THE CENTRE FOR
LONG-TERM RESILIENCE

Transforming risk governance at frontier AI companies

Ben Robinson and James Ginns

June 2024

www.longtermresilience.org

THE CENTRE FOR
LONG-TERM RESILIENCE

Transforming risk governance at frontier AI companies

Ben Robinson[1] and James Ginns

The Centre for Long-Term Resilience | June 2024

REPORT CONTENTS

[1] Contact: Ben Robinson, ben@longtermresilience.org

SUMMARY

This report explores how aspects of best practice risk governance – particularly the Three Lines Model (3LoD), which separates risk ownership, oversight and audit – could be effectively implemented at frontier AI companies to ensure safer model development and deployment. This aims to support the UK government's approach to potential future legislation mandating better practice risk management, as well as making the case to the companies themselves on the value of these approaches.

Overall, current risk management practices at frontier AI companies appear to be relatively ad hoc and lack overarching strategy and structure. While there has been some recent progress – such as responsible scaling policies and more comprehensive safety evaluations – this isn't supported by a holistic risk governance structure for how these initiatives interact and complement each other. **Risk governance provides the overarching framework, context and rules of engagement for the various risk management practices and functions in an organisation. In short, it is the 'glue' that holds various risk management practices together.** We argue that implementing best practice, like the Three Lines Model, would help AI companies better identify, assess and act on risks, thereby reducing the chance of harmful models being released.

The report is structured as follows:

- First, we outline the fundamental principles of risk governance and give case studies of where these have been effectively implemented in safety-critical industries such as nuclear, healthcare and aviation.
- Second, we apply these principles to a generic frontier AI company structure, showing what roles currently exist and what likely requires improvement to align with best practice.
- Third, we explore the importance of risk culture, arguing that these methods can operationalise the nascent safety cultures that already exist.
- Fourth, we outline legislative options to mandate forms of best practice risk management in law, such as distinct risk and assurance functions and external audit. We finish with concrete recommendations for governments and companies.

RECOMMENDATIONS

For Government

Require AI companies to establish and maintain an office of risk management and an internal audit function, submit an annual resilience statement demonstrating the efficacy of their risk management process, undergo an annual external audit, and establish a protected 'speak up' channel (whistleblowing equivalent) with appeal to external bodies where necessary.

Build consensus within business and civil society about the importance of more holistic risk management, including a specific focus on risk governance. This could include publishing papers or facilitating workshops.

For companies

Build consensus

Encourage internal discussions about how best practice risk management (overarching risk management framework, 3LoD, appetite statements) might be useful, in particular, how clarifying risk ownership and reporting lines could reduce/help manage risks from AI.

Champion and sponsor best practice and dynamic risk management in the organisation by board members and senior executives acknowledging that this will require buy-in across the organisation to be effective.

Implement better practice – an eight-point checklist:

1. Encourage a stronger sense of risk ownership in research, product and engineering teams through workshops, training and engagement with specialist risk and internal audit functions.
2. Experiment with an MVP version of 3LoD structure and related methodology, and explore ways of sharing learnings with other companies (e.g. via the Frontier Model Forum).
3. Introduce an office of risk management with a central risk management team reporting to a Chief Risk Officer or equivalent to provide challenge, a degree of independent oversight, and risk reporting to the board.
4. Introduce an independent internal audit team to provide assurance on the process.
5. Formulate/agree on risk appetite statements based on thorough risk identification and assessment.
6. Introduce measures to encourage and enhance healthy risk culture: leadership and 'tone from the top', regular pulse checks/surveys with results reviewed by the board, emphasis on 'just culture', a protected 'speak up' channel with independent review and board visibility, and regular cross-functional risk identification workshops at different levels.
7. Seek external assurance on the risk management process and overall compliance via external audit.
8. Produce an annual resilience statement demonstrating the efficacy of the risk management process.

Part 1
BACKGROUND AND INTRODUCTION

Background

Purpose

This is a guide to applying best-practice risk management, particularly the Three Lines Model governance structure, to frontier AI companies. This aims to a) support the UK government's approach to potential future legislation and b) make the case to the companies themselves on the value of these methods.

Scope

We focus on applying one aspect of risk management – the Three Lines Model risk governance structure – to AI companies developing foundation models (henceforth "AI companies"). We are particularly focused on frontier AI companies (e.g. OpenAI, Anthropic, Google DeepMind).[2] However, professionalising risk management should be a goal for all AI foundation model companies, whether or not at the frontier of capabilities, and so our report is aimed at those companies too.

Methodology

We draw on a mix of past professional experience implementing the Three Lines Model, a review of relevant AI risk management literature, and conversations with people in government, civil society and AI companies.

Acronyms and Glossary

3LoD	The Three Lines Model or Three Lines of Defence Model	**CRO**	Chief Risk Officer
KPI	Key Performance Indicator	**KRI**	Key Risk Indicator
RMF	Risk Management Framework	**RSP**	Responsible Scaling Policy
Risk Management	The discipline of managing uncertainty in an organisation in order to safely, reliably and ethically achieve objectives		
Risk Governance	The establishment of a risk management framework (RMF) and related accountabilities across an organisation, such that risk management activities are integrated and strategically aligned		
Three Lines Model	A risk management framework (RMF) model that aims to clearly separate risk ownership from oversight from audit (previously called the Three Lines of Defence, then updated to the Three Lines Model by the International Institute of Auditors[3])		
Risk Appetite	A statement of the appropriate and reasonable level of risk exposure for an organisation to take based on the nature of its business activity		
Risk Tolerance	The extent of risk an organisation is capable of taking		

[2] The UK Government defines this as a "subsection of AI developers ... invest[ing] large amounts of resource[s] into designing, building, and pre-training the most capable AI foundation models."
[3] 'The IIA's three lines model: An update of the three lines of Defense'

Introduction

AI companies are increasingly recognising that the development of novel, increasingly capable general-purpose AI systems could pose numerous risks to individuals and society, ranging from bias and stereotyping, to manipulation and misuse, to more hypothetical risks of agentic misaligned AI.

It is promising that companies are starting to take responsibility for assessing and managing these risks. Particular focus so far has gone into developing methodologies for identifying and assessing risks[4], identifying mitigations for responding to different levels of risks[5], and considering how organisational structures may affect safety.[6] Governments have also been supporting these efforts.[7]

However, it appears that many of these initiatives and approaches have been created organically over time, in response to improving capabilities and changing risk landscapes. As one policy and governance employee noted: "our approach to risk management is admittedly ad hoc". This sentiment is supported by other people we've spoken to within companies and in governments and civil society. We think this creates three clear risk management failure modes:

1. **A risk is not identified** meaning a company may be unaware of this potential harm to individuals, businesses or society. This could be because of things like ineffective governance structures, or misaligned incentives.

2. **A risk is identified but not properly assessed**, meaning a company may be aware of a risk but are not able to measure or effectively track it due to things like lack of appropriate methodology or lack of capability and coordination.

3. **A risk is identified and assessed but not properly acted on**, including both failing to act at all, or taking the wrong action. This could be because of a lack of clear mitigation plans, a lack of coordination between teams or misaligned incentives.

The aim of risk governance is to address and minimise these kinds of failure modes, ensuring that comprehensive risk assessment and mitigation plans can be deployed effectively, responsibilities for doing so are clear, conflicts of interest are avoided, and there's internal accountability for prioritising risk management even under other commercial pressures. The Three Lines (3LoD) Model is a specific approach to risk management, well established as best practice in many industries like finance and aviation, which involves clearly separating **risk ownership** from **oversight** from **audit**. Below are some high-level reasons why AI companies should consider implementing best-practice risk governance:

> **Greater strategy and structure.** Despite deep apparent concern about AI risk within these companies, as above, risk management practices are often ad hoc and lack overall strategy and structure. The Three Lines Model is a whole-of-organisation approach, ensuring risk identification, assessment and mitigation practices are governed and

[4] For example, safety evaluations at OpenAI, Google DeepMind and Anthropic.
[5] For example, Anthropic's Responsible Scaling Policy.
[6] For example, OpenAI has an unusual structure where a nonprofit controls a capped profit company (though it's not clear how this cap works); Anthropic is a public benefit corporation with a Long-Term Benefit Trust that independently selects board members (though beyond this, it's unclear how much power they have).
[7] For example, the US government's NIST AI Risk Management Framework emphasises the importance of effective governance for AI companies, with reference to 3LoD in its accompanying playbook and more recent Generative AI Profile. Similarly, the UK government has published a range of risk management related guidance, including its Emerging Processes for Frontier AI Safety document, which outlines nine 9 recommended safety practices for frontier AI organisations (though with a limited focus on risk governance).

implemented effectively and uniformly in practice. This could help to cement and operationalise good intent.

Greater efficiency. By improving oversight, identifying and closing gaps in risk coverage, and avoiding unclear risk accountabilities and task duplication, the Three Lines Model increases the effectiveness of risk management practices. It also enhances efficiency by overcoming the inefficiency of ad hoc arrangements, implicit assumptions (e.g. around risk tolerances), silos and blurred accountabilities.

Greater transparency and oversight. As well as improved oversight within the business (particularly between the board and management), demonstrating good practice externally will foster transparency with users and government stakeholders. The need for oversight is likely to only increase in the future, with greater downstream uptake of these models, and the demand from these organisations that risk is being effectively managed upstream.

