the IAEA with additional powers, most importantly an extension to the verification mechanisms commonly employed (see Enforcement). While APs are of a voluntary nature, they have been signed by over 120 countries and are seen as the "gold standard". *Figure 1* gives a cursory overview over the additional coverage embedded in APs (Source).

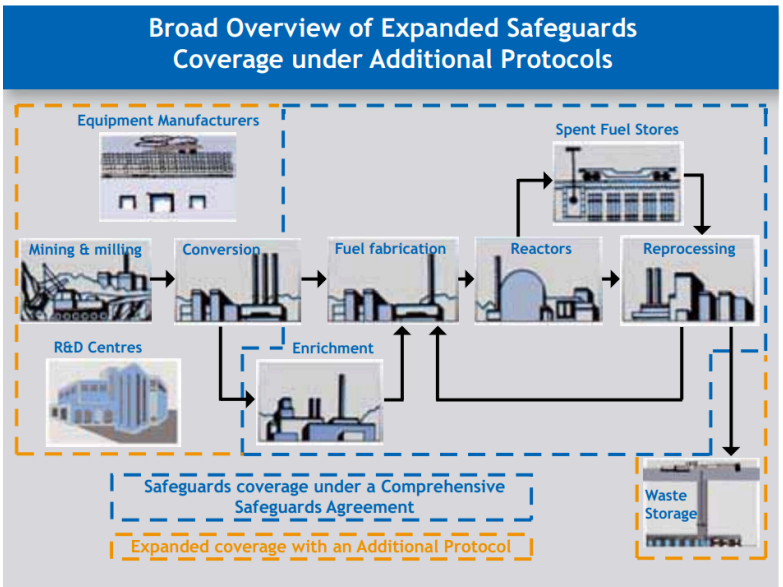

Figure 1: Expanded safeguards coverage under Additional Protocols (Source).

Process

The process of developing and monitoring safeguards approaches for each member state can be divided into four main stages (see *Figure 2*), although it's important to understand this process as cyclical rather than linear (Source, Source).

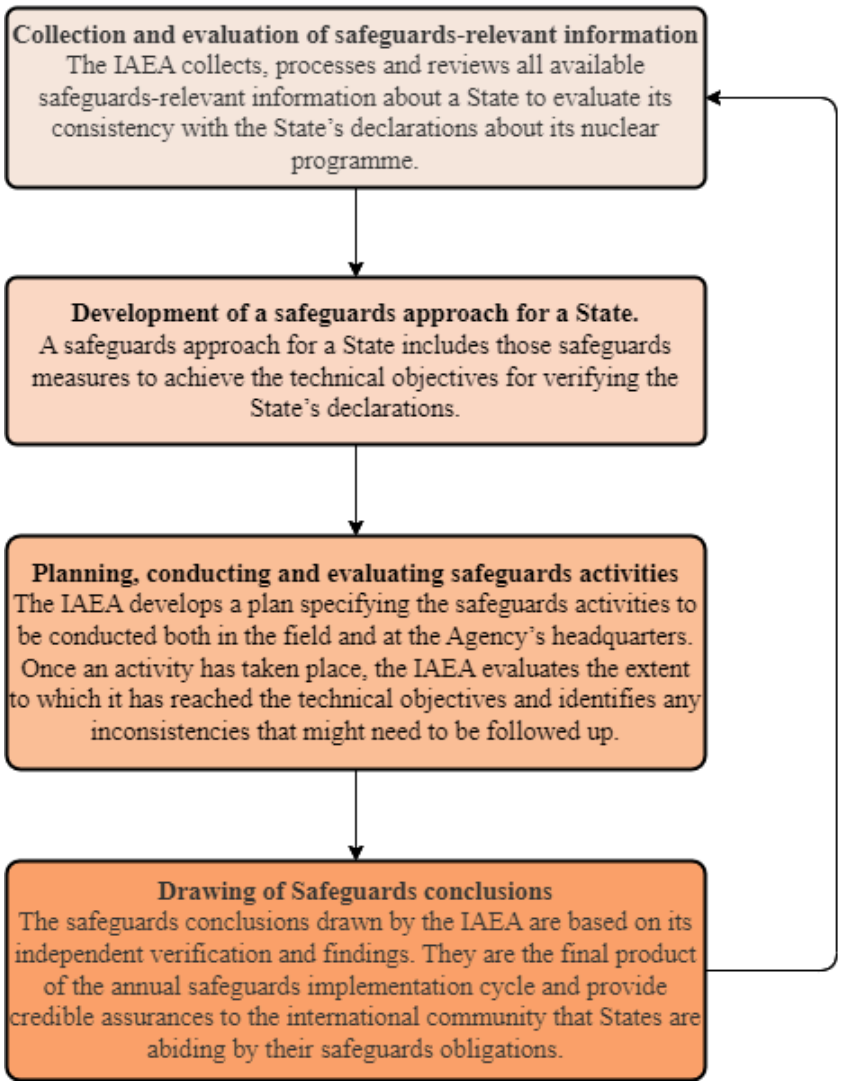

Figure 2: Overview of the IAEA safeguard process

As stated above, this process will be tailored to each individual member state, which will be given a "state-level safeguard approach: (SLA) and an annual implementation plan (AIP). More on the historical legacy of this approach can be found in the History section, and more details on how this gets put into practice can be found in the Implementation section.

3. History

While the history of IAEA safeguards is inseparably entangled with the history of the organisation that is housing this program, this case study is primarily concerned with how *safeguards* were created and how they developed over time. Therefore, covering the history of the IAEA is considered out of scope. Interested readers may want to look at this book-length history of the IAEA from the 1950s to the 1990s.

Figure 3 shows a high-level overview of key events between 1959 and 2005. The following sections will cover these and events after 2005 in greater depth.

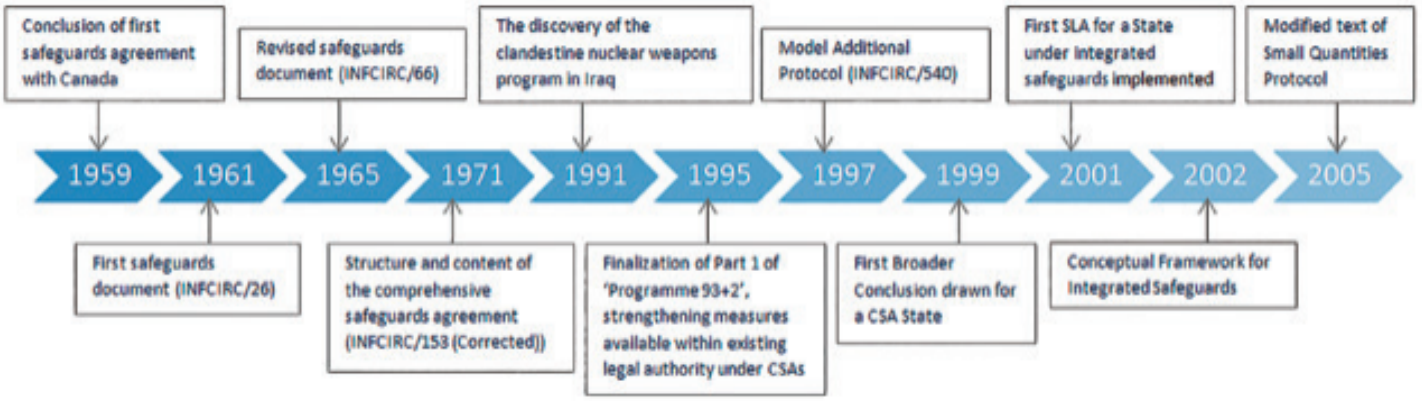

Figure 3: Key events in the IAEA safeguards history from 1959 to 2005 (Source)

3.1 Early days (1950s to 1970s)

"My country wants to be constructive, not destructive. It wants agreements, not wars, among nations. It wants itself to live in freedom and in the confidence that the peoples of every other nation enjoy equally the right of choosing their own way of life."

