simeon@safer-ai.org

How Can Nuclear Safety Inform AI Safety?

Simeon Campos, James Gealy

Abstract

Recent advancements in general-purpose AI systems (GPAIS) and foundation models from leading organisations have seen AI capabilities surpass human abilities in diverse tasks. However, as these systems rapidly integrate into our daily lives, their potential risks grow, underscoring a pressing need for a robust regulatory framework at the international level. Drawing inspiration from the high-risk nuclear power industry, this paper explores lessons from nuclear safety governance, particularly the International Atomic Energy Agency (IAEA), to inform both the development and deployment of GPAIS. Key insights include the importance of early international coalitions for safety standards, the structure and processes of standards development, and post-accident investigation methodologies. The paper delves into nuclear safety principles such as "defence in depth" and risk evaluation techniques like probabilistic risk assessments (PRA), suggesting adaptations for AI. Central to the discourse is the significance of safety culture, leadership responsibility, and the need for a clear global regulatory body analogous to the IAEA for AI. By harnessing lessons from the nuclear sector, we aim to pave the way for the responsible development and deployment of GPAIS globally.

Executive Summary

As the growth in capabilities of general-purpose AI systems (GPAIS) and foundation models continues to accelerate, the risks from these systems will increase in lockstep. With GPAIS already matching human performance in many domains, expediting sensible regulation by drawing upon expertise and experience from other high-risk industries is prudent. This paper reviews the hard-won safety lessons from the nuclear power industry and identifies the most actionable and applicable for GPAIS regulation.

To begin, the formation of the International Atomic Energy Agency (IAEA) and its role in coordinating and improving nuclear safety was not inevitable at the dawn of the nuclear age. International coalition building **began with the sharing of safety practices among a small group of nations** and steadily grew, and a similar approach can be taken with GPAIS safety. In addition, developing safety standards at the international level may grant a degree of independence to the process in order to avoid safety being compromised by individual countries writing standards favourable only to their strategic interests.

The safe development and deployment of highly-capable GPAIS requires that multiple best practices be implemented. Similar to the nuclear power industry, the foremost of these is that the providers have a strong organisational safety culture. It is therefore concerning that the Silicon Valley start-up culture mantra of "move fast and break things" is currently driving the paradigm of scaling GPAIS without limit. **A strong safety culture should be the top priority** of the leadership of GPAIS providers as this has been key to reducing risks from nuclear power.

As the experience of the nuclear power industry has also shown, regulatory outcomes can be improved and innovation encouraged through the **graded approach** and **performance-based regulation**. Performance-based regulation is a promising basis for the core regulatory framework used in AI as it sets safety standards without dictating the specifics of implementation. It thus encourages innovation to meet safety requirements without excessive burden. Similarly, the graded approach—where the amount of safety scrutiny devoted depends on the degree of dangerousness from a failure—should be applied in order to reduce regulatory overhead.

Safety principles from nuclear safety that should be considered for the development and deployment of GPAIS include **post-accident investigations**, where safety practices undergo continuous improvement, and **Probabilistic Risk Assessments (PRA)**, which could reduce the likelihood that small failures combine to cause severe problems by evaluating the probability of negative outcomes in a piecemeal fashion. Other principles include safety margins—an essential part of safety in complex systems, which may require the development of new GPAIS architectures beyond transformers—as

well as using defence in depth to ensure that no single failure or error can lead to an accident. These principles should be applied to AI safety to **improve safety practices over time**.

With the capabilities of GPAIS increasing at an ever faster pace, we should look for practical ways to reduce the time to implement sensible safety measures that will reduce the associated risks. Leveraging our experience in the field of nuclear safety is likely one of the best ways to do so.

simeon@safer-ai.org

Introduction

Within the past few months, general-purpose AI systems (GPAIS) and foundation models[1] from OpenAI, Anthropic, and Google DeepMind have reached or exceeded human ability at many tasks, such as passing bar exams and coding, and their capability and range is expanding quickly; while OpenAI's GPT-3 could only handle text inputs, GPT-4 can take a photo of a sketch of a website on a napkin and output the HTML code for a functional website.[2] [3]

The breadth of capability of GPAIS is such that no single government agency today can regulate every aspect of their potential use cases. And as these models become integrated into daily life while continuing to increase in capabilities, the risks from their failure or misuse will grow substantially. Concerningly, in the last few months Yoshua Bengio and Geoffrey Hinton, two of the three 2018 Turing Award winners for their pioneering work in deep learning, have started raising the alarm about our lack of preparedness for the serious risks that AI could pose to humanity.[4] [5]

This is not the first instance where regulators have had to ensure the safety of high-risk technologies with potential life-or-death consequences. We should draw upon the experiences within safety-critical industries such as nuclear power and bioengineering to jumpstart the implementation of governance, regulatory structure, and industry best practices surrounding the development and use of GPAIS at the national and international levels.