Finally, we think there is a strong case for these companies to **learn from best practice.** While the size and complexity of AI risk are novel in some sense, AI companies should learn from other industries where these methods have worked before. These methods will need to be adapted to a start-up context, but the degree of novelty of these companies or of AI risk doesn't warrant a fundamentally different approach.

Next, we give an overview of best practice risk governance and 3LoD (section two) before discussing how this could be applied in practice to an AI company and where the biggest gaps seem to be currently in companies' application of these principles (section three). We then analyse the role of risk culture (section four) and legislative options for enforcement (section five) before finishing with recommendations for Government and companies (section six).

About the authors:

Ben Robinson was previously in Deloitte's Risk Advisory practice, focussed on governance issues relating to climate change and emerging technology. Before this, he was at The Ethics Centre, an applied ethics think tank. He holds a BA Hons (First Class), with a thesis on AI risk and automation.

James Ginns was previously a Chief Risk Officer in the aviation sector, and has held a number of other senior leadership positions in the industry. He has significant risk management experience in the private sector implementing the Three Lines model, and has served as an advisor to several non-profits.

Acknowledgements: Thanks to Emma Lawsen for designing the report, and to Alexis Harrell for copy-editing. We'd also like to thank a range of people within the UK government, AI companies and civil society for useful discussions and feedback throughout the writing process, particularly Malcolm Murray and Jonas Schuett from GovAI.

Part 2
RISK MANAGEMENT FUNDAMENTALS

Section insights

- There are inherent dangers in uncertainty and our response to it as we pursue objectives. Risk management is the discipline of managing the effects of uncertainty on objectives such that they can be safely and reliably achieved.
- Risk governance aims to ensure risk management initiatives are properly integrated and aligned with strategic objectives. This includes establishing an overarching risk management framework (RMF) and related accountabilities across an organisation.
- The Three Lines Model is a RMF that provides checks and balances allowing for defined risk ownership, oversight challenge and independent assurance.

Risk Management Principles

Risk management is the discipline of managing the effect of uncertainty on our objectives, such that they can be safely and reliably achieved. This means exposing and countering our worst impulses – tendencies like overconfidence, fear, ignorance and bias – by systematically encouraging accountability, transparency, challenge, balance and trust.

There are a litany of crises and accidents in both private and public sectors over the past half-century which highlight that grappling with uncertainty can bring out the worst in us. Some examples: the origins of the financial crisis of 2008-9 and the cavalier way in which the financial sector became reliant on instruments such as the credit default swap (CDS); the tragic loss of the space shuttle Challenger in 1986, which was found to be partly caused by NASA's desperation to launch despite unsuitable conditions; the role that groupthink likely played in the UK's initial response to the Covid pandemic, which was anchored in the appropriate response to a flu pandemic despite being a coronavirus.

One way of better managing uncertainty in an organisation is through ensuring a series of checks and balances in a common governance framework, somewhat akin to the separation of powers in a polity. The Rogers Commission found that an indirect cause of the Challenger accident was a lack of checks and balances in NASA's management structure. No function in any organisation should be responsible for reviewing its own work, which means separating management and oversight and having independent external audits. Expanding on this, below are some key risk management principles:

Clear front-line risk ownership: The front-line of an organisation is best qualified to manage its risks and should be held accountable for doing so.

Independent oversight with separate reporting lines: Oversight should be limited to an advisory role but should report outside the organisation's business reporting line to ensure maximum objectivity and independence.

Leadership knowledge and buy-in: Senior stakeholders should be provided with an understanding of the nature and scale of risks the organisation faces in order to evaluate whether levels of risk are reasonable and arrive at a balanced view.

Effective risk culture: The resulting transparency and accountability should inspire a culture of trust throughout the business, which in turn will facilitate effective and

imaginative risk reporting. The organisation as a whole is thereby enabled to be more resilient.

These principles are operationalised through frameworks such as the Three Lines Model, explored below.

Application of Risk Management Principles

The Three Lines of Defence or Three Lines Model (3LoD), formalised by the Institute of Internal Auditors in 2013 and updated in 2020, is seen as a best practice risk governance framework in the private and increasingly in the public sector[8] to achieve the required checks and balances in effective risk management. The energy and utilities, healthcare, pharmaceutical and aviation sectors all offer useful examples of the application of 3LoD beyond the financial services sector, where it originated.

Accountability for risk ownership is separated in first line functions from oversight in the second and assurance or audit in the third. Functions in all three lines collaborate closely, but they report separately, and the structure ensures that no function reviews its own work.

The 'first line' comprises all risk owners in the business, accountable for assessing and effectively managing all risks associated with their day-to-day activities and decisions.

The 'second line' comprises those involved in risk oversight, accountable for ensuring the risk management process is functioning as it should, developing appropriate policies, controls and processes within the business, and evaluating the overall levels of risk being taken by the organisation (Chief Risk Officer, risk management team, compliance and legal teams). Where levels of risk either exceed or threaten to exceed established tolerances (see below – 'Risk Appetite'), the second line ensures they receive appropriate executive attention and that any emerging risks are being identified and framed within these tolerances. They provide an independent view to the board that all risks are being contained within approved tolerances. The Chief Risk Officer provides a single point of accountability for the risk management process, usually reporting to both the CEO and to the independent chair of a board risk or audit committee, and thereby vested with a degree of independence. The second line provides a form of consultancy to the business, but not one to which the management of the business's risks can be outsourced.

The 'third line' comprises an independent audit function provided by internal and often – importantly – external auditors to test business controls and processes, including risk management and governance, and provide independent assurance to the board of their effectiveness. The Head of Internal Audit reports to the Chair of the Board Audit Committee.

Successful implementation of 3LoD requires that risk and audit functions are managed under a common framework cross-cutting the organisation, both have 'access all areas', no function reviews its own work, and clear accountability for decisions is assigned to a named officer on the 'first line'.

If implemented effectively, this separation of 'lines' ensures that it is clear who is accountable for identifying, assessing, reporting and mitigating different risks. This minimises the danger of a) risks going unidentified or unaddressed, or b) risks being not property acted on, even if they are identified, because of lack of clear mitigation plans, lack of coordination between teams or misaligned incentives. It also inspires trust and thereby encourages the internal reporting of concerns by team members.

[8] For example, IIA's application of 3LoD to the public sector, and the UK Government's Orange Book

Examples of the Implementation of 3LoD in Safety-Sensitive Sectors

The common elements in these examples are risk ownership strictly vested in the first line, independence of oversight and assurance functions, engaged board committees, and informed boards who are enabled to opine on whether the levels of risk being taken in the business match with agreed risk tolerances.

Nuclear

The principle of independent oversight and the role of challenge from the second line is stressed in the UK Nuclear Industry's Good Practice Guide. This includes "assisting the organisation to avoid complacency through encouraging a feeling of open 'chronic unease'". As a major player in the nuclear energy sector, EDF Energy has a stringent application of 3LoD, referred to as 'control lines'. The first line includes plant operators with robust procedures and safety culture, the second line comprises dedicated safety oversight and nuclear inspectors, and the third line is an independent audit by both internal auditors and external nuclear safety authorities.

Healthcare

NHS trusts have successfully applied a 3LoD approach to patient safety. The first line is the clinical staff who provide direct patient care, the second line includes the patient safety and quality assurance teams that develop and enforce safety protocols, and the third line is the internal audit function that evaluates the effectiveness of patient safety measures. This example highlights a comprehensive approach to patient safety and quality of care.

Pfizer is one example among many in the pharmaceutical industry that apply 3LoD, with a Chief Compliance, Quality, and Risk Officer responsible for overseeing a global Ethics & Compliance Programme. This position reports directly to the CEO and regularly reports to the Audit Committee and the Regulatory and Compliance Committee of the Board of Directors.

Aviation

The International Airlines Group (IAG) operates an Enterprise Risk Management Framework (ERM), with risk owners and management supported by a second line independent ERM function and a third line Audit and Compliance Committee.

In no case is legislation prescriptive to the extent of mandating 3LoD in these various sectors. A combination of broader principles of risk management in non-binding codes such as the UK's Code of Corporate Governance and industry-specific guidance such as that in the UK's Nuclear Industry has encouraged boards to be across risk in their organisations and stressed the importance of separate oversight and audit functions.

The Importance of Risk Appetite

Risk appetite describes the desired level of risk exposure that is appropriate and reasonable for an organisation to take. Appropriate risk tolerances make up risk appetite based on the nature of its business. The monitoring and reporting of risk levels related to agreed and stated tolerances within the organisation is intrinsic to the implementation of effective risk governance. This allows the board to consider the nature and scale of risks reported in the light of agreed tolerances and recommendations made by the second line risk management team.

Risk appetite statements are rarely published in full, may contain quantitative and qualitative elements, and often cover financial, reputational, marketplace and strategic risks.