Dwight D. Eisenhower, 34th president of the USA (1953)

First steps

The story of nuclear safeguards begins in 1953, when Dwight Eisenhower first called for the establishment of an international atomic energy agency in his famous "Atoms for peace" speech (Source). 4 years later, the IAEA was established, originally "intended to be a kind of broker for controlled nuclear assistance and trade" (Source). The first application of safeguards was formalised in 1959, 2 years before a principled approach was laid out in official documents. It focused on uranium trade between Canada and Japan (Source). In 1961, the safeguards system was officially approved by the IAEA Board of Governors, and only one year later the first in-field verification of research reactors was conducted

(Source, Source). At the time, requirements were restricted to "reactors with less than 100 MW(th) output, to the source material and special fissionable material used and produced in these reactors, and to small research and development facilities" (Source). An amendment was suggested in 1964 and approved in 1965 which extended coverage to reactors of all sizes. Over the following 3 years, further additions and extensions were added to also cover fuel fabrication plants and facilities for reprocessing and conversion (Source). Initially, these agreements were restricted to a few countries, but interest quickly grew:

> "Throughout the 1960s, more and more countries began to request the IAEA to apply safeguards to nuclear material and facilities which they received under bilateral nuclear cooperation agreements. Those countries concluded with the IAEA item-specific safeguards agreements [...], also known as INFCIRC/66- type agreements, under which the IAEA applies safeguards to items subject to the agreements to verify that such items are used only for peaceful purposes." (Source)

The natural next step was further extension and institutionalisation, and the 1970s offered just the right environment for this.

3.2 Embedding safeguards into nonproliferation treaties

Treaties

In 1970, the non-proliferation treaty (NPT) entered into force, providing an opportunity to embed safeguards into a more institutionalised approach. While the IAEA was not a party to the Treaty itself, under Article III members of the NPT had to submit themselves to IAEA safeguards, signing so-called Comprehensive Safeguards Agreements (CSA). This way, the bilateral nature of safeguard agreements could be maintained while rapidly expanding the scope of countries that were covered by the safeguards. The integration into the NPT also marked a change in strategy for the IAEA safeguards:

> "It introduced a major change from the previous, item-specific safeguards system—which applied only to specified nuclear materials, facilities, or items, recognizing the legitimate rights of a state to possess other nuclear material outside of safeguards—to full-scope safeguards, which cover all peaceful nuclear material and activities in a state having a safeguards agreement with the IAEA in connection with NPT. This change did not influence the basic safeguards approaches very much, and the transition to full-scope safeguards, or comprehensive safeguards, occurred quite successfully." (Source)

As future chapters will show, this was just a first step towards a more all-encompassing approach, but it was an important force in paving the way and assuring buy-in from member countries for more comprehensive systems. The first CSA (concluded between

Finland and the IAEA) entered into force two years after safeguards were implemented into the NPT (Source). Less famously, the Tlatelolco Treaty - a non-proliferation treaty with a regional focus on Latin America and the Caribbean - also required signatories to "negotiate multilateral or bilateral agreements with the International Atomic Energy Agency for the application of its safeguards to its nuclear activities" (Source). In fact, this treaty opened for signatures a year before the NPT did and has sometimes been credited as a "trailblazer for non-proliferation" (Source).

3.3 Learning from failure: Iraq, Romania, North Korea

"It's fine to celebrate success but it's more important to heed the lessons of failure."

Bill Gates

The problem

The 1990s were a formative time for the IAEA safeguards, during which they encountered some of their biggest challenges but consequently underwent some of their most transformative evolutions towards more effective safeguarding of nuclear materials.

The most prominent incidents at the time included:

- 1993: *North Korea* was found to have extracted plutonium clandestinely (Source, Source).
- 1991: *South Africa* dismantles its nuclear weapons programme. While often seen as a success story of non-proliferation, this presented a challenge to the IAEA since it had to verify the complete termination, a process that had no precedent and which revealed some of the shortcomings of current IAEA processes and procedures (Source).
- 1992: *Romania* declared after their revolution to have previously separated 150mg of plutonium in one of their reactors (Source).
- 1991: *Iraq* was found to be pursuing a secret nuclear weapons program[1] after the First Gulf War, in stark violation of their CSA with the IAEA as an NPT party (Source, Source).

The Iraq case was the most incisive of the three, prompting an IAEA investigation into what had gone wrong. What had happened was that IAEA safeguards were reliant upon the declaration of relevant facilities - any nuclear programs that had no obvious links to these declared facilities were able to "fly under the radar" and go undetected for long

[1] This was later destroyed by a special UN team (Source).

periods of time (Source). This also highlighted a shortcoming of the design of CSAs, which were focused exclusively on declared facilities, allowing Iraq to exploit this loophole (Source). This observation "provided the catalyst for strengthening IAEA safeguards" (Source).

The solution

Faced with these (sometimes blatant) violations and contradictions of what the IAEA safeguards and the NPT set out to achieve, the IAEA developed plans to combat non-compliance and update its strategy to reflect the increasing need to reliably verify commitments that member states had made.

Several proposals were developed in 1992, until it was decided that "Program 93+2" was how the IAEA was going to assure it had increased access to information about facilities to prevent future failures similar to the ones mentioned above (Source). Implementing this program took more time than expected and was done in two stages: First (in 1996), new verification mechanisms and techniques were established, including "new monitoring measures, such as environmental sampling, no-notice inspections at key measurement points within declared facilities, and remote monitoring and analysis" (Source, Source). All of these measures were intended to solve the problem of detecting undeclared nuclear material and its potential misuse. In 1997, the second step was completed when the "Model Additional Protocol" was approved by the Board of Governors (Source). Unlike the previously mentioned additions and reinforcements that could be implemented as part of existing agreements, the Additional Protocol served as a new legal authority on which the IAEA's mandate is based (Source):

> "An additional protocol (AP) enables the IAEA to obtain a much fuller picture of a state's current and planned nuclear activities, nuclear material holdings, and nuclear-related manufacture and trade. The AP increases the IAEA's ability to provide greater assurance of the absence of undeclared nuclear material and activities in a state. Unlike the transition to comprehensive safeguards in the 1970s, the introduction of APs required the IAEA to institute substantial changes in safeguards implementation practice and to integrate existing CSA-based procedures with procedures described in AP." (Source)

In summary, the Additional Protocol expanded the IAEA's mandate significantly by allowing it to visit any facility regardless of its declaration status. (A fuller account of additional verification measures under the AP can be found in the chapter on implementation.) While Additional Protocols are not mandatory for NPT members, they are seen as a best practice and recommended by the IAEA (Source, see also Source). To date, 140 states have voluntarily signed onto these protocols (Source).

Another addition was "Modified Code 3.1", which made it obligatory for countries to make transparent immediately the design of any new nuclear facilities that a state is planning to build. Previously, under the code introduced in 1976, states had to submit similar information only 180 days *before nuclear material was going to be introduced*. This gave the IAEA more time to respond and to adjust the safeguard agreements if necessary (Source). It also gave the IAEA the authority to verify design information over the lifecycle of a given facility, from its planning to its construction, operation and decommissioning (Source).

Paradigm shift I

These changes were more than small adjustments or tweaks to an existing system. Rather, they marked the beginning of what can be described as a paradigmatic shift in the IAEA's safeguards approach:

> "[T]he Board of Governors determined that under such agreements, the IAEA has the right and obligation not only to verify that State declarations of nuclear material subject to safeguards are 'correct', i.e. they accurately describe the type(s) and quantity(ies) of the State's declared nuclear material holdings, but that they are also 'complete', i.e. they include everything that should have been declared. This determination was a major catalyst for efforts to equip the safeguards system with important new tools to verify 'completeness'." (Source, see also (Source)

The change from a focus on "correctness" to one that considers "completeness" can be considered a milestone for safeguards advancement (Source) and it played a pivotal role in determining the IAEA's strategy after 1991, which was confirmed when the IAEA discovered the potential of incomplete rather than incorrect reporting of nuclear materials in North Korea in 1993.

Others have described the paradigmatic shift that occurred in the 1990s as one "from a quantitative system focused on accounting for known quantities of materials and monitoring declared activities to a qualitative system aimed at gathering a comprehensive picture of a state's nuclear and nuclear-related activities, including all nuclear-related imports and exports" (Source).

3.4 State-level concepts and the 2000s

Paradigm shift II

Closely connected with the learnings from the early 1990s was a second strategic shift - that to a "state-level approach". This was inspired both by the realisation that facility-based safeguards were very prone to the failures associated with the "correctness" approach and the experiences of the IAEA in the dismantlement of the South African

nuclear arms program. During this process it became obvious to the IAEA that close cooperation with member states on a national level would be crucial in avoiding the mistakes of the past (Source). Compressing these insights, the IAEA launched so-called "state-level concepts" (SLC). While this term was officially used only in the mid-2000s, the IAEA had started working with a similar approach back in the late 1990s[2] (Source, Source). The first "state-level safeguards approach" (SLA) was implemented in Australia in 2001 (Source). SLAs can be described as "customized approaches" (Source) tailored to a given state. Thanks to the greater oversight authority of the Additional Protocol, the IAEA was able to produce state evaluation reports, and started drawing so-called "Broader Conclusions" for entire states (Source).