Here, we will focus on nuclear power because the foundational technology is fundamentally high-risk, with failure resulting in life-or-death consequences for millions of people in many countries around the world. As such, security and safety is a bedrock cultural principle. Furthermore, the building and operation of a nuclear power plant requires hard-to-acquire resources, namely enriched uranium and specialized expertise, while the training of state-of-the-art GPAIS has similar bottlenecks such as the massive computing power and unique know-how required.

We will take a closer look at how the international governance frameworks and standards that regulate the global nuclear power industry can be followed and adapted to regulate the development and use of GPAIS. First, we describe the safety lessons learned, in particular the institutions and processes that were used in order to reach the current level of safety. Next, we describe the main principles and standards which are used by the nuclear safety industry. Finally, we dig further into the risk

[1] Though their definitions do not entirely overlap, we will herein refer to both GPAIS and foundation models simply as GPAIS

[2] https://openai.com/research/gpt-4

[3] https://www.youtube.com/watch?v=outcGtbnMuQ

[4] https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/

[5] https://yoshuabengio.org/2023/04/05/slowing-down-development-of-ai-systems-passing-the-turing-test/

assessment and safety assessment measures that are used and how similar ideas could be applied to the development and use of GPAIS.

We intend for this paper to serve as a source of background information for those thinking about how to regulate AI, rather than a piece of formal academic research. Our investigation into the safety practices of various high-risk industries has helped to inform our stance on AI safety and so we thought it would be constructive to write down our findings as a guide for others. This process took place over several months and included the use of GPT-4 to inform our study of the history of the IAEA and nuclear safety. This experience has only strengthened our belief that GPAIS would benefit from the lessons and experience of nuclear safety.

I. Lessons from Nuclear Safety and the IAEA

1. The Formation of the International Atomic Energy Agency (IAEA)

An initial attempt to create an international coalition for nuclear safety took place between 1945 and 1948 through the United Nations Atomic Energy Commission (UNAEC). Based on an Anglo-American proposal with agreement from the USSR, it sought to create a single international authority to manage the nuclear industry for peaceful purposes and to do away with nuclear weapons entirely.

However, the UNAEC was unable to resolve the fundamental issues between the great powers regarding the order in which inspections, disarmament and international control should be implemented, nor the underlying details. After two years of extensive negotiations and over 200 meetings, the commission reached an impasse and discontinued its work. The American nuclear monopoly ended in 1949 with the first Soviet nuclear test and the idea of a centralized international authority for nuclear regulation lost traction.

The IAEA as we know it today was created in 1957 within the United Nations family, with President Eisenhower's "Atoms for Peace" speech before the United Nations General Assembly in 1953 considered a seminal moment. Another pivotal event was the August 1955 conference on the peaceful uses of nuclear energy in Geneva held by the UN. This meeting was, "designed to lift the veil of atomic secrecy to a great extent," and included the USSR, who soon after joined the effort to create the agency.[6]

[6] https://www.iaea.org/sites/default/files/publications/magazines/bulletin/bull19-4/19401281219.pdf

Important regulatory measures were bolted on after incidents such as Chernobyl which, "led to a more receptive attitude towards proposals for expanding the IAEA's role in nuclear safety." Importantly, the IAEA promotes the peaceful use of nuclear technology through safeguards and standards, coordinating, but not controlling, national nuclear regulatory bodies.[7]

Implications for AI:

- The lengthy negotiation process required to establish the IAEA underscores the importance of implementing AI safety measures prior to the widespread proliferation of GPAIS, emphasising the need for pre-proliferation agreements, collaboration, and standards.
- Identifying mechanisms to move forward, such as sharing safety practices among allies, may be a promising way to increase trust and build a coalition for AI safety (e.g., sharing safety research and best practices with like-minded organisations).