A useful, comprehensive example of a risk appetite statement from Network Rail:[9]

*'We use company-wide risk appetite statements, split into four levels – **minimal, cautious, open, and eager** – to outline how much risk the Board is willing to take for Network Rail to achieve its strategic goals:*

Safety, health and environment	***Minimal:** We will seek to continually reduce safety, health, and environment risks across the system, reducing the likelihood of serious injury or loss of life to the public, passengers and workforce, or irreversible environmental damage.*
Political and stakeholder	***Open:** We are willing to accept some negative exposure to support high risk strategies, including national media coverage, political or regulatory scrutiny (i.e. our stakeholders).*
Financial	***Open:** Within our core business, we are willing to accept and invest in opportunities with inherent financial risks, where these are understood and proportionate to the expected benefits to passengers and freight users. Outside of our core business, we are only willing to accept and invest in opportunities with moderate inherent risks, where these are understood, proportionate to the expected benefits and undertaken with necessary external approvals. We are prepared to accept minimal risk of a breach of our agreed funding limits and will allocate funding to create buffers to mitigate the risk.*
Train performance	***Open:** We are open to new approaches and will work across the industry to build back better following the pandemic. Innovation will be supported where the risks are understood and proportionate to the expected benefits. Where risks are poorly understood, we will be cautious about making any decision that could negatively impact on train performance for passengers and freight users.*

The statements are built into our process and are integral to our rating and reporting of risks. Each risk is assessed against all four of these appetite areas, and we report on whether they're in or out of appetite'.

This final sentence of the above example is key – the appetite statements provide the board with a yardstick by which to assess an organisation's risk levels, reported through the oversight function.

Conclusion

The novelty and pace of change at the frontier of AI development hardly makes companies immune to the dangers of managing uncertainty. The degree of uncertainty and scale of the risks potentially involved with their business, as well as the benefits it brings, demands that they adopt best practice risk management, albeit tailored and adapted to their relative start-up culture. The clarity of accountabilities and board transparency provided by 3LoD-based governance, as well as the formation of agreed risk tolerances and appetite statements, should facilitate both efficiency and strategic direction if implemented effectively.

The following section explores how this might be approached for an AI foundation model company.

[9] Network Rail's Annual Report, 2022

Part 3

APPLYING THE THREE LINES MODEL TO AI COMPANIES

Section insights

- 3LoD and best practice risk management may be helpful in supporting AI companies to avoid failure modes of not identifying risks, not assessing risks and not acting on risks.
- An initial attempt at mapping 3LoD to a generic AI company structure is given, aiming to show how this model applies and where companies may need uplift.

In this section, we first outline why we think 3LoD is a promising route for AI companies wanting to better manage risk and uncertainty before showing how this could be applied in practice.[10]

Why AI companies should implement 'Three Lines'

Potential failure modes of current risk management approaches

While risk management and governance efforts within AI companies have increased recently,[11] these efforts **appear to be relatively ad hoc and unsupported by an overall governance approach and strategy**. See the below quote from a policy and governance employee at one of these organisations:

> *'Our overall approach to risk management is admittedly ad hoc.... We have some good initiatives, but we would like to make this more holistic in a way that's appropriate to us.'*
>
> – AI company Policy and Governance employee

This sentiment is supported by other people we've spoken to within the companies and in governments and civil society familiar with how these organisations operate. We think that a lack of overarching strategy and implementation of best-practice risk management frameworks, like 3LoD, exposes these companies to undue levels of risk to their business and potential harm to individuals and society.

Making this more tangible, there are **three broad categories of failure modes** that current AI companies may be exposed to, which risk management practices aim to avoid. Below are some examples, with further discussion given in the next section.

[10] We're grateful to Jonas Schuett and Malcolm Murray from GovAI for their support with this research, particularly the sections on identifying what roles currently exist at AI companies and conducting a high-level gap analysis. Both GovAI and CLTR have spoken to a range of people at AI companies to inform this analysis, but as highlighted in the text, this should be treated as fairly preliminary with more work needed within companies to both substantiate what currently exists and how to mould this to align with best practice.

[11] Some examples: in the last year, OpenAI have created a Preparedness team to evaluate future risks of frontier models, and have expanded their Safety Systems team focussed on evaluating current models. See Appendix A for a more detailed outline of what roles appear to exist currently, and how this maps to 3LoD.

1. A risk is not identified. A company may be unaware of this potential harm to individuals, businesses or society. This could be because of:

a. Ineffective governance structures, meaning policies and processes are not in place to identify risks, particularly cross-cutting risks, that no one individual team may be responsible for. This is the high-level coordination view of risk across the organisation; in the terminology of 3LoD, it would mainly be the second 'risk oversight' line, headed by a Chief Risk Officer. An example of this in the context of AI companies may be certain misuse risks: if there aren't effective safety evaluations for all the ways these models could be misused by bad actors (for instance, enhancing the capabilities of novice cyberattackers), unsafe models could be released without the company knowing this was a potential risk.

b. Misaligned incentives: Another reason that organisations can fail to identify risks is that the incentive structures do not facilitate comprehensive risk identification. Even if overall governance structures and reporting lines are effective, if individuals and teams aren't empowered and incentivised to identify and manage these risks on a day-to-day basis, they could be missed. An example of this in an AI company could be a first line product team wanting to get a model out to market as quickly as possible and failing to conduct routine safety checks.

2. A risk is identified but not properly assessed. A company may be aware of a risk but aren't able to assess it. This could be because of:

a. Lack of appropriate methodology. In the case of an AI company, this might look like companies not having developed thorough enough capability evaluations to test for certain risks they may be aware of (e.g. more diffuse, systemic risks); or not utilising diverse enough risk assessment methodologies (e.g. scenario planning and wargaming).

b. Lack of capability and coordination. Effective risk assessment also requires having appropriate staff capability and coordination between teams. Given the wide range of AI risks, from autonomy, to misuse, to more diffuse structural impacts, it's important companies can draw on a mix of skill sets when assessing risk, including technical, sociotechnical and risk management experience.

3. A risk is identified and assessed but not properly acted upon. This includes both failing to act on risks at all, or taking the wrong action, and could be because of:

a. Lack of clear mitigation plans: Without clear risk mitigation plans, including clearly assigned actions, responsibilities and timelines, risks can be identified but not acted upon, or acted on in the wrong way, allowing them to incubate and potentially worsen over time. In an AI company, this could include a range of risks, such as bias and stereotyping highlighted below or failing to act on potentially dangerous agentic capabilities. Given how quickly these models are improving, this type of risk is important to stop early, as letting it incubate could make it significantly worsen over time.

b. Lack of coordination between teams: Related to lacking clear mitigation plans, a common failure mode for why certain risks like misuse or bias may be inadequately acted upon is that it is seen to be another team's responsibility. This can occur because of unclear, poorly designed or conflicted reporting lines (e.g. risk teams reporting to heads of product or engineering), blurred responsibilities (e.g. sitting across different lines), or because of a lack of communication between teams about certain risks and how they should be actioned. Silos between teams can lead to incorrect actions (e.g. being ignored or deprioritised) even if they are identified.

c. Misaligned incentives: Incentive structures also affect how risks are acted upon, even if they are identified. For example, the incentive to get a product to market as soon as possible

may mean models are publicly released, even if there are known safety concerns that haven't been resolved. A potential example of this is bias in LLMs; it is well known that LLMs like ChatGPT often make stereotypical associations due to underrepresentative or biassed training data. This can be a hard problem to solve if the training data reflects biases in society writ large, and there aren't easy ways of fixing this. The decision to release a model in such situations should depend on the risk tolerance of the organisation, how the potential risk is weighted, and the degree to which individuals in business roles are empowered to make decisions about risk.

The above examples of failure modes are not exhaustive; there are various other reasons why risks may not be identified at all, identified but not properly assessed, or identified but not acted on. These include cultural elements (fear of speaking up), further capability failures (not having sufficient risk training or expertise to raise issues) or other structural factors like the fact that mitigations can be expensive to implement or have a negative impact on user experience (e.g. filters that block false positives). More examples are given in the following sections, as well as how 3LoD helps avoid these.

How 3LoD and better risk management can help avoid failure modes

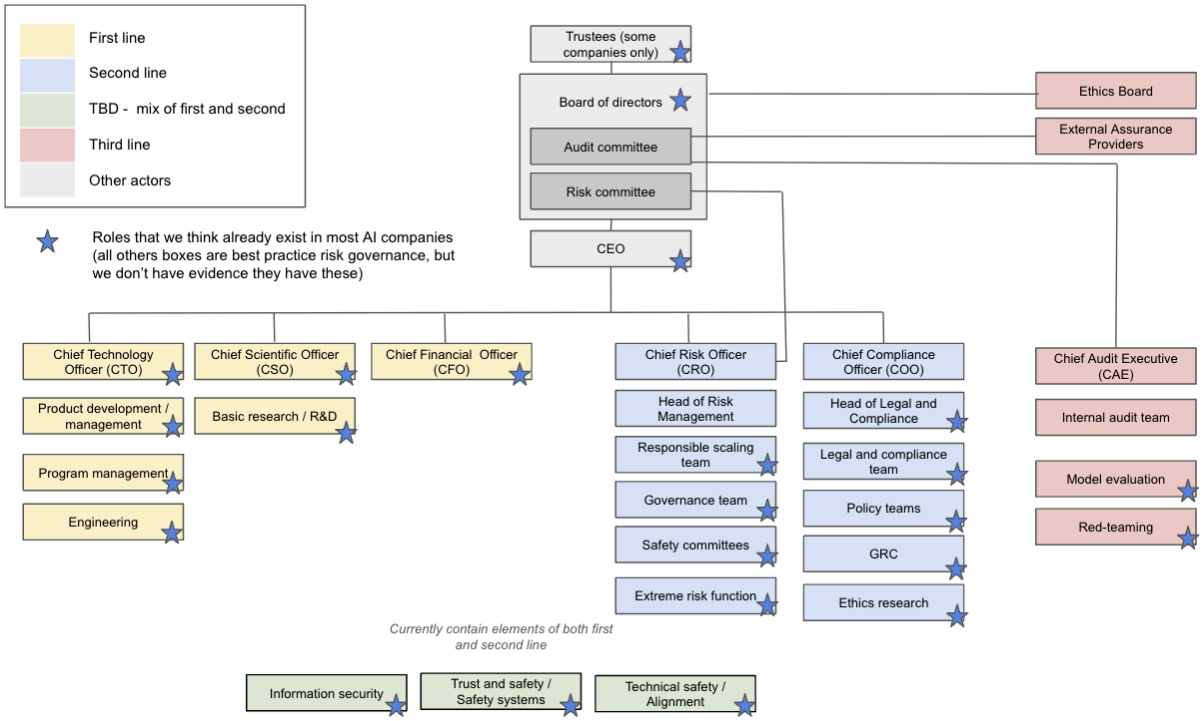
3LoD and best practice risk management can avoid the above failure modes by giving **greater structure around managing risk** by ensuring risk assessment and mitigation practices are governed and implemented effectively and uniformly in practice, and **greater transparency and oversight** by clearly separating risk ownership from oversight from audit, facilitating challenge, and being able to demonstrate this to external stakeholders.