> "This involved greater consideration of the State's nuclear fuel cycle as a whole (as opposed to IAEA safeguards primarily focused on nuclear material at declared facilities and locations outside facilities (LOFs)) for the purpose of ensuring that the IAEA is able to exercise its right and fulfil its obligation to ensure that IAEA safeguards are applied on all nuclear material in all peaceful nuclear activities in a State with a CSA" (Source)

This major shift from a facility-based approach to a state-level system "can rightly be characterized as a revolution, rather than an evolution" (Source) and marks an important milestone in the IAEA safeguards history. It also helped the IAEA to expand further, covering 145 countries by the year 2001 (Source). At the same time, new surveillance technologies such as the use of satellite imagery were integrated into the IAEA verification and monitoring repertoire.

Nonetheless, even in these years of steady progress and seemingly successful safeguarding, a few setbacks need to be noted. In 2003, Libya disclosed that they had engaged in research related to nuclear weapons that was previously undeclared, which highlighted "illicit supply networks" (Source) as a source of leakage that weakened the effectiveness of the IAEA safeguards. In the same year, a different country made the headlines: "In 2003, information came to light regarding previously undeclared nuclear material and activities that the Islamic Republic of Iran (Iran) should have declared but had not declared to the IAEA." (Source) The violations reported were grave enough to merit a transmission of a report on the Iran case to the UN security council in 2006 (Source).

Small quantities

Another loophole could (for all we know) be closed before any attempts of evasion were identified: that of the so-called "small quantities protocol". As described in an earlier

[2] A longer history of these developments can be found here.

chapter, these protocols exist for countries with little or no nuclear materials at their disposal and are meant to reduce their regulatory burden. In 2005, this perceived weakness was addressed and this protocol was revised to contain fewer exemptions (Source, see also Source):

> "The revised SQP reduces the number of provisions in Part II of the CSA the implementation of which is held in abeyance as long as the State meets the eligibility criteria. The eligibility criteria for an SQP based on the revised standard text are that the State has nuclear material in amounts less than those specified in para. 37 of [153] and no decision has been taken by the State to construct or authorize construction of a facility as defined in [153]. An SQP based on the revised standard text is unavailable to a State with a planned or existing facility. States with revised SQPs are required to provide to the IAEA an initial report on all nuclear material subject to IAEA safeguards, an annual report on imports and exports of nuclear material, and early design information; to accept the designation of IAEA inspectors; and to allow access for inspections." (Source)

All of the IAEA's efforts to continuously update and strengthen its mechanisms were finally rewarded in the same year (2005), when the IAEA received the Nobel Peace Prize for its contribution to nuclear non-proliferation.

3.5 The 2010s - Now

Pushback

Not all of the IAEA's reforms were met with excitement, however. Particularly the state-level approach became the topic of heated discussions between member states, since it opened the door to a potential politicisation of the previously rather un-politicised process of nuclear safeguarding:

> "In 2012, several Member States questioned what they saw as IAEA management shortcomings and a lack of communication and transparency with Member States in this process. Their concern was that the concept was not properly described and that the IAEA Secretariat had not sought the Board of Governors' formal approval of this concept before implementing it. Additionally, some Member States were concerned that the practical implementation of the SLC could potentially result in subjective and politically motivated safeguards conclusions, which could jeopardize the stability of IAEA safeguards implementation." (Source)

Fortunately, these fears could be hushed by the IAEA Board of Governors and the General Conference, which committed to investigating and addressing these concerns. This resulted in a resolution which stated clearly that all state-level concepts needed to conform with existing agreements and that the states needed to be involved in the implementation of the state-level assessments, pacifying those who were concerned with

the potential effects of a state-based approach (Source). Even then, progress was slow, and in 2014 only 53 states had the "integrated safeguards" that were the new gold standard inspired by the state-level concept (Source). The IAEA also engaged in several projects trying to strengthen their oversight and building confidence, for example when they defined performance targets for SLAs in 2019 (Source). These initiatives to strengthen the "demand" side of the equation were met with attempts to improve the "supply" side as well. For instance, the IAEA launched a capacity building initiative called COMPASS in 2020 which was targeted at the respective national authorities that were responsible for implementing the safeguards agreements for their countries and sought to provide support and assistance to these entities (Source).

Steady growth

Over the years and thanks to this proactive and supportive approach, the IAEA was able to steadily register new participants to the safeguard agreements, and in 2016, 182 states had safeguards agreement in place (Source), a number that further rose to 189 in 2022 (Source). We turn to the role of member countries and their cooperation with the IAEA in the next chapter.

4. Cooperation with member countries

As outlined above, the IAEA safeguards program was able to continuously attract new partners and extend its coverage across the globe. While some countries withdrew and then re-joined the programme, the overall trend is one of expansion and integration across continents.

An outlier

The one notable exception to this rule is North Korea. In April 2009, the North Korean government decided to end all cooperation with the IAEA, which resulted in a somewhat hasty departure of all IAEA inspectors from North Korean facilities, including the removal of safeguards equipment (Source). To this day, there is no operational safeguards agreement with the country, and efforts to reach out to North Korean decision makers have not proven fruitful.

SSACs

For all other members that have CSAs, VOAs or item-specific agreements, the IAEA emphasises close collaboration with the respective entities on the ground, acknowledging its limited ability to conduct all necessary activities on the ground itself. The most important mechanism by which this coordination and cooperation is ensured in practice are so-called "State systems of accounting for control of nuclear material" or SSACs. The European Union and Argentina-Brazil have a regional variety of this called an RSAC

(Source). Under comprehensive safeguards agreements (CSAs), countries are obligated to install these accounting systems as a way of centralising recording, reporting, controls, legislation and accounting for nuclear materials. SSACs are technically not required for countries with item-specific agreements, however in practice these countries also need to prove an organisational arrangement that acts as an intermediary between the state and the IAEA with a duty to report things accurately in a centralised manner, making some version of an SSAC de facto necessary. SSACs are entrusted with a wide range of tasks:

> "They are responsible for submitting design information to the IAEA, making sure that nuclear facility operators maintain the records that the IAEA requires and providing for IAEA inspectors to gain physical access to facilities and other locations. They must also make sure that nuclear plant operators are able to measure quantities and types of nuclear material precisely and accurately and that their equipment and measuring systems meet the highest international standards. SSACs can also help to resolve any problems that arise during in-field verification activities. For States with additional protocols in force, it is generally the SSAC that is entrusted with ensuring that the IAEA receives the additional information and access that this entails." (Source)

On top of these accounting and reporting responsibilities, SSACs serve an additional purpose, which is to act as "the chief point of contact between the State and the IAEA for operational issues" (Source). In the case of a need for adjustments or additional inspections, the IAEA relies on the SSAC to be their point of contact and representative on the ground. On the other hand, SSACs can count on the IAEA to provide support to assure technical proficiency and adequacy of their respective systems. To hold up its end of the bargain, the IAEA has introduced an advisory service called "ISSAS" which assists member states and their SSACs with advice and recommendations, highlighting the importance of close links between the IAEA and member states (Source, Source). Through the ISSAS program, small expert teams are sent to member states to better understand the conditions on the ground and to help them improve the efficacy of their respective SSACs (Source).

This level of support is also reflected in other programmes and initiatives the IAEA engages in to provide member states with the necessary expertise and the resources needed to effectively put IAEA safeguards into practice. In service of this goal, the IAEA issues "safeguards implementation practices guide" to share "best practices" and "lessons learned", although it's important to note that their application is voluntary and can not be mandated (Source). They publish recommendations on how to set up a state authority responsible for implementation and annually circulate a glossary to establish a shared terminology (Source). They also offer technical cooperation in the form of funding, expertise, training and equipment.

This cooperative approach has made it possible for the IAEA to move away from the original image of a "watchdog" that is to be feared and towards the role of a supportive partner that provides assistance to those who need it. This has incentivised more and more countries to sign onto to safeguards agreements, with some of them explicitly mentioning the role-model and standard-setting function of safeguards as a key incentive to join: "The U.S. implements the U.S.– IAEA Safeguards Agreement to demonstrate to other countries that the implementation of IAEA safeguards will not place their nuclear facilities at an economic disadvantage compared to the U.S. nuclear facilities." (Source)

5. Implementation

Earlier sections of this report have already touched on the high-level circular process the IAEA employs to determine safeguards needs and verify compliance. This section aims to elaborate and elucidate the steps of this process in greater detail, with a particular focus on Enforcement and Compliance.

Four distinct steps can be identified (see *Figure 4*): I) *Information collection and evaluation*, II) *Development of a safeguard approach*, III) *Verification and other activities* and IV) *Drawing safeguards conclusions* (Source). We will take each of these in turn.

