Foundation

Careful Industries

Levers for the Safe Adoption of AI

Foresight Review Positioning Paper

Rachel Coldicutt OBE,
Uchenna Anyamele,
Madhuri Karak,
Dr. Odongo Oduor Joseph

February 2026

Lloyd's Register Foundation

About Lloyd's Register Foundation

Our vision

Our vision is to be known worldwide as a leading supporter of engineering-related research, training and education, which makes a real difference in improving the safety of the critical infrastructure on which modern society relies. In support of this, we promote scientific excellence and act as a catalyst working with others to achieve maximum impact.

Lloyd's Register Foundation charitable mission

- To secure for the benefit of the community high technical standards of design, manufacture, construction, maintenance, operation and performance for the purpose of enhancing the safety of life and property at sea, on land and in the air.
- The advancement of public education including within the transportation industries and any other engineering and technological disciplines.

About the Lloyd's Register Foundation Report Series

The aim of this Report Series is to openly disseminate information about the work that is being supported by Lloyd's Register Foundation. It is hoped that these reports will provide insights for research, policy and business communities and inform wider debate in society about the engineering safety-related challenges being investigated by the Foundation.

T +44 20 7709 9166
E info@lrfoundation.org.uk
www.lrfoundation.org.uk

About this document

This Foresight Review into the Safe Adoption of AI aims to understand the current and future impacts of Frontier AI on engineered systems and develop recommendations for future safe adoption. This paper draws on the literature review, The AI Safety Paradox (published in parallel), and sets out the midpoint perspective of the project.

Credits

Writing and additional research:
Rachel Coldicutt

Worker Safety literature search and analysis:
Uchenna Anyamele

Environmental Safety literature search and analysis:
Madhuri Karak, PhD

Critical Infrastructure literature search and analysis:
Dr. Odongo Oduor Joseph

About Careful Industries

Careful Industries is a UK-based inclusive innovation studio. Through research, foresight, and prototyping we understand the impacts of technologies and create more inclusive futures through policy development and training.

www.careful.industries

Glossary

AI System:	A machine-based system designed to operate with varying levels of autonomy that generates outputs such as predictions, recommendations, decisions, or content. An AI system typically comprises data, algorithms or models, and computational infrastructure working together to perform tasks with varying degrees of autonomy.
Artificial Intelligence (AI):	Artificial intelligence is a broad, multi-disciplinary field of computer science that makes possible a number of advanced computing functions. These computer functions include analysing and processing data, using rules to organise and output information, and organising and completing tasks. Artificial intelligence systems also "learn" as they undertake these tasks, and so can make progress without requiring human intervention.
Critical Infrastructure:	Physical and digital systems essential to national security and public welfare, such as power grids, transportation networks, and water systems.
Generative AI:	AI systems capable of creating new content, including text, images, and code, based on learned patterns from training data.
General Purpose AI (GPAI):	An artificial intelligence system designed to perform a wide range of tasks across multiple domains, rather than being built for a single, narrowly defined function. The term has gained particular prominence in regulatory contexts, notably the EU AI Act, where it refers to AI models trained on broad data at scale (such as large language models) that can serve as a foundation for many downstream applications.
Large Language Models (LLMs):	Advanced AI systems trained on massive amounts of text data to understand and generate human language.

and there are ongoing disagreements as to whether the term "safety" relates to current societal harms or to future existential risks, with some critics venturing that the "safe" use and adoption of current forms of AI is not possible.

1. AI Safety

AI is a complex field of computer science in which "safety" is still an emergent and disputed concept. As the UK AI Security Institute noted in 2023:

> *since our understanding of AI safety is nascent, it is not yet possible to build full safety cases that scale to risks posed by models significantly more advanced than those of today.*[1]

Meanwhile, as the technical discipline of AI assurance develops, there are ongoing disagreements as to whether "safety" relates to current societal harms or to future existential risks, with some critics venturing that the "safe" use of AI is not possible.[2]

Our contention is that the safe adoption of AI relies on safety throughout the supply chain, specifically that safe use and deployment of any AI system must be accompanied by its safe creation and ongoing operation.

Our research indicates that AI technologies can deliver measurable improvements in efficiency, safety monitoring, and predictive capability across critical infrastructure systems. However, the use of General Purpose AI systems in particular also risk introducing new failure modes, vulnerabilities, and governance challenges that are often poorly understood and unevenly regulated. This "safety paradox" is exacerbated by many factors, including the pace of technological development;, the global AI race that leads developers to prioritise rapid releases over good governance; and the lack of international regulatory alignment.

While it is beyond the scope of this project to suggest fixes to these political and economic factors, our research so far suggests there are more intrinsic changes to the AI lifecycle that could increase the prospects of safe adoption in engineered systems.

General Purpose (GPAI) refers to AI systems that are not designed for a single task or domain but are capable of performing a wide range of distinct tasks across multiple applications. Large language models and foundation models are among the most prominent examples of general purpose AI, in contrast to narrow or task-specific AI systems built for defined purposes such as fault detection or fraud monitoring.