Below is a high-level overview of how 3LoD could be applied to AI companies, with the following sections expanding on each of these points in more detail:

First line: risk ownership	Second line: risk oversight	Third line: risk audit
The first line is the "core business" of any organisation; the people who deliver products and services and who are primarily responsible for owning the associated risks of those products and services. In AI companies, this includes product and engineering, research and development, and program management teams.	The second line oversees the first line to provide additional risk expertise and monitor and challenge risk practices. In AI companies, this includes risk and governance teams and safety committees.	The third line audits the first and second lines to test the efficacy of the risk management process. In AI companies, this currently includes model evaluation and red-teaming processes.

How 3LoD could be applied to an AI company

To demonstrate how this might work in practice, below is a proposed structure for how 3LoD could be applied in an AI company.[12] The specific roles and teams are based on a broad understanding of some of the kinds of functions that exist in some AI companies already (boxes with blue stars in the bottom right-hand corner) and those that would be needed in addition to implement 3LoD effectively (non-starred boxes).



Below are some example rationales for why certain teams are placed in different lines, with Appendix A providing a more detailed description of each of the above teams and the rationale for placement. Appendix B evaluates the extent to which current risk activities align with best practice, and Appendix C gives advice for implementation.

- **Engineering (first line):** Engineering teams design, develop, and maintain the software infrastructure and systems for AI foundation models. They are in the first line because they are responsible for managing technical risks and ensuring the robustness and reliability of AI systems.
- **Governance team (second line):** The governance team develops and implements governance frameworks and policies for the responsible development and deployment of AI foundation models. They are in the second line because they provide oversight and support to ensure AI development aligns with legal, ethical, and societal expectations and that governance practices are effective.
- **Red-teaming (third line):** Red teams conduct adversarial testing and simulated attacks on AI foundation models to identify vulnerabilities and weaknesses. They are in the third line because they provide independent assurance of the efficacy of the risk management process by this process of vulnerability and weakness detection. While red-teaming has come to mean various things within the context of AI companies (including sometimes as an internal form of testing a specific system for risks, which would be a first line function),

[12] This diagram is a more detailed extension of a diagram given in Schuett (2023) Three Lines of Defense Against Risks From AI (p. 6), to which James Ginns contributed.

they ideally should be seen as independent assurance function rather than a risk identification function, and are therefore third line.

First line: Risk Ownership

What this ideally looks like: This is the "core business" of any organisation. It's where decision-making happens and strategy is set. In AI companies, this includes teams responsible for product and engineering, research and development, and program management.

- **Risk ownership**: People and teams in the first line should be risk owners. That is, they need to be incentivised and empowered to be aware of the types and levels of risk that their decisions create and to take deliberate actions in managing these risks. The key behaviours associated with ownership include *accepting* risk (taking a risk as it is deemed to be within tolerance), *mitigating* risk (preparing for and reducing the impact of a risk if it materialises), *transferring* risk (shifting a risk to another party) or *avoiding* risk (removing or eliminating a risk entirely).

What currently exists: There are various teams within the first line at AI companies. However, it's unclear the extent to which these teams have the capability, structure, incentives or culture to facilitate effective risk management. Below are some examples of these teams, followed by a table providing a preliminary gap analysis evaluating the extent to which ideal risk activities are being met.

- **Product management:** All companies have forms of product management teams that define the strategy and development of new models and related products.
- **Program management:** All companies have forms of program management teams that support the rollout of cross-functional projects.
- **Engineering**: All companies have engineering teams that design, develop, and maintain the software infrastructure and systems for AI foundation models.
- **Basic research / R&D:** While under different names, all companies have Research and Development teams that conduct basic research into state-of-the-art AI and ML.

Preliminary gap analysis:[13]

Line	Activity	Implementation?	Explanation
First line	Risk ownership	Limited	There is little evidence of a strong risk culture and risk ownership in "the business" — i.e. the delivery teams of these organisations. While certain individuals and teams clearly care deeply about risk, the overarching culture of risk ownership and reporting lines appears to be ad hoc, fragmented and lacking structure. It's unclear the extent to which any of these teams see themselves as 'risk owners' or have clear Key Risk Indicators and reporting lines that enable best practice risk management.

[13] The gap analysis table here and in the following sections are to be treated as a basic, preliminary evaluation of the extent to which AI companies are meeting the ideal risk activities for different lines in the Three Lines model. This is necessarily preliminary as there are limitations to how much information can be found about the internal operations of companies, as well as difficulties generalising across companies. See Appendix B for this table in full.

Second line: Risk Oversight

What this ideally looks like: The second line oversees the first line to provide additional risk expertise as well as monitor, support and challenge risk practices. An outline of these activities is given below:

- **Monitoring:** Risk monitoring means defining KRIs (Key Risk Indicators) and metrics to track, finding the right data sources for these, ensuring the data is kept up to date, creating dashboards that display the data, etc.
- **Advice:** The owners of the risks sit in the "business", but even the most risk-aware business will not have as much risk knowledge as the specialist risk and assurance functions that are normally referred to as the second line. A key role of these functions therefore is to advise the business on all risk matters. The specialist risk and assurance functions need to be the experts on each risk domain, coupling intimate risk knowledge with knowledge of the business and strategy and ensuring that their advice is offered to decision-makers at the right time to be valuable.
- **Challenge:** Although they should not make the final decisions, the specialist risk and assurance functions should play a "challenger" role, pressure testing the business's plans and decisions to ensure they are risk-informed. This requires sufficient independence, objectivity and subject expertise.
- **Policy creation:** Policies should typically be created in the specialist risk and assurance functions, as these functions tend to have more expertise in what makes for effective policies and have better oversight of the business as a whole. When creating policies, these teams should collaborate closely with first line functions who likely have the deepest knowledge of specific risks, but keeping this process centralised in a second line function ensures effective integration between policies and better overall coverage of risk across the business.
- **Aggregator:** A key function of the specialist risk and assurance functions is to be the aggregators. Each individual risk owner in the business can only be expected to see part of the fuller picture. It is imperative that the specialist risk and assurance functions aggregate data across risks and create a holistic view of all the various risks that the organisation is taking as a result of its decisions.
- **Reporting:** The specialist risk and assurance functions need to provide sufficient risk information in their reporting to senior management and the board for them to have a holistic view of the risk picture.
- **Control advisory:** The specialist risk and assurance functions are experts on controls and mitigation measures and should provide that expertise as new projects are executed so that risk management is built into processes from the start.

What currently exists:

- **Policy teams**: All companies have policy teams. While their size and mandate vary, most of them seem to conduct policy research and engage with policymakers. These teams sometimes have several subteams, such as OpenAI's Policy Frontiers Team.
- **Ethics research:** Google DeepMind has an Ethics Research Team that conducts and publishes academic research.
- **Governance, Risk, and Compliance (GRC):** OpenAI has a GRC team. This is likely squarely in the second line, but since it's fairly new, its remit is unclear.

- **Governance teams**: Google DeepMind has an AGI Strategy and Governance team. This team seems to mostly perform second line activities such as providing advice.
- **Safety committees:** OpenAI recently announced its new Safety Advisory Group (SAG), and Google DeepMind has several councils, such as the Responsibility and Safety Council (RSC) and the AGI Council. These are more senior committees that provide advisory and potentially some oversight.
- **Responsible scaling team:** Anthropic's Responsible Scaling Policy (RSP) states that there is a Responsible Scaling Officer in place who will offer oversight.

Preliminary gap analysis:

Line	Activity	Implementation?	Explanation
Second line	Monitoring	Some evidence	There is some evidence of monitoring taking place. For example, OpenAI's Preparedness team aims to "track, evaluate, forecast and protect against catastrophic risks".
	Challenge	Limited	There is limited evidence that the business is sufficiently challenged by an independent oversight function. Anthropic's newly announced Alignment Stress-testing team may be positioned to provide some challenge, but it's unclear yet how effective this will be, including whether it's adequately independent.
	Advice	Some evidence	The policy and governance teams seem to have some ability to offer advice to the business. However, their level of influence is unclear. This often comes from having reporting lines to senior people.
	Policy creation	Substantial evidence	These companies all have policy teams that are regularly creating new policies. E.g., see the variety of policies outlined in responses by Anthropic, Google DeepMind and OpenAI prior to the UK AI safety summit. It's unclear the extent to which these policies are integrated into day-to-day operational decision-making in relevant teams or supported by an overarching risk management framework, but there is no doubt substantial evidence that policies are being created.