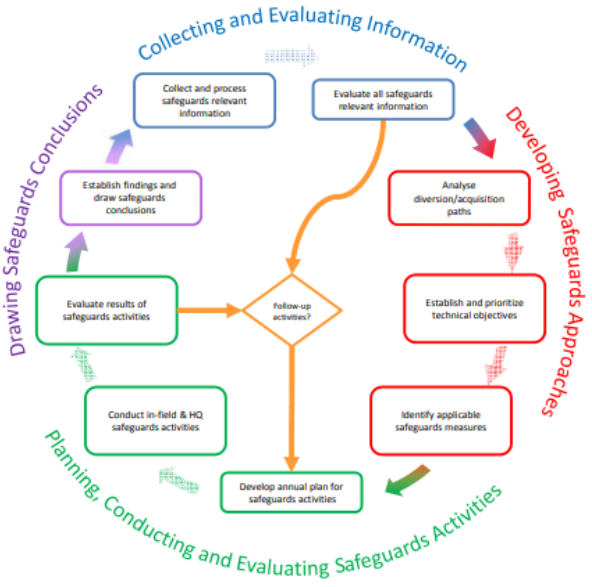

Figure 4: IAEA safeguards process (high-level overview) (Source)

Four steps

I. Information collection and evaluation

In order to determine the necessity and degree of oversight by the IAEA, information needs to first be collected on the current state of nuclear material use in a given country. While collecting this information and evaluating the requirements for safeguards, the IAEA can draw from 3 main sources (Source):

- Information provided by the state itself (in the form of reports or declarations)
- Activities conducted by the IAEA, both in their HQ and in-field
- Other relevant information that is either open-sourced or surfaced by third parties (this also includes methods like the use of satellite imagery)

When evaluating this information, the IAEA is particularly paying attention to potential incongruencies between different sources of information. The degree to which different sources of information are used and weighted respectively will vary from case to case.

II. Development of a safeguard approach

Based on the findings from stage I, the IAEA establishes technical objectives in line with the information it collected. For every state, a state-level safeguard approach (SLA) is crafted, based on plausible pathways through which a country could acquire the nuclear materials necessary to construct a nuclear weapon (Source). Increasingly, states are involved in this process.

III. Verification and other activities

Step 3 describes what is mostly associated with the work that the IAEA is doing - verification and monitoring. Making use of a plethora of different methods and tools, the IAEA checks whether states comply with the safeguard approach that has been developed for and with them in step 2. Further details on how this is done in practice and what devices are available to inspectors to fulfil this obligation can be found in the following section.

IV. Drawing safeguards conclusion

Safeguards conclusions are the closest thing to a "product" or "tangible output" that the IAEA has, and they are published in an annual report. Depending on the type of safeguards agreements a state has with the IAEA (*CSA with AP, CSA without AP, VOA or item-specific*), the type of conclusion issued by the IAEA will vary (see *Figure 5*).

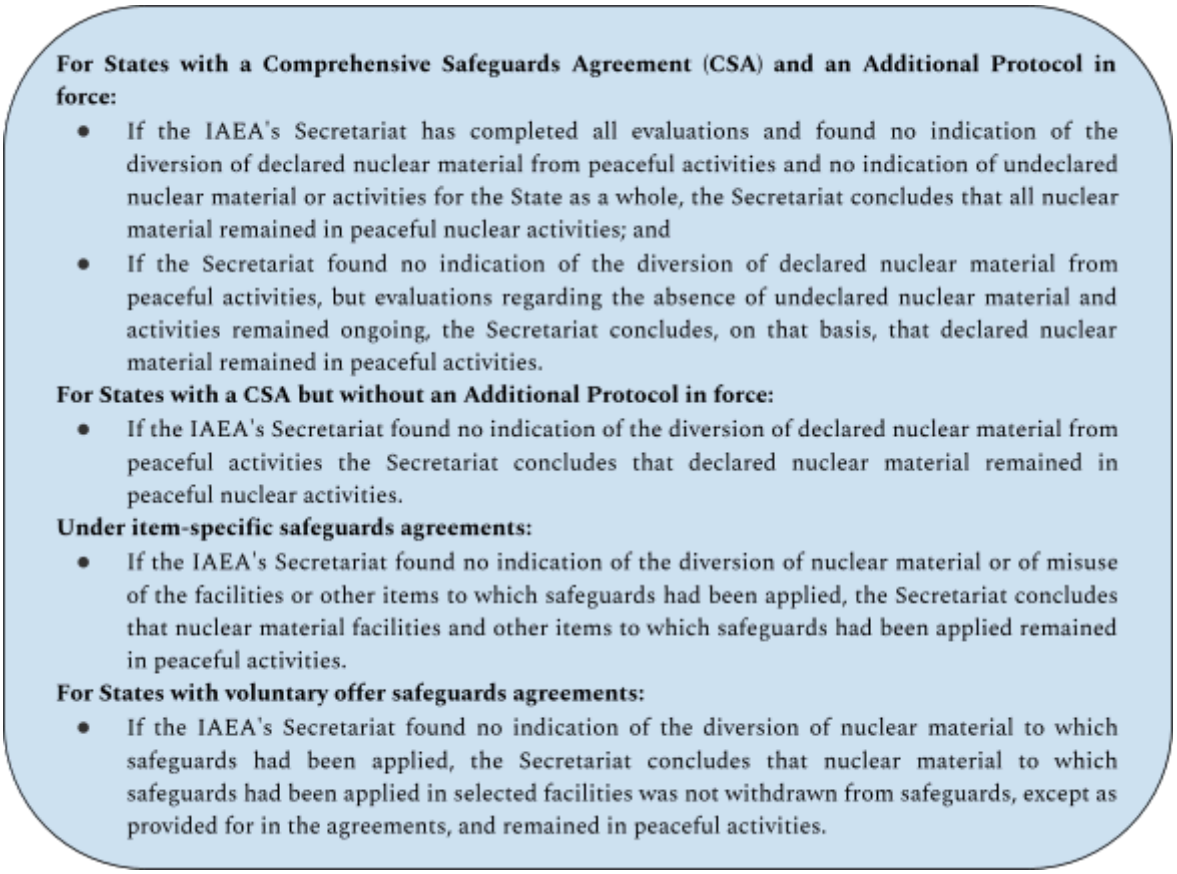

Figure 5: Possible safeguards conclusions under different agreements (Source)

5.1 Enforcement

We will now zoom in on a particular part of this process, which is the verification and monitoring. The IAEA employs more than 800 inspectors (Source), all of whom receive special training for the job, highlighting the emphasis the IAEA puts on assuring that state's efforts are controlled adequately. An important point to note is that inspectors look for signs of potential misuse but do NOT investigate matters of nuclear safety generally, since this is out of scope for the IAEA safeguards system.

Verification measures

Key verification measures include on-site inspections, visits and ongoing monitoring and evaluation (Source). In practice this means that inspectors are tasked with "auditing the facility´s accounting and operating records and comparing these records with the State´s accounting reports to the agency; verifying the nuclear material inventory and inventory changes; taking environmental samples; and applying containment and surveillance measures (e.g., seal application, installation of surveillance equipment)" (Source, see also Source). This has often been compared to a procedure in the financial system, the audit of

a bank: "the inspector compares the information stated on the nuclear material accounting records, books and reports of a facility with what has been reported by the State to the IAEA and, crucially, verifies that the nuclear material is actually present at the facility as declared" (Source). Measurement techniques for nuclear material can be differentiated into two categories: non-destructive and destructive. Non-destructive methods like counting, weighing or measurements with radiation detectors allow for verification without the need to alter the investigated item. Destructive assays, on the other hand, require that something is taken from a facility and destroyed in the process of measurement, and is hence only employed when other methods of verification are deemed inconclusive or insufficient (Source).

Similarly, when verifying the design of a facility, inspectors focus on whether what the state has submitted matches onto the observations they make during visits (both with regard to accuracy as well as completeness to rule out the possibility of misuse (Source). Additionally, inspectors may take environmental samples to verify use-as-intended: "These samples allow an analysis of traces of materials that can reveal information about nuclear material (for example, separated plutonium or highly-enriched uranium at a facility) or activities that have not been declared to the IAEA" (Source). These samples are often sent to a laboratory in Europe to be analysed.

A third category of measures could be classified as "containment and surveillance techniques". Apart from cameras and detectors, this includes the use of seals[3] to create additional transparency and accountability, sometimes also referred to as "knowledge continuity" (Source). Measures of this kind also allow for the monitoring process to be less concentrated around physical visits and to allow verification to be an ongoing process. Satellite imagery is another example of a verification method that does not require the physical presence of inspectors (Source). More and more, innovative technologies like robots and automated systems are able to take over some of these verification and surveillance tasks (Source, Source), increasing the capacity of the IAEA.