1 Geoffrey Irving, 'Safety Cases at AISI | AISI Work', AI Security Institute, 23 August 2023, https://www.aisi.gov.uk/blog/safety-cases-at-aisi.

2 Emily M. Bender and Alex Hanna, *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want* (Bodley Head, 2025); Adam Becker, *More Everything Forever: AI Overlords, Space Empires, and Silicon Valley's Crusade to Control the Fate of Humanity* (Basic Books, 2025); Dan McQuillan, *Resisting AI: An Anti-Fascist Approach to Artificial Intelligence* (Bristol University Press, 2022).

2. Literature Review Findings

To establish a clear starting point for foresight activities, we conducted a literature review to summarise significant and emerging themes relating to AI safety in the fields of critical infrastructure, worker safety, and environmental safety. 333 sources were analysed between November 2025 and January 2026, including established and peer-reviewed academic literature plus "grey" material such as news and media reporting and analysis, think tank reports, social media posts and newsletters, and pre-print research.[3]

The literature review concluded that:

i) The opacity of AI supply chains is a structural barrier to safety.

The lack of transparency over the provenance of training data, the invisibility of human labour in ostensibly automated systems, and the absence of standardised environmental impact measurement mean that organisations procuring or deploying AI-enabled tools frequently lack the information necessary to conduct meaningful risk assessment or due diligence. This opacity is intensified by proprietary systems, corporate publication policies, and fragmented global supply chains that combine to obscure the full range of upstream and downstream impacts.

ii) Existing assurance and governance frameworks are insufficient to address the sociotechnical complexity of AI safety.

The literature consistently shows that technically focussed assurance methodologies fail to capture the broader societal, environmental, and labour impacts that accompany AI deployment. Risk frameworks that treat AI safety as a primarily technical problem miss more systemic effects: for instance, the rebound effects that undermine the environmental case for autonomous vehicles, the mental health impacts of human–robot collaboration, or the ways algorithmic bias in financial services deepen existing social divisions. A sociotechnical approach that integrates technical, social, environmental, and economic dimensions would improve safety assurance by offering a more complete picture of the end-to-end impacts of adopting a given system.

3 A full bibliography is available in *The AI Safety Paradox: A Literature Review on the Safe Adoption of Artificial Intelligence in Engineered Systems*

iii) The international regulatory landscape is fragmented in ways that compound rather than mitigate risk.

The fundamentally divergent approaches of the European Union, United States, and China reflect competing visions of AI's role in economic and societal development. None of the frameworks examined fully address supply chain transparency, worker protections, or environmental justice, and the competitive pressures between nations and firms continue to accelerate product release cycles in ways that outpace regulatory capacity. Top-down global alignment appears unlikely in the short term, suggesting that alternative mechanisms — including sectoral standards, procurement requirements, and multi-stakeholder governance initiatives — will need to play a more prominent role.

iv) The distribution of AI's costs and benefits is profoundly unequal.

Across every domain reviewed, the harms of AI development and deployment are disproportionately borne by those with the least power to shape its trajectory: workers in Low and Middle-Income Countries who label data and moderate content under exploitative conditions; communities in climate-vulnerable regions whose water sources are contaminated by mineral extraction or whose air quality is degraded by data centre operations; and populations subject to biased automated decision-making in domains such as finance, welfare, and justice. The literature makes clear that AI safety cannot be meaningfully assessed without attending to these distributional questions, and that governance frameworks which fail to centre the rights of affected communities will entrench rather than address existing inequalities.

v) The pressure to rapidly deploy emergent and untested technologies can displace both governance and assurance.

The growing body of evidence on environmental, labour, and societal harms has not so far served to limit the scope of General Purpose AI development; as such, without the development of new technical models or a pivot to a "safety-by-design" approach, the continued expansion of GPAI systems will likely continue to generate safety deficits that are displaced onto the most vulnerable.

3. Exploring Safety Levers

The second half of the Foresight Review will explore potential safety levers in more depth, focussing on:

a) Provenance and Assurance

For organisations that buy AI products, build tools and systems that sit on top of externally created large language models (LLMs), or rely on generative AI tools to create software, the scrutiny that comes with purchasing power is a significant lever for ensuring improved safety outcomes. These include:

- **AI Supply Chain Transparency,** characterised by high-quality data about provenance that runs both upstream and downstream from development and deployment. In addition to technical transparency regarding data and model composition, this could include data on workplace conditions for those involved in hardware production and data labelling, clear data on environmental impacts, and detailed risk assessments regarding future social impacts.
- **Sociotechnical Evaluations and Assurance Methods** that run "end-to-end" across the AI supply chain, treating safety not as a fixed property of a system but as an emergent characteristic.