Teams that mix first and second line

Companies also have a number of teams that seem to perform a mix of first and second line activities. These include:

- **Information security:** Most AI companies have information security teams. Like in most organisations, the Chief Information Security Officer (CISO) and information security teams tend to conduct both first and second line activities (ownership and oversight of risk).
- **Technical safety:** Superalignment Teams (previously at OpenAI, now at Anthropic) are an example of a technical safety team that can be seen as straddling first and second line activities, given it both makes decisions about risk through AI-powered alignment initiatives, and oversees risk by creating policies to ensure risk processes are effective.
- **Trust and Safety (T&S):** Most companies, like many other online platforms, have T&S teams. OpenAI calls this 'Safety Systems'. These teams mostly, but not exclusively, conduct first line activities insofar as they test systems and manage risk. However, they also conduct second line activities due to the fact they create policies and oversee the risk assessment process.

A 'purist' approach to 3LoD would aim to separate these roles to avoid confusion arising from overlapping responsibilities. However, in practice, overlapping first and second line responsibilities in these teams can exist in transition towards 3LoD, so long as responsibilities and reporting lines are made clear to people within the relevant teams so there is no conflict between ownership and oversight. Oversight roles need to retain independence via reporting lines to second line management. There also needs to be clearly defined risk ownership, adequate oversight challenge and a central risk management function.

Third line: Risk Audit

What this ideally looks like: The third line auditing the first and second lines to test the efficacy of the risk management process. While the second line monitors and challenges the risk management practices of the first line, the third line independently assesses this process, supervising the supervisors. For all organisations, it's crucial these assessments are truly independent to be effective (i.e. not able to be influenced by others within the organisation).

There are two kinds of audit functions:

- **Internal independent assurance:** Some of the specialist risk and assurance functions should have full independence in order to be able to provide fully independent assurance. This assurance should be provided to the board as well as external stakeholders where appropriate. This independent assurance should range from assurance over important individual processes to assurance over the risk management system as a whole. The former means checking that key individual business processes are well-controlled and managed in line with risk tolerance. The latter means checking that all risk management processes function as designed, e.g. that information flows appropriately, etc.
- **External independent assurance:** In addition to internal independent assurance, it is often valuable and necessary to also have external independent assurance. Depending on the risk in question, this can be required due to the need for specific expertise that doesn't exist in the organisation or if the risk calls for there to be additional checks and balances and 'layers of defence' (external audit is sometimes referred to as a fourth line of defence).

What currently exists:

- **External model evaluations:** METR conducts model evaluations on some of OpenAI's and Anthropic's models. In doing so, they provide external assurance.
- **Red-teaming:** Similarly, all three companies have stated that external experts have red-teamed some of their models, which can also be seen as an example of external assurance. Red-teaming is, by definition, a third line activity, as it is concerned with assurance in terms of testing resilience and identifying vulnerabilities and gaps as opposed to standard product testing.
- **Non-board governing bodies:** Anthropic has its Long-Term Benefit Trust (LTBT), which is an additional governing body that has the power to select some of the board members.

Preliminary gap analysis:

Line	Activity	Implementation?	Explanation
Third line	Aggregation	Missing	There does not appear to be any central risk team in these organisations that aggregates all risk information into a holistic picture.
	Reporting	Missing/Limited	Naturally, there is some level of reporting to senior management, but it is unclear how focused it is on risk. It's also unknown if the specialist risk and assurance functions get time with the board.
	Control advisory	Missing	This is normally performed by an internal audit team, which does not appear to exist in these organisations, so there is likely no control advisory in place.
	Internal assurance	Missing	There do not appear to be any internal audit teams in these organisations, which is typically quite synonymous with internal independent assurance. Anthropic's newly announced Alignment Stress-testing team might be seen as providing some assurance over the alignment teams.
	External assurance	Limited	There is some external independent assurance taking place since most of these organisations invite external parties to design evaluations and to conduct red-teaming on the models. This seems likely to be limited to just model capabilities at this time. Another area where external assurance is often provided is penetration testing in information security. The labs have not stated whether they use external parties for this, but it would be likely.

Case study: risk is identified with and without 3LoD

To demonstrate how this framework works in practice, consider a hypothetical example:

A product team wants to publicly launch the latest version of a model because they have a Key Performance Indicator (KPI) attached to a specific delivery date, despite the fact the red-teaming process has identified safety concerns relating to stereotyping and bias.

Without 3LoD: the product team might have a vague sense that this is an important risk to be aware of, but without a framework or processes for balancing this over other competing commercial incentives, the model is likely released. Alternatively, the model isn't released because an individual product team manager makes a call, but in both situations, it is up to individual preferences rather than being a structured, transparent business decision aligned with organisational strategy (and ideally a stated risk appetite set by management and the board).

With 3LoD: the product team has a Key Risk Indicator (KRI) built within their KPIs, meaning they take risk seriously and have processes in place for balancing this with other commercial incentives. They can draw on existing frameworks, like a risk matrix that considers various financial and non-financial elements, to make a transparent all-things-considered decision. The second line oversight function monitors the effectiveness of first line processes and KRIs, in line with management's stated risk appetite. The third line audit function independently assesses the extent to which risk ownership and oversight processes are effective; red-teaming within the third line is separate from commercial incentives and is able to escalate findings to the board if necessary. Outcome: a decision is made about releasing the latest version of the model that is transparent and consistent with broader business strategy and process. The inefficiency and 'chronic unease' of unstated risk assumptions are avoided and rationales for decisions can be clearly communicated to all stakeholders.

Part 4

HOW AI COMPANIES CAN BUILD A HEALTHY RISK CULTURE

Section insights

- Culture can be understood as "the way things are done around here." It is a critical component of effective risk management, and various mechanisms exist to improve it.
- AI companies appear to have some elements of a strong risk culture, but this needs to be built upon and operationalised. Currently, risk culture appears to vary across teams and individuals, consistent with there being no overarching framework for how to manage.
- There may be internal cultural barriers to improving risk management at AI companies, especially an aversion to perceived bureaucracy. These can be overcome by persuasively making the case about why these methods are effective at reducing risk and improving efficiency, and how they can be implemented in ways that are agile, don't require significant overheads, and are adaptable to specific organisational contexts.

Risk culture overview

What is risk culture? Organisational culture is colloquially known as "the way things are done around here." More officially, in a risk context, this is the set of collective values, attitudes, beliefs, norms, and behaviours regarding risk awareness, risk-taking, and risk management.[14]

Behaviours characterising a healthy risk culture:

- **Clear tone from the top:** Senior management and leaders actively promote a culture of risk awareness, transparency and accountability.
- **Openness and transparency:** Employees at all levels feel comfortable raising concerns and discussing potential risks. There's a culture of open communication and reporting of risks without fear of retribution or 'shooting the messenger'.
- **Effective cross-functional collaboration:** Risk management is seen as a shared responsibility across the organisation (ideally with a clear understanding of the differences between risk ownership, oversight and audit). There is strong collaboration and communication between teams, especially first line technical research and engineering teams and second line risk oversight teams.
- **Psychological safety:** Employees at all levels feel psychologically safe to express dissenting opinions, challenge assumptions and raise concerns about risk without fear of consequences. Diversity of thought and constructive debate are encouraged.
- **Continuous learning and improvement:** A culture of continuous learning and improvement, especially as it relates to risk management, supported by organisation-led initiatives relating to incident reporting, near misses and lessons learned.
- **External engagement:** An open and active engagement with external stakeholders, including industry peers and bodies, academics, policy makers and the public, to share best practices and collaborate on risk management initiatives.

[14] The Institute of Risk Management: Risk Culture - Resources for Practitioners.

Interventions to improve risk culture:

- **Speak-up channels:** The establishment and use of a 'speak up' channel and related procedures for confidential reporting of safety or risk concerns or suggestions, as well as an independent review process with board visibility. 'Speak up' is preferred to whistleblowing as branding here – it should enjoy the same protection but has a more general and positively perceived application. An appeal to external bodies such as regulators could be included in the process.
- **Cross-functional workshops**: Regular cross-functional risk identification workshops conducted at various seniority levels of the company under the Chatham House Rule. These help keep the culture risk-attuned and focused.
- **Risk management training:** Training for leaders and employees throughout business about how to manage risk within their roles; regular internal communications about risk reporting, facilitating the creation of a 'just culture' that encourages risk reporting and discourages a culture of blame or avoidance.
- **Surveys and pulse checks:** Regular risk culture pulse checks and surveys, with results reviewed by the board.

All of the above contribute to a dynamic risk culture. There's an opportunity for AI companies to graft this into their existing start-up culture in a way that keeps the risk management dynamic and avoids the sadly all-too-common static and bureaucratic application of risk registers and control self-assessments, which, as a result, become 'box-ticking' exercises and end up harming risk culture.

Finally, great care must be taken with incentives to ensure they align with safety and risk objectives and appetites. This is particularly the case where stock options are widely used in remuneration, and stockholders are present across different functions and levels of the business.