[3] "[O]ne of the main tools is sealing. The new passive seal comes in the form of a device that is no larger than a coin. With it, an IAEA safeguards inspector can seal a container, a hatch or a nuclear material cask, and the inspector can return at a later date to verify whether it was opened. A passive seal is one way to ensure the continuity of knowledge for nuclear material – if the seal is verified as having not been tampered with, the inspector knows that the material has not been touched. A passive seal also guarantees the integrity of the IAEA's on-site verification tools and equipment." (Source)

Inspection types

When it comes to inspecting facilities, the IAEA has several different types of inspections at their disposal (Source):

- *Ad hoc inspections*: These inspections are used to verify a state's report of nuclear materials.
- *Routine inspections:* These are the most frequent inspections, and they can be scheduled or unannounced. Routine inspections are limited to so-called "strategic points", i.e. locations within a nuclear facility or locations containing nuclear material.
- *Special inspections*: These will be carried out if the IAEA determines that the information provided by the state is not sufficient to assure the implementation of the agreed upon safeguards.
- *Safeguards visits*: These focus on facility design throughout the entire lifecycle of a plant, from construction to maintenance to decommissioning. Increasingly, the IAEA has moved towards encouraging "safeguards by design" - "an approach whereby international safeguards requirements and objectives are fully integrated into the design process of a nuclear facility" (Source).

In addition to this and to avoid the exploitation of possible loopholes, the IAEA can conduct simultaneous inspections which make it more difficult for potentially non-compliant states to move nuclear material around to appear compliant while actually possessing more material than agreed upon (Source). On the other hand, to protect states and facilities from overregulation and undue burdens, there is a limit on how much any given facility can be inspected (Source).

Additional powers

Finally, it's important to note that the verification methods which can be used during inspections and continuous monitoring depend on the underlying agreement that has been made with the state. Additional Protocols, for example, furnish inspectors with additional competencies and facilitated access, as *Table 2* shows.

Comprehensive Safeguards Agreements	Additional Protocol
• IAEA collection of environmental samples in facilities and at locations where inspectors have access during inspections and design information verification (with sample analysis at the IAEA Clean Laboratory and/or at certified laboratories in Member States). • IAEA use of unattended and remote monitoring of movements of declared nuclear material in facilities and the transmission of authenticated and encrypted safeguards-relevant data to the Agency. • IAEA expanded use of unannounced inspections within the scheduled routine inspection regime. • IAEA enhanced evaluation of information from a State's declarations, IAEA verification activities and a wide range of open sources. • State provision of design information on new facilities and on changes in existing facilities as soon as the State authorities decide to construct, authorize construction or modify a facility. The IAEA has the continuing right to verify the design information over the facility's lifecycle, including decommissioning. • State voluntary reporting on imports and exports of nuclear material and exports of specified equipment and non-nuclear material. (Components of this reporting are incorporated in the Model Additional Protocol). • Closer co-operation between the IAEA and the State (and regional) systems for accounting for and control of nuclear material in Member States. • Provision of enhanced training for IAEA inspectors and safeguards staff and for Member State personnel responsible for safeguards implementation.	• State provision of information about, and IAEA inspector access to, all parts of a State's nuclear fuel cycle - including uranium mines, fuel fabrication and enrichment plants, and nuclear waste sites - as well as to any other location where nuclear material is or may be present. • State provision of information on, and IAEA short-notice access to, all buildings on a nuclear site. (The Protocol provides for IAEA inspectors to have "complementary" access to assure the absence of undeclared nuclear material or to resolve questions or inconsistencies in the information a State has provided about its nuclear activities. Advance notice in most cases is at least 24 hours. The advance notice is shorter - at least two hours - for access to any place on a site that is sought in conjunction with design information verification or ad hoc or routine inspections at that site. The activities carried out during complementary access could include examination of records, visual observation, environmental sampling, utilization of radiation detection and measurement devices, and the application of seals and other identifying and tamper-indicating devices). • IAEA collection of environmental samples at locations beyond declared locations when deemed necessary by the Agency. (Wider area environmental sampling would require IAEA Board approval of such sampling and consultations with the State concerned). • IAEA right to make use of internationally established communications systems, including satellite systems and other forms of telecommunication. • State acceptance of IAEA inspector designations and issuance of multiple entry visas (valid for at least one year) for inspectors. • State provision of information about, and IAEA verification mechanisms for, its research and development activities related to its nuclear fuel cycle. • State provision of information on the manufacture and export of sensitive nuclear-related technologies, and IAEA verification mechanisms for manufacturing and import locations in the State.

Table 2: Verification measures under CSAs and the AP (Source)

5.2 (Non-)Compliance

Non-compliance can occur at different levels of severity. Smaller problems regarding a lack of cooperation with inspectors such as denied access to facilities or unexpected issues with visas for inspectors are not unheard of, while larger violations or complete non-compliance is significantly more rare (Source).

Procedure

If irregularities are found, the respective member state is first contacted with a request for clarification. If more serious anomalies are detected or the issue can not be resolved by the state and the supplication of additional information, the IAEA Board of Governors will be informed, as they have the power to authorize additional measures. In extreme cases, the matter can be taken to the UN Security Council:

> "In theory, the way that the IAEA determines noncompliance with nuclear safeguards agreements is clear, straightforward, and automatic. Article XII.C of the agency's 1957 statute provides that safeguards inspectors "shall report any non-compliance" to the director-general, who "shall" in turn report such noncompliance to the 35-member Board of Governors, the agency's policymaking body. If the board determines that noncompliance has occurred, it "shall" report it to the agency's membership at large (most readily through the annual General Conference), the UN General Assembly, and the UN Security Council.[1] The word "shall" makes such steps legally binding. Action to bring a state back into compliance may be taken by the board, although its powers are limited, and by the Security Council, which has enforcement powers under the UN Charter." (Source)

What is noncompliance?

In practice, however, things are often less clear and straightforward. Part of this is due to the fact that a mutually agreeable definition of noncompliance remains missing (Source). Generally speaking, there is a lack of a clear and unambiguous legal basis on which the IAEA could act:

> "Comprehensive safeguards agreements address noncompliance in just two paragraphs (18 and 19) and do not use the word 'non-compliance.' Instead of the apparent automaticity of the statute, they give the agency flexibility in deciding what constitutes noncompliance, what actions should be taken by a state to redress it, and whether it should be reported to the Security Council." (Source)

Even when it is decided that a state is in violation of its agreement, the IAEA has only limited power to react to such a finding.

> "If the IAEA finds a State is in violation of its safeguards agreement, the IAEA's principal function is to sound the international alarm — to its Member States, to the United Nations General Assembly and, in particular, to the Security Council. The IAEA has neither the legal authority nor the means to seek physically to

> prevent diversion; in fact it is not within its power or its statutory authority to compel governments to take or to desist from any action." (Source)

The IAEA may temporarily suspend assistance to the state or revoke any rights and privileges that are associated with membership in the IAEA, but these are rather symbolic in nature and the IAEA ultimately has to defer to the United Nations, where only the Security Council has the right to impose political or economic sanctions (Source). (Needless to say, military action is also reserved to decisions made by the Security Council.) It's also worth noting that the IAEA doesn't have to respond to a case of non-compliance at all, and it's ultimately up to the Board to decide whether to take any action, including reporting it to the Security Council: "In two cases, those of Romania and Libya, it has even taken to reporting to the council 'for information purposes only,' presumably signalling that the council should take no action. The council took the hint and did not act." (Source) This arrangement makes any attempt to sanction violations susceptible to politicisation[4]:

> "As experience has taught us, for example, in the cases of North Korea and Iran, one of the greatest difficulties in deterring states from violating their non-proliferation undertakings or from ignoring legally binding UNSC resolutions is their hope that for geo-political or economic reasons at least one of the five veto-wielding members of the UNSC will oppose the adoption of effective sanctions." (Source)

Politicisation

This vulnerability to geopolitical manoeuvring constitutes a key weakness in the enforcement of IAEA safeguards. This is true even within the IAEA, before any referral to the UN Security Council is even considered:

> "Political controversy has attended the handling of every noncompliance case. This is inevitable. Governors representing member states, with all of their divergent interests and allegiances, make the ultimate judgment on noncompliance. Unless the evidence is overwhelming. which seldom is the case, the board is unlikely to rush to a decision. Because of the political implications, the board has a natural reluctance to report a fellow member state to the Security Council. Instead, the board is prone to ask for more verification and more reporting. Some governors may be unconvinced by the evidence and be reluctant, whether for political or substantive reasons, to declare a state in noncompliance. Some may wish to give the state a chance to explain itself or return quickly to compliance. Although the statute is silent on the matter, comprehensive safeguards agreements oblige the board to "afford" the state "every reasonable opportunity to furnish the Board of Governors with any

[4] One example of "the political complexities of finding a state in non-compliance" (Source) is that of Iran in the 1990s.