2. The Standards Development Process

IAEA safety standards are designed to reflect an international consensus and achieve a high level of protection of people and the environment. This consensus starts with the structure of the IAEA itself.

Publication of the IAEA's safety standards is approved by the IAEA Board of Governors, one of the two policy-making bodies of the IAEA. It is currently composed of 35 member states of the IAEA and also offers guidance to other programs, including those promoting the global application of the standards.[8]

The Board of Governors appoints the Director General, who serves for a four-year term and is the head of the Secretariat. This is the administrative body of the IAEA and the core body of its staff members.

Finally, the Director General appoints the Commission on Safety Standards (CSS).[9] [10] Consensus on safety standards is achieved through the CSS and the establishment of Committees, along with five standing Safety Standards Committees.[11]

[7] https://www-pub.iaea.org/MTCD/publications/PDF/Pub1032_web.pdf

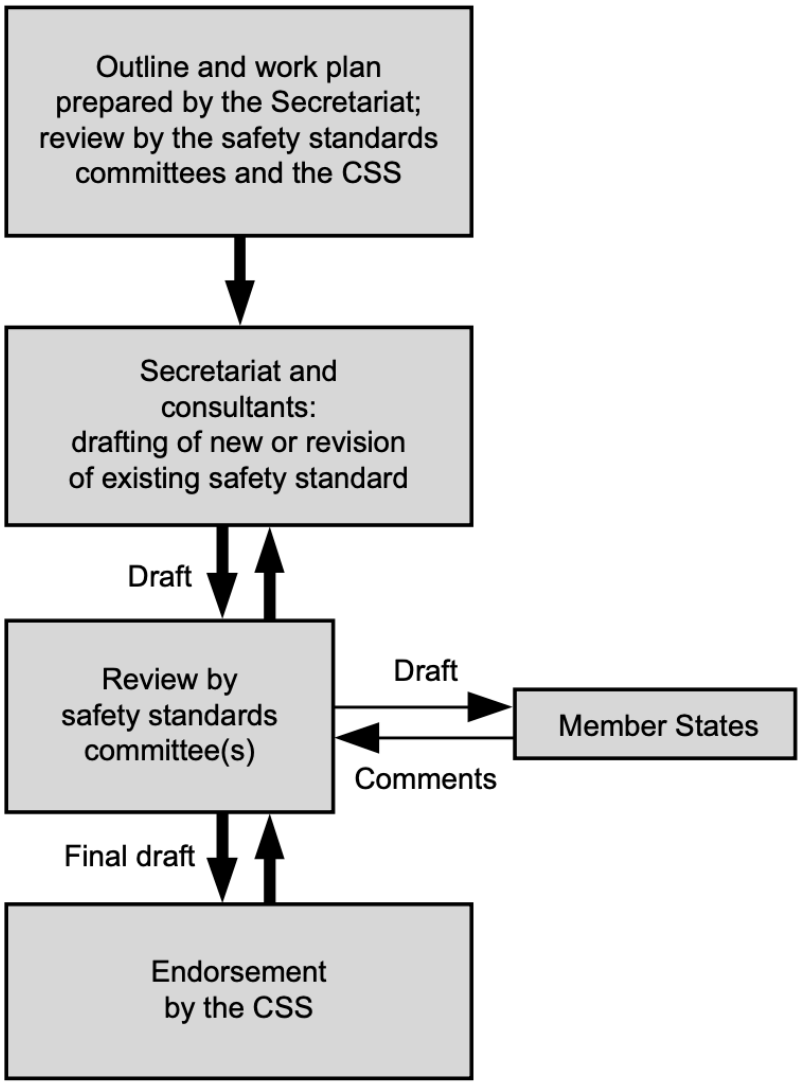
[8] https://www.iaea.org/about/governance/board-of-governors

[9] https://media.nti.org/pdfs/iaea_secretariat_EpHQ23M.pdf

[10] https://www.iaea.org/about/policy/board/rules-and-procedures-of-the-board-of-governors

[11] https://www-ns.iaea.org/downloads/standards/ss-committees-tor.pdf

The CSS is responsible for providing guidance on the approach and strategy for establishing safety standards, resolving issues raised by the Committees, and reviewing components of the safety standards before submission to the Board of Governors. The CSS emphasises the importance of member states' comments and feedback in the safety standards process, with experts from member states reviewing draft standards. This helps ensure that the standards are of requisite quality and also represent consensus among member states and address their key safety issues.[12]