Building on the nascent risk culture in AI companies

1. **New risk governance mechanisms (like 3LoD) are most likely to succeed if they can build on the risk cultures that already exist within AI companies.** The stated mission of all these companies is to create safe AGI that benefits humanity. There have been various reports about the unique culture towards AI risk within these companies, and staff appear to be drawn to these places because of this. However, these risk cultures aren't necessarily supported by robust and holistic risk governance practices, and certain people and teams appear to care more than others. It's difficult to think of other industries where there is such an apparent focus on the risks of the core business offering but without the safeguards put in place to manage these. This is perhaps slightly changing recently with the introduction of Responsible Scaling Policies (RSPs) and other policy initiatives, though the overarching theme is that these are relatively ad hoc measures and not holistic. See the expanded quote below from a policy and governance employee from one of these companies:

   "Our overall approach to risk management is admittedly ad hoc. We would like to change this to make it more holistic, but it needs to be done in a way that's appropriate for our organisation. We find that any time we bring in external policies or frameworks, they never work. People need to be empowered to do their job better."

– Policy and Governance employee

2. A likely tension for attempts to operationalise the nascent safety culture and risk governance is a strong **aversion to bureaucracy** that appears to exist. A common refrain from talking to people at companies was that there might be significant resistance to new teams/names introduced, especially ones that seemed like risk management terms.

   "It's unlikely there will be appetite for new roles or terms that sound too much like risk management – better to frame these ideas within the context of roles that already exist. People want to move quickly, and teams change often"

   – Policy and Governance employee

3. As the above quote suggests, there is a possibility that the 3LoD might be interpreted as overly bureaucratic and burdensome if implemented ineffectively, slowing down/restricting a pivotal part of their success: their pace of action. The challenge for governments, regulators, and risk management advocates is how to advocate for these initiatives in a way that is framed as **promoting innovation, efficiency and progress.** The next section addresses how to overcome some of these potential barriers.

Potential barriers to introducing more structured risk management

Below are some high-level hypotheses we've developed from talking to people within these companies about why there may be barriers to implementing more structured risk management approaches. We imagine this could be helpful for governments in understanding barriers to success for voluntary commitments or legislation or for advocates within companies wanting to push for these initiatives internally.

Startup ethos

These companies' success comes from their ability to be agile and dynamic. Risk management is stereotypically seen as time-consuming, bureaucratic and burdensome and wouldn't work in a start-up context; if implemented, it might put these companies at a competitive disadvantage.

Can this be overcome?
Yes – 3LoD and other measures can be implemented in a relatively dynamic, agile way and could put companies at a competitive advantage by having safer products and being able to talk about this publicly. 3LoD can promote efficiency by overcoming the inefficiency of ad hoc arrangements, implicit assumptions (e.g. around risk tolerances), silos and blurred accountabilities. It doesn't always require the creation of new teams and roles; often, the most effective means of ensuring the effective implementation of these approaches is to train existing teams over time rather than introducing too much change at once.

A relevant example is the *fintech sector*, particularly the provision of payment services – these are subject to rapid change, relatively immature in terms of structure compared to established banks, high risk to the point that risk management becomes a differentiator, and where the challenge has been to tailor and rightsize the risk management framework whilst adhering to the same guiding principles.[15] Like in fintech, there are opportunities to scale the framework to match

[15] This paper from Oliver Wyman ("Right-Sizing Three Lines of Defense") is instructive in this regard.

the size of the business – for example, second line risk management and compliance functions can be combined, and joint board-level audit and risk committees can be established. Ultimately, the Three Lines Model will always need to be tailored to the specific context where it is being implemented, and an agile, staged approach is often the most effective.

AI exceptionalism

The risks from frontier AI are so novel (and potentially uncertain) that past risk management and governance approaches aren't fit for purpose or useful here.

Can this be overcome?
Yes – risk management is fundamentally about managing uncertainty, and while AI provides some new aspects, there doesn't seem to be anything particularly novel about AI risk that would justify a totally different approach to other industries. Indeed, if firms were to adopt more holistic risk management and governance, they would be better positioned to understand and respond to the changing nature of AI risk as capabilities improve.

Lack of knowledge/understanding

It appears that some companies haven't implemented a more holistic risk management approach because they don't know what this entails, and such approaches aren't common in startups.

Can this be overcome?
Yes – this requires making the case to these companies about why these methods are effective and strategic for their business. Case studies and applied examples will be important here, given 3LoD isn't common in startups.

The challenge for governments or risk management advocates is to tailor 3LoD and risk management methods to the business context where they are being implemented. While this piece has tried to generalise across these companies, effective implementation will ultimately require an understanding of the business with the aim of empowering people in their existing roles to better understand their risk responsibilities.

Part 5

LEGISLATIVE OPTIONS FOR MANDATING BEST PRACTICE RISK MANAGEMENT

Section insights

There's a case for AI companies to be required by law to adopt the principles of best practice risk management, including distinct risk and assurance functions, resilience reporting and external audit.

Legislative Options

Options to mandate best practice risk management at frontier AI companies range from prescribing a detailed 3LoD-based risk management framework and related methodology, to mandating broader risk management principles and methodology, to issuing a voluntary code of best practice. These could be achieved via primary legislation or through delegation to empowered regulators.

Ultimately given the degree of uncertainty, the scale of the potential risks to society, the sheer pace of development and the rapidly compounding financial incentives involved, we think there is a compelling case for AI companies to be required by law to adopt at least the principles of best practice risk management, including distinct risk and assurance functions, resilience reporting and external audit. In our view, this is likely to require primary legislation, at least initially, given the relative immaturity of the regulatory environment at this point.

Available examples to draw on here lie largely in the financial sector. A primary legislation approach could draw upon the requirements of the UK Companies Act (2006), the guidance of the UK Corporate Governance Code (2018, updated in 2024) and, in particular, the more stringent requirements of the UK Draft Companies (Amendment) Regulations of 2023. Although abandoned late last year, this draft legislation would have required companies to submit an annual resilience statement demonstrating the efficacy of their risk management process. The updated Code requires them to establish an office of risk management and an internal control function and review their effectiveness. These requirements would seem to be sensible to mandate for AI companies. Given the potential risks involved, the 'comply or explain' basis of the Corporate Governance Code would seem overly 'light touch' if applied to foundation model AI development.

A requirement to undergo external auditing would be a helpful addition to legislation; this could be delivered by a range of national regulators as well as by the industry itself (the example of the International Air Transport Association's (IATA's) Operational Safety Audit (IOSA) scheme in the aviation sector is worthy of study).

The challenge is to ensure the risk management process is dynamic and nimble and requirements don't mire developers in bureaucracy (which could be counter-productive to effective risk management, as this would likely encourage a 'box-ticking' approach to satisfying what's required).

Here it may be instructive to learn from the implementation of successful integrated safety management systems in different industries – for example, aviation – and emulate ways in which these have contributed to an organisation-wide risk-attuned culture whilst meeting rigorous regulatory reporting and external audit requirements.

Another contributor to a more dynamic and agile process in the private sector has been the use of regular cross-functional workshops conducted under the Chatham House Rule to review risk registers, identify and assess emerging risks and propose new mitigations. This can be combined

with a 'just culture'[16] approach to encourage reporting of safety issues. The example of 'crew resource management' in aviation could also provide inspiration in its encouragement of challenge in cockpit communications leading to improved safety outcomes within a hierarchical command structure.

The key to successfully implementing any recommended or mandated structure lies in maintaining a healthy and dynamic risk culture in which the risk management framework is embedded.

[16] See for example this paper outlining the approach as applied to a safety-sensitive industry

Part 6
RECOMMENDATIONS

Below are recommendations for governments and companies on how to ensure better AI risk management and governance. These apply across foundation model companies but are particularly relevant for companies developing frontier models.

For Government

Require AI companies to establish and maintain an office of risk management and an internal audit function, submit an annual resilience statement demonstrating the efficacy of their risk management process, undergo an annual external audit, and establish a protected 'speak up' channel (whistleblowing equivalent) with appeal to external bodies where necessary.

Build consensus within business and civil society about the importance of more holistic risk management, including a specific focus on risk governance. This could include publishing papers or facilitating workshops.

For companies

Build consensus

Encourage internal discussions about how best practice risk management (overarching risk management framework, 3LoD, appetite statements) might be useful, in particular, how clarifying risk ownership and reporting lines could reduce/help manage risks from AI.

Champion and sponsor best practice and dynamic risk management in the organisation by board members and senior executives acknowledging that this will require buy-in across the organisation to be effective.

Implement better practice – an eight-point checklist:

1. Encourage a stronger sense of risk ownership in research, product and engineering teams through workshops, training and engagement with specialist risk and internal audit functions.
2. Experiment with an MVP version of 3LoD structure and related methodology, and explore ways of sharing learnings with other companies (e.g. via the Frontier Model Forum).
3. Introduce an office of risk management with a central risk management team reporting to a Chief Risk Officer or equivalent to provide challenge, a degree of independent oversight, and risk reporting to the board.
4. Introduce an independent internal audit team to provide assurance on the process.
5. Formulate/agree on risk appetite statements based on thorough risk identification and assessment.