> necessary reassurance." The board also may collectively wish to keep the case within its own control rather than hand it over to an unpredictable Security Council, especially one subject to the veto of any of the five permanent members." (Source)

This concern about the tractability of reacting to reported noncompliance is corroborated by some evidence that has been collected in recent years. It was found that an increase in polarisation within the Board has led to a steady decline in votes approving of the decision to find a state in noncompliance (see *Figure 6*), although some of this may be explained by other factors such as ambiguous data or the complexity of a case (Source).

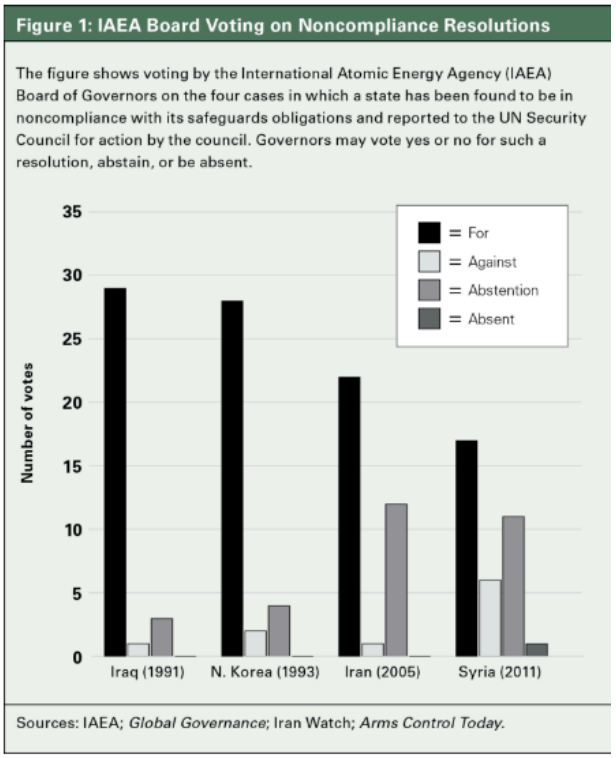

Figure 6: Historical track record of IAEA Board Voting on Noncompliance Resolutions (Source)

Historical examples of non-compliance

In this section, we examine a few select cases of non-compliance. Their selection is somewhat arbitrary, although out of the total of 8 safeguards noncompliance cases since 1991 (Source), these are the more commonly known and serious ones. One metric by which this could be measured is the level of reporting by the IAEA: In some cases (Romania, South Korea, Egypt), a single report was filed, in others (Libya and Syria), a limited amount of reports have been issued. On the other hand, the most notorious instances of noncompliance (Iraq, North Korea, Iran) have been the catalyst of numerous

and lengthy reports over a long timespan, and have attracted huge public attention throughout the process (Source).

5.2.1 Iraq (Source, Source, Source)

Date	Description
April 1991	The UN Security Council requests that inspections be carried out in Iraq by the IAEA.
July 1991	In a report to the Board of Directors, the Director General informs them about the fact that Iraq had failed to declare nuclear material, finding them noncompliant with their safeguards agreement. The Board adopts a resolution condemning Iraq due to its noncompliance.
September 1991	The Board notes that Iraq remains noncompliant and requests that the Director General reports this to the Security Council. The IAEA seizes documents in Iraq that document the extent of the violations.
October 1991	The UN Security Council passes a resolution that approves IAEA plans for monitoring. Iraq considers this to be unlawful.
July 1995	Iraq threatens to end all cooperation with the IAEA unless sanctions are lifted.
February 1998	Iraq signs a memorandum in which it assures full cooperation with the IAEA.

5.2.3 Korea (Source, Source)

Date	Description
February 1993	The Board of Governors notes inconsistencies in Korea's reporting and requests access to facilities to verify their concerns.
March 1993	Korea withdraws from the NPT.
April 1993	The IAEA finds itself unable to confirm that no diversion had taken place, and reports Korea's non-compliance to the UN Security Council.
June 1994	Korea withdraws from the IAEA.

5.2.4 Iran (Source, Source, Source)

Date	Description
2002	Allegations are making the rounds that Iran is constructing undeclared facilities in Natanz and Arak.

September 2003	The Board of Governors expresses concerns in a resolution that Iran failed to accurately report material and activities in violation of their safeguards agreement. A deadline is set for Iran to prove that it is not engaging in a nuclear weapons programme.
November 2003	The Board of Governors reiterates their concern and calls on the Director General to "take all steps necessary to confirm that the information by Iran on its past and present nuclear activities is correct and complete as well as to resolve such issues as remain outstanding" (Source).
June 2004	The IAEA adopts a resolution which, while not finding Iran in noncompliance, notes that Iran is not in full compliance.

Table 3 provides an additional overview of all reported cases of noncompliance to date, as well as the response by the IAEA.

The table shows, for each of the cases of safeguards noncompliance to date, the term used by the International Atomic Energy Agency (IAEA) director-general to characterize the IAEA Secretariat's findings, the term used by the Board of Governors in its response, the form of the board's response, and whether a report was sent to the UN Security Council.

Case	Term used in director-general reports	Term used by Board of Governors	Response by board	Report by board to UN Security Council
Iraq (1992-2005)	"non-compliance"	"non-compliance"	resolution	n/a[1]
Romania (1992)	"non-compliance"	"non-compliance"	chair statement	yes (information only)
North Korea (1992-present)	"non-compliance"	"non-compliance"	resolution	yes
Iran (2002-present)	breach	"non-compliance"	resolution	yes
Libya (2003-2008)	breach	"non-compliance"	resolution	yes (information only)
South Korea (2004)	failures to report; serious concern	failures to report; serious concern	chair statement	no
Egypt (2004-2005)	repeated failures; matter of concern	repeated failures; matter of concern	chair statement	no
Syria (2008-present)	lack of cooperation	"non-compliance"	resolution	yes

1. Because the UN Security Council tasked the IAEA director-general rather than the IAEA itself with conducting verification in Iraq, he reported on compliance directly to the council.

Sources: IAEA; United Nations; Pierre Goldschmidt, "Exposing Nuclear Non-Compliance," *Survival*, Vol. 51, No. 1, February 2009.

Table 3: Overview of noncompliance cases (Source)

An extended case study of Iran and its noncompliance can be found here.

6. Evaluation

This section seeks to evaluate to what degree the IAEA safeguards have been successful in aiding nonproliferation and increasing international nuclear safety and cooperation.

One source of evidence in favour of this thesis would be expert consensus, since many have noted that "there would be far less nuclear cooperation or trade if safeguards did not exist" (Source). The fact that the IAEA received a Nobel Peace Prize in 2005 may be seen as corroborating this account of history. With that said, expert consensus and the awarding of a prize are at best weak proxies for the IAEA's counterfactual impact.

Growth

A second way to evaluate the IAEA's track record would be to use publicly available data about membership numbers over time. The IAEA notes that the amount of nuclear material under safeguards has steadily increased over the years (between 2010 and 2022, the number rose by 24%), as has the number of member states that have safeguards agreements. New additional protocols have also contributed to increases in safety (Source). This has happened alongside increases in the IAEA safeguards budget over time (see *Figure 7*), although once again, an increase in budget should not be seen as a reliable indicator of success alone.