The process for developing a new safety standard or revising an existing standard.[13]

Implications for AI:

[12] https://www.iaea.org/sites/default/files/21/12/ds522_nv.pdf

[13] IAEA Safety Standards, © IAEA https://www-ns.iaea.org/downloads/standards/iaea-safety-standards-brochure.pdf

- An international structure for coordinating the regulation of AI consisting of a Board of Governors, a Commission, and specialised Committees could promote a high degree of agility, moving more quickly than would be possible if coordinating among all countries directly, while countries would provide comments on drafts via committees and maintain oversight via a Board of Governors (e.g., an AI Safety Commission overseeing the development of AI safety standards and policies).

3. Post-Accident Investigation

Although the responsibility for investigating nuclear accidents rests with the state where the accident occurred, the IAEA holds advisory powers and can provide assistance to member states. This includes fact-finding missions, technical support, peer review missions, coordination of international efforts, development of guidelines and standards, and capacity building and training.[14]

There are several methods to investigate safety events in order to improve the quality of a nuclear power plant, such as Assessment of Safety Significant Event Team (ASSET), Human Performance Enhancement System (HPES), or Management Oversight and Risk Tree (MORT).[15] The key steps and elements of a safety event investigation typically include:

- Root cause analysis: A root cause analysis is conducted to identify the underlying causes of the accident. This may involve the use of various investigation techniques that are meant to elicit possible scenarios, such as event tree analysis, fault tree analysis, and failure modes and effects analysis (FMEA), to systematically examine the interactions of human, organisational, and technical factors.
- Assessment of the organisation:The investigation team evaluates the processes within the organisation and the affected facility, looking for any weaknesses or deficiencies that may have contributed to the accident. This may include examining factors such as communication, attitudes toward safety, and the effectiveness of safety management systems.
- Identification and implementation of corrective actions: Based on the findings of the investigation, the team develops recommendations for corrective actions to address the identified root causes and contributing factors. The operator of the affected facility, under the oversight of the regulatory body, is responsible for implementing the recommended corrective actions.

[14] https://www.iaea.org/topics/nuclear-safety-conventions/convention-assistance-case-nuclear-accident-or-radiological-emergency

[15] https://www-pub.iaea.org/MTCD/Publications/PDF/te_1278_prn.pdf

- Monitoring and follow-up: The regulatory body and the operator of the affected facility closely monitor the implementation of the corrective actions and assess their effectiveness in addressing the identified issues. This may involve conducting periodic reviews, inspections, and audits, as well as tracking the progress of the facility in meeting established safety performance indicators. All that information is recommended to be stored in one single database.
- Sharing lessons learned: The results of the investigation and the lessons learned should be shared with other nuclear facilities, regulatory bodies, and the international community to promote continuous improvement in nuclear safety. This typically involves mechanisms like the Joint IAEA/NEA Incident Reporting System (IRS) and the event information exchange of the World Association of Nuclear Operators (WANO).

Implications for AI:

- AI engineers and researchers should be upskilled in safety to better understand and address potential risks and hazards.
- Knowledge sharing mechanisms should be established to promote continuous improvement in AI safety (e.g., sharing lessons learned from AI incidents and accidents with the broader AI community).
- Post-accident investigation and peer review missions could be employed after significant AI incidents, with external experts assessing the measures taken to prevent similar occurrences in the future.
- Safety standards and practices should be updated based on lessons learned from AI incidents and accidents. Iterating over safety practices through incidents, near-misses, and low-stakes accidents can lead to continuous improvement in AI safety.
- The implementation of corrective actions should be closely monitored to ensure their effectiveness.

II. Key Safety Principles & Responsibilities in Nuclear Safety

1. Safety Principles

The IAEA provides guidance to member states to develop a “risk-informed” approach to regulation, defining the risk of a potential event as a combination of the probability of the event and its possible

consequences.[16] Several safety principles are used in their risk-informed regulation and we will briefly touch on some of the most important ones.

- Defence in depth is a crucial principle that involves securing systems with multiple independent layers of protection to ensure that no single failure or error can lead to an accident.[17]
- Safety margins are designed to account for uncertainties and provide a buffer against potential risks, ensuring that facilities and activities can withstand conditions more severe than those expected during normal operation.
- Performance-based regulations impose performance requirements for safety standards without necessarily specifying how they must be achieved.[18] This gives flexibility to operators while maintaining potentially ambitious goals for safety.