6. Introduce measures to encourage and enhance healthy risk culture: leadership and 'tone from the top', regular pulse checks/surveys with results reviewed by the board, emphasis on 'just culture', a protected 'speak up' channel with independent review and board visibility, and regular cross-functional risk identification workshops at different levels.

7. Seek external assurance on the risk management process and overall compliance via external audit.

8. Produce an annual resilience statement demonstrating the efficacy of the risk management process.

Overall, holistic risk management and governance at AI companies is in its infancy, exposing people and society to undue risk. While all of the above measures are unlikely to be applied immediately, governments should consider how to mandate them, and companies should consider how to implement them, potentially in a staged approach.

This report has attempted to more practically show how to improve risk governance at AI companies, but further work is needed from a range of actors including civil society and researchers (e.g. making the case for different aspects of risk management like annual reporting or risk appetite statements), governments (e.g. considering how these approaches best fit into an effective overall regulatory regime for AI) and companies (e.g. testing these methods and publicly sharing findings).

APPENDICES

Appendix A: Placement of AI company teams in Three Lines Model

The table below outlines how different teams within an AI company may fit within the Three Lines Model. The specific teams and descriptions are based on a general understanding of some of the functions currently existing in AI companies, but there may be some differences between organisations.

First line – risk ownership		
Role	**Description**	**Rationale for placement**
Chief Technology Officer (CTO)	Oversees the development and implementation of the company's technical strategy and roadmap.	The CTO, and roles below, are in the first line because they are responsible for delivering and operating products and owning the associated risks of these products. This includes risks from product design, delivery, and implementation, to risk related to technical, financial, and research processes.
Product development/ management	Defines the product vision and strategy for AI foundation models and related products.	
Program management	Plans and coordinates cross-functional projects and initiatives related to AI development and deployment.	
Engineering	Designs, develops, and maintains the software infrastructure and systems for AI foundation models.	
Chief Scientific Officer (CSO)	Leads the company's scientific research strategy and oversees the research and development (R&D) teams.	
Basic research / R&D	Conducts fundamental research to advance the state-of-the-art in AI and machine learning.	
Chief Financial Officer (CFO)	Oversees the company's financial strategy, planning, and operations.	
Mixed first and second line – ownership and oversight		
Role	**Description**	**Rationale for placement**
Information security	Develops and implements strategies to protect the company's data, intellectual property, and AI systems from cyber threats.	Provide oversight and support to ensure the company's AI systems and data are protected from cyber risks. This could be considered first and second line because they are directly involved in implementing and managing security controls, monitoring systems for potential breaches, and responding to security incidents (first line). They also establish security policies,

		standards, and guidelines and provide oversight and guidance to ensure that the company's AI systems and data are protected from cyber risks (second line).
Trust and safety systems	Designs and implements systems and processes to ensure the safe and responsible development and deployment of AI foundation models.	Provide oversight and support to ensure AI systems are developed and deployed in a safe and responsible manner, mitigating risks related to AI misuse and bias. This could be considered first and second line because they are directly involved in the development and deployment of AI systems, ensuring that safety and ethical considerations are integrated into the design and implementation process (first line), and they have an oversight function of establishing policies, guidelines, and best practices for the responsible development and deployment of AI systems (second line).
Technical safety / alignment	Conducts research and develops techniques to ensure AI foundation models are aligned with human values and goals.	Provide oversight and support to ensure AI systems are designed and implemented in a way that aligns with human values and mitigates risks related to reward hacking and scalable oversight. This could be considered first and second line because they are directly involved in the research, development, and implementation of AI systems, ensuring that technical alignment approaches are in place (first line). They also establish technical standards, guidelines, and best practices for ensuring the safety and alignment of AI systems and provide oversight and guidance to ensure that these standards are being followed (second line).
Second line – risk oversight		
Role	**Description**	**Rationale for placement**
Responsible scaling team	Develops strategies and processes for responsibly scaling AI foundation models and managing associated risks.	These are all second line roles because they provide an oversight and support function, ensuring that first line risk ownership roles are operating effectively. These roles collectively support the development and deployment of AI systems by helping to identify, assess, and mitigate any risks, while also ensuring adherence to best principles and ethical guidelines.
Governance team	Develops and implements governance frameworks and policies for the responsible development and deployment of AI foundation models.	
Safety committees	Provide oversight and guidance on AI safety and ethics issues.	

Extreme risk function	Identifies and assesses low-probability, high-impact risks related to advanced AI systems, such as existential risks.	
Legal and compliance teams	Ensure the company's AI development and deployment activities comply with relevant laws, regulations, and industry standards.	
Policy teams	Develop and advocate for policies that support the responsible development and deployment of AI foundation models.	
GRC	Develop and implement integrated strategies for managing governance, risk, and compliance issues related to AI.	
Ethics research	Conduct research on the ethical implications of AI foundation models and related technologies.	
Third line – risk audit		
Role	**Description**	**Rationale for placement**
Model evaluation	Design and conduct evaluations of AI foundation models to assess their performance, fairness, and safety.	Provide oversight and support to ensure AI models are thoroughly evaluated for performance, fairness, and safety, and areas for improvement are identified. They are in the third line because they ideally provide independent assurance about the efficacy of the risk management process rather than as a primary source of risk identification.
Red teaming	Conduct adversarial testing and simulated attacks on AI foundation models to identify vulnerabilities and weaknesses.	Provides independent assurance by conducting adversarial testing and simulated attacks to identify vulnerabilities and weaknesses in AI systems and collaborating with other teams to ensure a comprehensive approach to AI risk management. They are in the third line because they provide independent assurance of the efficacy of the risk management process by this process of vulnerability and weakness detection. While red-teaming has come to mean various things within the context of AI companies (including sometimes as an internal form of testing a specific system for risks, which would be a first line function), they ideally should be seen as independent assurance function rather than a risk identification function, and are therefore third line.

Appendix B: Preliminary gap analysis

A preliminary evaluation of the extent to which frontier AI companies are meeting the ideal risk activities for different lines in the Three Lines model.

Line	Activity	Implementation?	Explanation
First line	Risk ownership	Limited	There is little evidence of a strong risk culture and risk ownership in "the business" — i.e. the delivery teams of these organisations. While certain individuals and teams clearly care deeply about risk, the overarching culture of risk ownership and reporting lines appears to be ad hoc, fragmented and lacking structure. It's unclear the extent to which any of these teams see themselves as 'risk owners' or have clear Key Risk Indicators and reporting lines that enable best practice risk management.
Second line	Monitoring	Some evidence	There is some evidence of monitoring taking place. For example, OpenAI's Preparedness team aims to "track, evaluate, forecast and protect against catastrophic risks".
	Challenge	Limited	There is limited evidence that the business is sufficiently challenged by an independent oversight function. Anthropic's newly announced Alignment Stress-testing team may be positioned to provide some challenge, but it's unclear yet how effective this will be, including whether it's adequately independent.
	Advice	Some evidence	The policy and governance teams seem to have some ability to offer advice to the business. However, their level of influence is unclear. This often comes from having reporting lines to senior people.
	Policy creation	Substantial evidence	These companies all have policy teams that are regularly creating new policies. E.g., see the variety of policies outlined in responses by Anthropic, Google DeepMind and OpenAI prior to the UK AI safety summit. It's unclear the extent to which these policies are integrated into day-to-day operational decision-making in relevant teams or supported by an overarching risk management framework, but there is no doubt substantial evidence that policies are being created.
Third line	Aggregation	Missing	There does not appear to be any central risk team in these organisations that aggregates all risk information into a holistic picture.
	Reporting	Missing/Limited	Naturally, there is some level of reporting to senior management, but it is unclear how focused it is on risk. It's also unknown if the specialist risk and assurance functions get time with the board.
	Control advisory	Missing	This is normally performed by an internal audit team, which does not appear to exist in these organisations, so there is likely no control advisory in place.
	Internal assurance	Missing	There do not appear to be any internal audit teams in these organisations, which is typically quite synonymous with internal independent assurance. Anthropic's newly announced Alignment Stress-testing team might be seen as providing some assurance over the alignment teams.
	External assurance	Limited	There is some external independent assurance taking place since most of these organisations invite external parties to design evaluations and to conduct red-teaming on the models. This seems likely to be limited to just model capabilities at this time. Another area where external assurance is often provided is penetration testing in information security. The labs have not stated whether they use external parties for this, but it would be likely.

Appendix C: Three Lines Model keys to success and potential failure modes

Effectively implementing the Three Lines Model will require carefully considering the capabilities, structure, incentives and culture within each line. Below are keys to success and potential failure modes for each of these aspects.