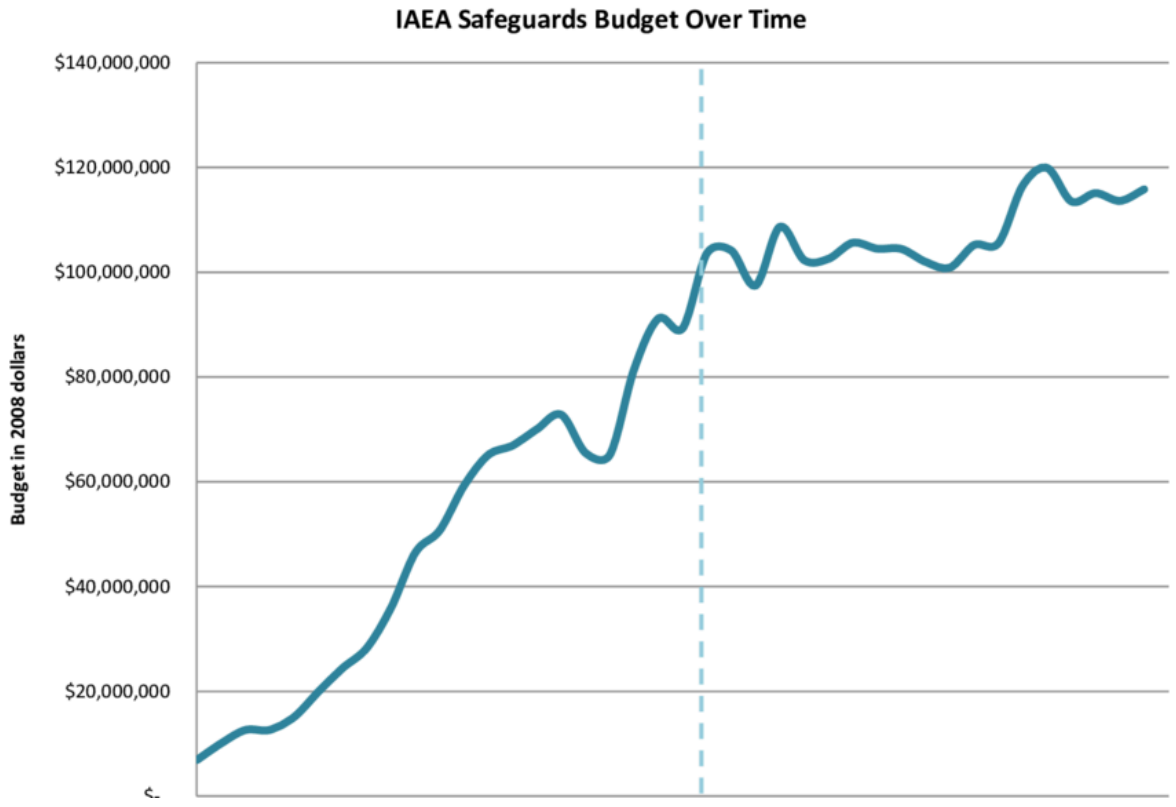

Figure 7: IAEA safeguards budget from 1970 to 2009 (Source)

Externalities

A third method by which the success of the IAEA could be determined would be to investigate positive externalities and indirect effects it has had on the nuclear safety and nonproliferation field. The most notable precedent for this would be the accelerating effect that safeguards had on bringing about the NPT: "The fact that a working IAEA safeguards system was already at hand in the late 1960s and that the IAEA provided an established means of international nuclear co-operation also made it easier to reach agreement on the content of the NPT and on the means of verifying it." (Source)

Criticism

On the other hand, a number of objections have been raised with regards to the IAEA safeguards' efficacy:

- The voluntary nature of some components of safeguards agreements (such as Additional Protocols) has been named as a key weakness of the program, since the IAEA lacks any formal power to compel states to comply (Source). This has, however, been remedied to some degree by its integration into other treaties which are more binding and more widespread with regards to their coverage. A condition of non-nuclear weapons states (NNWS) ratifying the NPT is agreement to be subject to an IAEA safeguards agreement. Ratification of the NPT makes these safeguards agreements legally binding in NNWS (Source). Nonetheless, countries like North Korea have shown that there are few mechanisms or levers to stop countries from stopping adherence to the safeguards altogether.
- Even when safeguards are in place, the IAEA has limited power or leverage to compel countries to comply, as was evidenced by some of the failures in the 1990s and 2000s (Source), and when IAEA inspectors were asked by North Korea to leave the country in 2009 (Source).
- The ability for member countries to maintain covert operations of nuclear reactors remained alive and flourishing even after the IAEA pivoted to an approach focused on completeness, and even after it strengthened its oversight through Additional Protocols. One of these cases involved Syria, which in 2011 was determined to have built but not declared a nuclear reactor prior to its destruction by Israel in 2007 (Source). This type of "leakage" represents a significant risk not only on an object level, but also because it damages trust and the confidence in the IAEA safeguards system as a whole.
- The IAEA has been criticised for not focusing on establishing "safeguard culture" enough (Source).

- The transition to a state-level concept has been controversial among some members who argue that it's an instance of the IAEA overreaching beyond its mandate (Source).

The IAEA is well aware of these weaknesses, and acknowledges its limitations in providing a strong guarantee of safety: "The IAEA made that point again in 2003 in its reports on Iraq to the UN Security Council, in which it acknowledged that proving a negative was not possible even with the authority granted under Security Council resolutions." (Source) An additional challenge the IAEA is facing in its attempt to increase nuclear safety through nonproliferation is the existence of considerable time lags in its operations. Months and even years can pass between the acceptance and the ultimate and complete implementation of a protocol, which often means there is a large number of states with accepted agreements but which are not actually in force yet (see *Figure 8*).

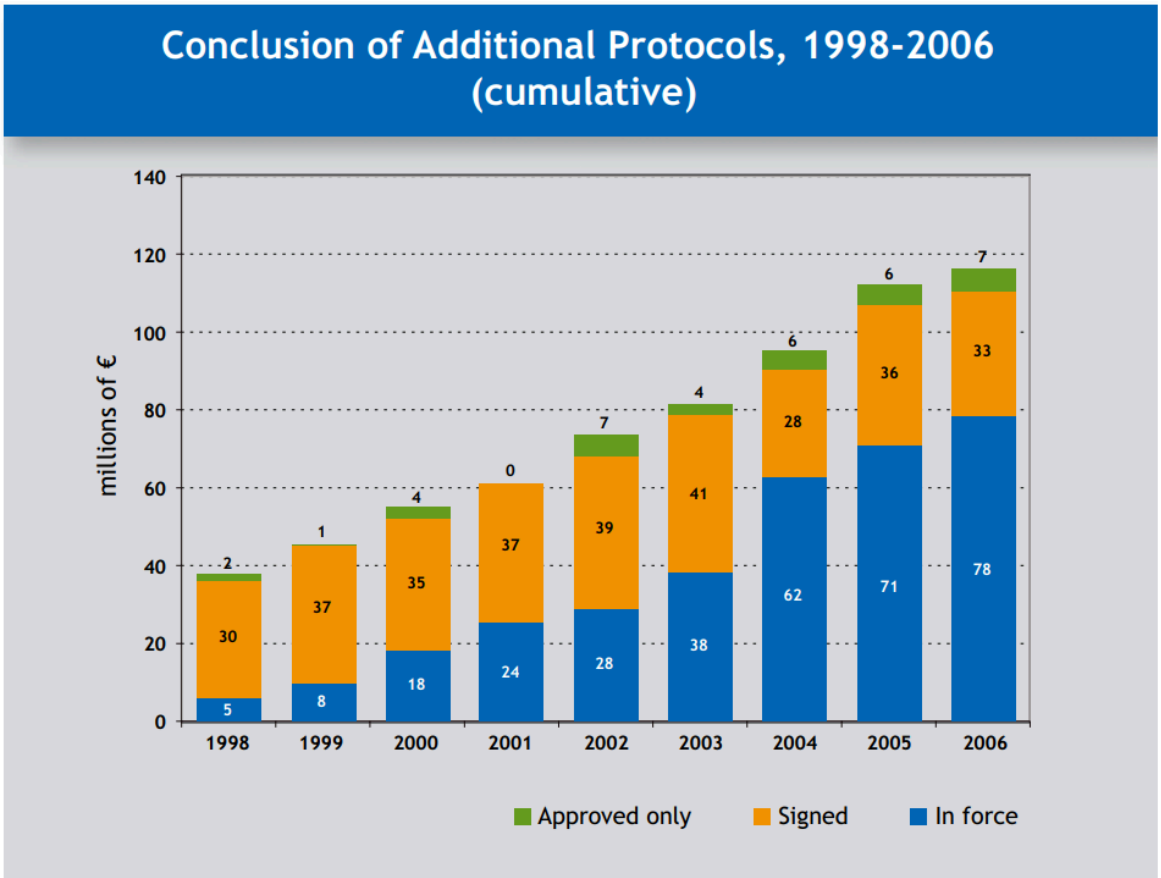

Figure 8: Overview of Additional Protocols between 1998 and 2006 (Source)

At the same time, the IAEA has introduced measures to become more efficient by introducing cost-benefit analyses and installing remote transmission monitoring systems (Source).

In summary, the degree to which the IAEA will be successful in advocating for and verifying the compliance with nonproliferation is a function of 3 different factors :

- "the extent to which the IAEA is aware of the nature and locations of States' nuclear and nuclear-related activities;
- the extent to which IAEA inspectors have physical access to relevant locations for the purpose of providing independent verification of the exclusively peaceful intent of a State's nuclear programme; and
- the will of the international community, through the United Nations Security Council, to take action against States that are not complying with their safeguards commitments to the IAEA." (Source)

As this case study has shown, limitations and challenges can occur on all three levels (and have been noted to have happened in the past).