Implications for AI:

- Risks from AI systems should be evaluated based on their likelihood and possible consequences.
- Defence in depth could be used in all aspects of AI development and operation and should be implemented in AI by incorporating multiple layers of protection for both models and APIs. However, one core difference for accidental risks in GPAIS is the potential for adversarial optimisation that occurs from the AI itself, which may require designing layers that can counteract that specifically. In particular, one may want to set up layers of defences such that the first layer of defence is not the strongest one and that the model does not know what its layers of defences are.
- To develop reliable safety margins for AI systems, new architectures may be necessary, as current techniques do not allow us to create reliable bounds on AI capabilities, especially emergent ones.
- Performance-based regulation may reduce the need or volume of overly-specific safety standards and allow the providers of GPAIS to innovate and develop novel approaches to ensure system safety.

16

https://www.iaea.org/publications/10677/risk-informed-approach-for-nuclear-security-measures-for-nuclear-and-other-radioactive-material-out-of-regulatory-control

17 https://www.iaea.org/publications/4716/defence-in-depth-in-nuclear-safety

18 https://inis.iaea.org/collection/NCLCollectionStore/_Public/29/032/29032958.pdf?r=1

2. Responsibility Structure

"Nuclear facility operators are ultimately responsible for the safety of their facility."[19] The licensee retains prime responsibility for safety throughout the lifetime of their facilities and all activities within them, and this responsibility cannot be delegated. This responsibility includes long-term damages. "Since radioactive waste management can span many human generations, consideration must be given to the fulfilment of the licensee's (and regulator's) responsibilities in relation to present and likely future operations." [20]

Implications for AI:

- A strong responsibility and liability structure will be needed, potentially involving insurance for damages, especially given that AI development is largely driven by the private sector.
- Long-term aspects of safety, such as the protection of future generations, should be included and accounted for in AI risk analysis.

III. Risk Analysis & Safety Assessment in Nuclear Safety

1. Assessing Safety Culture

Finally, the IAEA defines a strong safety culture as the "assembly of characteristics, attitudes and behaviours in individuals, organisations and institutions which establishes that, as an overriding priority, protection and safety issues receive the attention warranted by their significance."[21] A strong safety culture influences an organisation's structure and style, as well as the attitudes, approaches, and commitment of individuals at all levels in the organisation.

Safety culture can be assessed through self-assessments, peer reviews, culture surveys, observing behaviour, benchmarking, and the monitoring of safety indicators. Self-assessments themselves evaluate an organisation's safety culture through interviews, surveys, observations, and document reviews. Safety culture assessment is crucial to identify key failure points and suggest improvements.[22]

[19] https://www.iaea.org/publications/factsheets/nuclear-facility-safety
[20] https://one.oecd.org/document/NEA/RWM/RF(2008)4/PROV2/en/pdf
[21] https://www.iaea.org/topics/safety-and-security-culture
[22] https://www.iaea.org/publications/10742/performing-safety-culture-self-assessments

Peer reviews have external experts evaluate safety culture using criteria such as leadership commitment to safety and accountability for safety. These reviews provide insights and facilitate the sharing of experiences between organisations. The IAEA offers peer review services like Operational Safety Review Teams (OSART) and Integrated Regulatory Review Service (IRRS) to help states assess safety culture. The mission of OSART is to conduct in-depth, peer reviews of the operational safety performance of nuclear power plants.[23] The mission of IRRS is to conduct peer reviews of a country's regulatory infrastructure for nuclear and radiation safety who assess the regulatory framework against IAEA safety standards and good practices.[24] Culture surveys and interviews gauge employee perceptions of culture and safety to provide insight into barriers and areas for improvement. Observing behaviour during routine and emergency operations can identify positive and negative practices.

Benchmarking compares an organisation's safety culture to that of other organisations, as well as industry standards, to find best practices and areas for improvement. Participating in industry groups and reviewing reports and guidelines enables benchmarking. Safety culture indicators are tracked and monitor culture over time. Leadership commitment, accountability, trust, and learning are crucial to promote safety culture. Leaders set expectations, provide resources, and hold accountability, while commitment actions can indicate the overall culture.