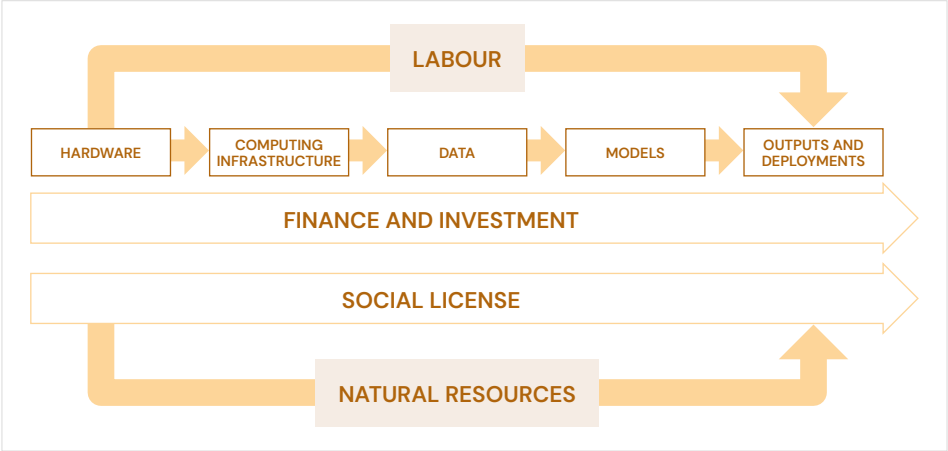


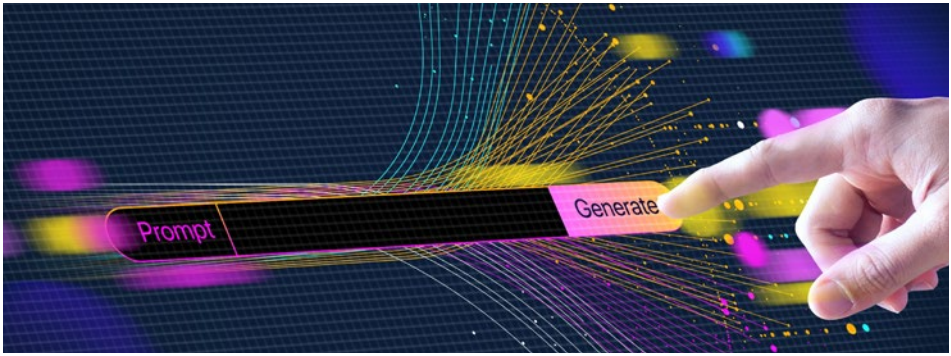
Figure 1 — A Sociotechnical View of the AI Supply Chain

b) Future Development Models

General Purpose AI systems are purposefully developed to be broadly useful in a range of different settings. This broadness creates a number of barriers to robust quality assurance of GPAI for use in safety-critical, narrow use cases. At the time of writing, these barriers include "alignment faking", when GPAI models behave strategically to pass assessments, and hallucinating, when models incorrectly "guess" outcomes.[4] Not only are GPAI systems difficult to assure, they are also dependent on large amounts of data and on natural resources for their continued operation.

It is likely that the major AI labs will continue to pursue refinements to their own GPAI systems over the coming years, and create new ones. As such, the opportunity for invention may most sensibly lie elsewhere: in developing and refining alternative development models. Areas for further investigation include:

- **Future Developments in Narrow AI,** such as applied machine learning, Frugal AI, and the use of small models, that may be easier to safely assure.
- **Computing within Planetary Boundaries,** sustainable approaches to hardware and software development
- **"Safety-by-Design"** as a development approach to current and future AI hardware and software development.



4 Greenblatt, Ryan, Denison, Carson, Wright, Benjamin, Roger, Fabian, MacDiarmid, Monte, Marks, Sam, Treutlein, Johannes, Belonax, Tim, Chen, Jack, Duvenad, David, Khan, Akbir, Michael, Julian, Minderman, Sören, Perez, Ethan, Petrini, Libra, Uesoto, Jonath, Kaplan, Jared, Shlegeris, Buck, Bowman, Samuel. R, Hubinger, Evan, "Alignment Faking in Large Language Models." *arXiv.Org*, 18 December 2024, https://arxiv.org/abs/2412.14093; Kalai, Adam Tauman, et al. "Why Language Models Hallucinate." arXiv:2509.04664, arXiv, 4 Sept. 2025. *arXiv.org*, https://doi.org/10.48550/arXiv.2509.04664.

4. Next Steps

In the final stage of this project, we will use expert interviews and participatory foresight workshops to understand what future developments are in train, or could be initiated, to make an end-to-end approach to the safe adoption of AI tenable.

Future workshops and research activities include (March to June 2026) include:

- Qualitative research to understand good work and safety for data workers (partner: Data Labelers Association)
- Workshop on embodied AI safety (partner: Global Center on AI Governance)
- Workshop on the development of sustainable AI systems (partner: Royal Academy of Engineering)
- Workshop on sociotechnical approaches to assurance
- Research on future developments in Narrow AI and Safety-by-Design.

If you would like to share your work or expertise as a part of this process, please contact hello@careful.industries.

Foundation

Careful Industries

February 2026

Lloyd's Register Foundation