7. Lessons for AI standards

From this case study, we may draw a few conclusions that are relevant to both the content as well as the procedures of generating, implementing and enforcing safety standards in the domain of Artificial Intelligence:

Evolution

1. **Standards and oversight programmes like the safeguards agreements can evolve (and improve) considerably over time. Lock-in effects are minimal.**

At least in the case of IAEA safeguards, we can safely conclude that steady improvements were possible and could be implemented without strong opposition to the suggested changes. Some adaptations required a previous "failure", but overall, there was a great willingness to change and adapt safeguards agreements and IAEA procedures more generally to the current needs, which inspires hope for the fast-moving space of AI development. Or, in the words of Carl Robichaud:

> "[W]e should not despair if initial rules and regulations appear too weak. Assuming sufficient warning signs, regulations can become stronger with time. It took three decades — and wake-up calls such as India's nuclear test (1974) and the discovery of a secret weapons program in Iraq (1991) — to galvanize action toward a stronger and more universal regime." (Source)

Level of oversight

2. **Deciding on the level at which oversight is executed (facility-based, state-based, internationally) can be a hugely consequential decision that is worth investigating before introducing safeguards.**

The update from a "correctness-focused" and facility-based approach to one focusing on the completeness of accounts on a national level was one of the most significant ways in which the IAEA safeguards were modified (and, arguably, improved) (Source). For the case of AI development, a similar strategic decision will need to be made, although it seems important to note that the degree to which AI development is centrally coordinated and overseen on a national level varies greatly by country (compare, for instance, the US and China). It is currently unclear to me what the right unit is for administering verification and monitoring for AI development.

Benchmarks

3. **It's important to get benchmarks right - if you don't, then whatever you end up measuring will likely be misguided and expose you to significant risk.**

Before 1991, "inspection goal attainment" was the key metric used by the IAEA to measure compliance. Iraq reached 100% of their inspection goals, making them as compliant as a member state could be (Source). IAEA safeguards were, by their own measure, very, very effective. Yet at the same time, behind closed doors, a secret nuclear weapons programme was launched, the precise thing the IAEA was tasked to prevent. For AI development, it's now established wisdom that some measurements are very easy for AIs to "game" (Source) in a similar way to how Iraq appeared compliant, but was actually pursuing different goals. Hence, the IAEA safeguard experience can serve as proof that identifying and operating with the right measurements and benchmarks will be crucial in ensuring actual, not just measured, compliance[5].

Non-discrimination

4. **A key success factor in achieving such a wide and global buy-in was the "non-discriminatory approach" taken by the IAEA. A perceived lack of power can also help to incentivize stakeholders to join a new safety-oriented program.**

Safeguard agreements were NOT set up to only monitor "the dangerous countries" - they were a universal system that didn't take into consideration geopolitical factors too much (for more on this, see point 5 on this list) (Source). This universality was helpful to get the buy-in from countries who otherwise might have perceived this as a move by powerful Western countries to control and regulate them. If safeguards agreements for AI are to be successful, they also need to have this universal and non-discriminatory nature, although it is currently unclear to me how fitting the analogy is given the differences between nuclear weapons and general purpose systems.

Politicisation

[5] Some helpful ideas for the monitoring of training runs through hardware (chips) and analogies to the IAEA verification processes can be found in this paper.

5. **Politicisation can jeopardise progress and safety, and it can undermine the credibility and trust-building goal of an oversight agency.**

The dependency of the IAEA on the UN Security Council has been described as a flaw in the system that exposes the theoretically geopolitically neutral IAEA safeguards to a nonzero amount of vulnerability to politicisation, given the veto power some countries hold in the Security Council. Should an AI agency of a similar kind have the same structures, problems of this kind could even be amplified given the high concentration of technological power in the US and China.

Non-compliance

6. **Clear definitions of what "noncompliance" entails and what happens when it is established are needed to assure accountability and predictability.**

A lack of clarity around the idea of "noncompliance" has been a recurring point of criticism. Similarly, clear thresholds for what does and doesn't qualify as "noncompliance" would be needed for AI developers. Possible candidates for this might be the use of a specified amount of compute in excess of what was previously declared (this could be expressed by a percentage), or proof that larger models with more parameters are being trained and/or deployed[6].

Integration

7. **Integrating safeguards into existing agreements can help with member acquisition and compliance, especially if the actual safeguard agreement is voluntary.**

IAEA safeguards benefited hugely from their integration into the NPT and other regional nonproliferation treaties. I am uncertain what parallel treaties may currently exist in the AI space, but the AI Act of the European Union or similar projects like the GDPR may be worth looking into. Alternatively, safeguards for AI may be understood as a means of accelerating the development of such (currently non-existent) treaties such as an NPT for AI.

Tailoring

8. **Instead of employing a "one-size-fits" approach common in regulation, AI safeguards could be tailored to labs in the same way that IAEA safeguards agreements are tailored to countries.**

A key selling point is the IAEA's willingness to work with signatories of safeguards agreements to work out the details of a given agreement, which is helped by the bilateral

[6] There is a second strategic question here regarding the timing, and whether an "IAEA" of AI would focus on a specific stage (development, deployment, post-deployment) or on all of these.

nature of the agreements. While the relevant unit for AI may be "labs" rather than "countries", it is possible such an approach will reduce fears of overregulation from corporations and key labs, thereby helping widespread adoption and helping with confidence- and trust-building.

Material

9. **An approach based on material accountancy and quantity-based regulation could be applicable to AI development.**

The closest analogy to nuclear material in the case of AI would be compute as the key "input" to AI development. Quoting Carl Robichaud once again:

> "[T]he key to effective regulation is understanding which capabilities are harmful and what the key choke points are. For nuclear weapons the limiting factor is highly enriched uranium and plutonium. These materials are produced in fuel-making and reprocessing plants — facilities that require special scrutiny. With AI, the limiting factor is most likely computational resources. Current models take hundreds of millions of dollars to train, and the requisite server clusters can be located and observed. It is hard to distinguish whether these computational resources are being used for good or ill, or ways to identify which applications pose the greatest risks." (Source)

Another mechanism that might be copied over from the IAEA is that of exceptions for small quantities. It's previously been said that small-scale developers could be crushed and pushed out of the market by sweeping regulation, and that this could have disastrous effects on the AI ecosystem. Exceptions for developers who only use small amounts of compute or training data could be a part of such an agreement, especially in cases where a "significant quantity" (to be determined) of compute is not used.

SSACs

10. **SSACs could provide a blueprint for AI agencies and organisations on a national level.**

While some countries have established AI agencies or organisations as part of their state apparatus by integrating them into existing ministries or outsourcing this work to government-affiliated think tanks, SSACS could serve as a model that member states want to adopt to provide a central point of contact. An "IAEA for AI" could provide technical assistance and support their establishment.

Buy-in

11. **Buy-in from big players can serve as a strong signal that acceptance of safeguards agreement does not automatically mean compromising competitiveness.**

Similar to how the US signed safeguards agreements to send a strong signal to other countries that perceive them as an economic competitor, a leading AI lab's decision to join a similar program could provide other actors with some evidence that safety doesn't need to trade off against economic interests. This might be even more true for AI than it was for nuclear materials, since there is a "consumer-facing" element to AI models: A company that is compliant with a rigorous and recognized safeguards agreement could use this to advertise their model as one that is particularly safe and user-friendly.

Life cycle

12. **The IAEA's focus on "life cycle approaches" for their facility design safeguards could be a precedent for auditing and evaluations throughout a model's life cycle.**

Previous research (Source, Source) has highlighted the importance of evaluations and audits throughout the entire life cycle of a model, and not just during training and/or before deployment. Rather, post-deployment monitoring and other measures are also needed. The IAEA measures for facilities and their design also extend all the way to their decommissioning, which could contribute to safety culture overall.

Continuous monitoring

13. **Facilities are continuously monitored, rather than inspecting them only irregularly or at intervals with long breaks in between site visits.**

The IAEA has automated a lot of their verification and monitoring, which has not just reduced their cost and increased their capacity (which might become increasingly important as AI models scale) but which has also increased safety. That's because as monitoring intervals become smaller, the cost of even temporary noncompliance rises significantly given the higher risk of being found in noncompliance. Yearly audits, which are common in other fields (Source) would be rather inadequate for the breakneck speed at which AI advancements are announced these days. Continuous (and automated) monitoring, on the other hand, appears more appropriate.