Implications for AI:

- Strong safety culture is one of the most important levers to decrease the probability of accidents. Therefore, it is crucial to ensure that organisations developing frontier AI systems meet at least minimum requirements for sensible and responsible safety cultures. It is important to note that the startup culture at many organisations developing frontier AI is grounded in principles that are strongly antagonistic to safety culture, such as the ethos "move fast and break things".
- Safety culture should be evaluated using the techniques described above and include most of the objective criteria that those reviews encompass, e.g., self-assessment, peer reviews and benchmarking.

2. Graded Approach

Nuclear safety implements the graded approach, which matches an appropriate level of safety assessment, precaution, and process or product control to the potential risks of a facility or activity.

[23] https://www.iaea.org/services/review-missions/operational-safety-review-team-osart

[24] https://www.iaea.org/services/review-missions/integrated-regulatory-review-service-irrs

This helps allocate resources efficiently and prioritise safety changes. Periodic safety reviews assess facilities periodically and help identify necessary changes. They review safety performance, ageing, experience, analyses, and identify upgrades.[25]

Implications for AI:

- The graded approach could be applied to AI in order to make sure that risks are well prioritised, and that the safety overhead does not prevent the development and innovation of small models, while maintaining a high degree of safety for the most dangerous models.

3. Probabilistic Safety Assessment Techniques

Probabilistic Safety Assessment (PSA)—also called Probabilistic Risk Assessment (PRA)—systematically analyses potential accidents and failures in order to estimate their likelihood and consequences, essentially creating a map of what could go wrong after an initial problem occurs. It involves the identification of possible initiating events, the development of accident sequences, and the evaluation of their consequences. This method also looks at how small failures can combine to cause severe problems, using probabilities. PSA results inform safety improvements and operational decisions.[26] In contrast, deterministic safety analysis evaluates if safety systems can respond to anticipated events. It identifies necessary safety margins and the effectiveness of safety measures.[27]

To calculate risks, PSA recognises potential starting issues, develops possible sequences of events, studies the possible results, and quantifies the risks. When specific data is limited, expert opinions and comparable historical events are used. It is essential to use a variety of methods for the best estimates, especially for events that are rare but could have serious consequences.

Various techniques are used to estimate how often safety systems and other components may fail and how human errors can impact these initiating events. This includes understanding how an operator's decision-making process can impact their performance in critical tasks. Overall, PSA is a critical tool that helps identify weaknesses and prioritise safety improvements in nuclear facilities.

Implications for AI:

[25] https://www.iaea.org/publications/10643/use-of-a-graded-approach-in-the-application-of-the-management-system-requirements-for-facilities-and-activities

[26] https://www.iaea.org/publications/3789/probabilistic-safety-assessment

[27] https://www.iaea.org/publications/12335/deterministic-safety-analysis-for-nuclear-power-plants

- Just like in nuclear facilities, AI systems can face complex challenges and potential failures. By creating a map of possible faults and understanding how small errors might combine into significant issues, we can improve system design and anticipate problems before they occur. This approach is expected to guide AI systems towards more "fail-safe" operation. Also, human interaction with AI introduces another layer of complexity and potential error. The analysis of human reliability, such as how human decisions impact the functioning of the AI, can enhance the interaction between users and the system, reducing potential errors.

Conclusion

The lessons learned from the nuclear power industry and the IAEA provide valuable insights for the development and implementation of safety measures for GPAIS and their providers. The establishment of a global regulatory body for AI safety, similar to the IAEA, would help to ensure the safe and responsible development use of AI technologies.

Moreover, the responsibility for the management of GPAIS safety should be clearly defined and rest with the organisations providing these technologies. Safety needs to be a core focus of the leadership of providers and it should be ingrained in their culture, while national and international regulatory agencies should provide guidance and feedback.

Finally, risk analysis and safety assessment techniques used in the nuclear industry should be adapted for use in GPAIS safety assessments. These techniques can help to identify potential risks and hazards, prioritise safety improvements, and ensure that GPAIS "fail safe". However, it is important to note that the unique characteristics of AI, such as the potential for goal misgeneralisation, require additional considerations and may necessitate the development of new architectures for both GPAIS and their safety standards. Given that GPAIS are already exceeding the ability of humans at many tasks, it is crucial that we draw upon the lessons learned from these other high-risk industries to ensure their safe and responsible development and use sooner rather than later.