First line		
	Keys to success	**Potential failure modes**
Capabilities	• Expertise in first line 'core' activities • Teams and individuals involved understand that if they are involved in decisions related to risk in any way, they are first line risk owners • Teams and individuals involved understand their responsibilities and accountabilities for risk ownership • Teams and individuals involved are active in mitigating risks and seeking the support and guidance of the second line in doing this	• Lack of understanding of how • decision-making roles relate to risk • Lack of risk awareness • No process for identification of emerging risks
Structure	• Reporting lines exclusively to first line management	• Blurring of ownership and accountability • Blurring of risk ownership and oversight roles (which undermines risk ownership and leads to conflicts in reporting lines, damaging checks and balances)
Incentives	• People within the first line are incentivised to take direct ownership of risks, and this is included within KPIs or performance conversations	• No risk responsibilities included within KPIs or other forms of formal or informal incentives
Culture	• Teams and individuals involved are involved in regular cross-functional risk identification and assessment workshops • A 'speak up' channel for risk concerns to be raised and independently reviewed	• No protected channel to raise concerns • A punitive/blame assignment approach to safety issues
Second line		
	Keys to success	**Potential failure modes**
Capabilities	• A degree of technical knowledge of first line activities, but effective risk oversight doesn't require subject-matter expertise in all areas • Specialist risk management expertise • Availability of data and metrics for monitoring • Imaginative questions and challenge	• Blurring of risk oversight and ownership roles (which undermines independence in oversight and leads to conflicts in reporting lines, damaging checks and balances) • Lack of specialist risk management expertise • Unclear risk tolerances or misalignment over these at the executive/board level
Structure	• A degree of independence in the oversight functions (for example, a Chief Risk Officer	• Fragmented risk teams with no overarching purview • Limited access across the organisation

	(CRO) with dual reports to an independent board or board risk/audit committee chair) • Access to all areas of the business • One office of risk management which oversees all areas of the organisation	• Lack of monitoring data and metrics • 'Policing'-style oversight vs supportive, guiding, influencing, constructive challenging, building trust
Incentives	• Strong executive sponsorship and board-level support and interest • Clear risk appetite statement(s) and agreed risk tolerances	• 'Box-ticking'-style risk management leading to a 'static' risk register and unnecessary/unpopular bureaucracy
Culture	• Dynamic vs static risk management: 'live' interactions and discussion of risk in the business vs 'box-ticking' culture	• Uninterested senior executive and board membership with limited/no sponsorship

Third line

	Keys to success	Potential failure modes
Capabilities	• Expertise in audit and internal audit reinforced by external assurance where helpful, including red-teaming	• Lack of expertise (technical or otherwise)
Structure	• Independent audit structure, including independently chaired audit committee and independent red-teaming input • Collaboration with second line to probe areas highlighted as posing potential issues and agreed mitigation actions • Access to all areas	• Lack of empowerment and senior/board sponsorship in the organisation • Audit team disconnected from or misaligned with second line risk oversight • Lack of external assurance support (this link talks about independent red-teaming)
Incentives	• Strong senior executive and board membership support, interest and involvement to empower the audit team	• 'Box-ticking'-style risk management leading to a 'static' risk register and unnecessary/unpopular bureaucracy • Lack of independence, for example if their reporting line is not fully independent
Culture	• Constructive challenging vs 'policing' and blaming, 'present' and actively engaged on the ground in the organisation	• Bureaucratic, static, box-ticking, passive approach

References

Anthropic, 'Anthropic's Responsible Scaling Policy', *Anthropic*, 2023, https://www.anthropic.com/news/anthropics-responsible-scaling-policy (accessed 29 April 2024).

Anthropic, 'Our Response to the UK Government's Internal AI Safety Policy Enquiries', *Anthropic*, 2023, https://www.anthropic.com/uk-government-internal-ai-safety-policy-response (accessed 29 April 2024).

Anthropic, 'Red Teaming and Model Evaluations', *Anthropic*, 2023, https://www.anthropic.com/uk-government-internal-ai-safety-policy-response/red-teaming-and-model-evaluations (accessed 29 April 2024).

Department for Science, Innovation & Technology, 'A pro-innovation approach to AI regulation: government response', *GOV.UK*, 2024, https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response (accessed 10 May 2024).

Department for Science, Innovation & Technology, 'AI Safety Institute Approach to Evaluations', *GOV.UK*, 2024, https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations (accessed 29 April 2024).

Department for Science, Innovation & Technology, 'Emerging processes for frontier AI safety', *GOV.UK*, 2023, https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety#responsible-capability-scaling (accessed 29 April 2024).

Field, H., 'OpenAI dissolves team focused on long-term AI risks, less than one year after announcing it', CNBC, 2024, https://www.cnbc.com/2024/05/17/openai-superalignment-sutskever-leike.html (accessed 15 June 2024).

Frontier Model Forum, 'Frontier Model Forum: Advancing Frontier AI Safety', *Frontier Model Forum*, 2024, https://www.frontiermodelforum.org/ (accessed 29 April 2024).

Google DeepMind, 'AI Safety Summit: An Update on Our Approach to Safety and Responsibility', *Google DeepMind*, 2023, https://deepmind.google/public-policy/ai-summit-policies/ (accessed 29 April 2024).

Government Finance Function and HM Treasury, 'Orange Book: Management of Risk - Principles and Concepts', *GOV.UK*, 2023, https://www.gov.uk/government/publications/orange-book (accessed 29 April 2024).

Independent Oversight Working Group, 'The UK Nuclear Industry Good Practice Guide To: Independent Oversight', *Nuclear Institute*, 2018, https://www.nuclearinst.com/write/MediaUploads/SDF%20documents/IOWG/SDF_sub-group_Good_Practice_Guide_Issue_2.pdf (accessed 29 April 2024).

Institute of Risk Management, 'Risk Culture: Resources for Practitioners', *IRM,* 2012, *https://www.google.com/url?q=https://www.theirm.org/media/7236/risk-culture-resources-for-practitioners.pdf&sa=D&source=docs&ust=1715086855130230&usg=AOvVaw24IBQ1llBP4X-tdD8pdrNc* (Accessed 29 April 2024).

Leike, J. and Sutskever, I., 'Introducing Superalignment', *OpenAI,* 2023, https://openai.com/blog/introducing-superalignment (accessed 29 April 2024).

Maritime and Coastguard Agency, 'A "Just Culture": Improving Safety and Organisational Performance', *GOV.UK*, 2014, https://www.gov.uk/government/publications/a-just-culture-improving-safety-and-organisational-performance (accessed 29 April 2024).

National Institute of Standards and Technology, 'Artificial Intelligence Risk 4 Management Framework: 5 Generative Artificial Intelligence 6 Profile', *NIST*, 2024 https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf (accessed 14 May 2024).

National Institute of Standards and Technology, 'AI Risk Management Framework', *NIST*, 2023, https://www.nist.gov/itl/ai-risk-management-framework (accessed 29 April 2024).

National Institute of Standards and Technology, 'AI RMF Playbook'', *NIST*, 2023, https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook (accessed 14 May 2024).

Narayanan, A. and Kapoor, S., 'AI Safety Is Not a Model Property', *AI Snake Oil,* 2024, https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property?r=ksgr7&utm_medium=ios&triedRedirect=true (accessed 29 April 2024).

Network Rail, 'Network Rail Limited's Annual Report and Accounts 2022', *Network Rail*, 2022, https://www.networkrail.co.uk/wp-content/uploads/2022/07/Network-Rail-Annual-report-and-accounts-2022.pdf (accessed 29 April 2024).

Oliver Wyman, 'Right-Sizing the Three Lines of Defense', *Oliver Wyman - Impact-Driven Strategy Advisors*, 2020, https://www.oliverwyman.com/our-expertise/insights/2020/jan/three-lines-of-defense.html (accessed 29 April 2024).

OpenAI, 'Careers: Machine Learning Engineer, Moderation', *OpenAI*, https://web.archive.org/web/20231120155223/https://openai.com/careers/machine-learning-engineer-moderation (accessed 29 April 2024).

OpenAI, 'OpenAI's Approach to Frontier Risk', *OpenAI*, 2023, https://openai.com/global-affairs/our-approach-to-frontier-risk (accessed 29 April 2024).

OpenAI, 'Preparedness', 2023, *OpenAI,* https://openai.com/safety/preparedness (accessed 29 April 2024).

Piers, C., 'Even Chatgpt Says Chatgpt Is Racially Biased', *Scientific American*, 2024, https://www.scientificamerican.com/article/even-chatgpt-says-chatgpt-is-racially-biased/ (accessed 29 April 2024).

'Report to the President By the Presidential Commission: On the Space Shuttle Challenger Accident ', *NASA*, 1986, https://sma.nasa.gov/SignificantIncidents/assets/rogers_commission_report.pdf (accessed 29 April 2024).

Schuett, J., 'Three Lines of Defense against Risks from AI', *AI & Society*, 2023, https://doi.org/10.1007/s00146-023-01811-0

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., & Dafoe, A., 'Model Evaluation for Extreme Risks', DeepMind, 2023, https://doi.org/10.48550/arXiv.2305.15324

The Institute of Internal Auditors, 'Applying the Three Lines Model in the Public Sector', *The Institute of Internal Auditors*, 2022, https://www.theiia.org/en/content/articles/2022/applying-the-three-lines-model-in-the-public-sector/ (accessed 29 April 2024).

The Institute of Internal Auditors, 'The IIA's three lines model: An update of the three lines of Defense', *The Institute of Internal Auditors: Position Papers*, 2020 https://www.theiia.org/en/content/position-papers/2020/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense/#content (Accessed: 29 April 2024).

Weidinger, L. and Isaac, W., 'Evaluating Social and Ethical Risks from Generative AI', *Google DeepMind*, 2023, https://deepmind.google/discover/blog/evaluating-social-and-ethical-risks-from-generative-ai/ (accessed 29 April 2024).

Wiggers, K., 'Anthropic hires former OpenAI safety lead to head up new team', *TechCrunch*, 2024, https://techcrunch.com/2024/05/28/anthropic-hires-former-openai-safety-lead-to-head-up-new-team/?guccounter=1 (accessed 15 June 2024).