AI & SOCIETY
https://doi.org/10.1007/s00146-023-01811-0

MAIN PAPER

Three lines of defense against risks from AI

Jonas Schuett[1,2,3]

Received: 15 April 2023 / Accepted: 31 October 2023

Abstract
Organizations that develop and deploy artificial intelligence (AI) systems need to manage the associated risks—for economic, legal, and ethical reasons. However, it is not always clear who is responsible for AI risk management. The three lines of defense (3LoD) model, which is considered best practice in many industries, might offer a solution. It is a risk management framework that helps organizations to assign and coordinate risk management roles and responsibilities. In this article, I suggest ways in which AI companies could implement the model. I also discuss how the model could help reduce risks from AI: it could identify and close gaps in risk coverage, increase the effectiveness of risk management practices, and enable the board of directors to oversee management more effectively. The article is intended to inform decision-makers at leading AI companies, regulators, and standard-setting bodies.

Keywords Artificial intelligence · Risk management · Three lines of defense · Internal audit

Abbreviations

3LoD	Three lines of defense
4LoD	Four lines of defense
5LoA	Five lines of assurance
AI	Artificial intelligence
API	Application programming interface
BCBS	Basel Committee on Banking Supervision
CAE	Chief audit executive
CCO	Chief compliance officer
CEO	Chief executive officer
CFO	Chief financial officer
CLO	Chief legal officer
COSO	Committee of Sponsoring Organizations of the Treadway Commission
CRO	Chief risk officer
CSO	Chief scientific officer
CTO	Chief technology officer
EBA	European Banking Authority
ERM	Enterprise risk management
IEC	International Electrotechnical Commission
IIA	Institute of Internal Auditors
ISO	International Organization for Standardization
KPI	Key performance indicator
NIST	National Institute of Standards and Technology
OECD	Organisation for Economic Co-operation and Development
PAI	Partnership on AI
RQ	Research question
RSP	Responsible scaling policy
SEC	Securities and Exchange Commission

1 Introduction

Organizations that develop and deploy artificial intelligence (AI) systems need to manage the associated risks—for economic reasons, because accidents and cases of misuse can threaten business performance (Cheatham et al. 2019); for legal reasons, because upcoming AI regulation might require them to implement a risk management system (Schuett 2023a); and for ethical reasons, because under most moral theories they have an obligation to prevent harm (Mohamed et al. 2020; Hagendorff 2022; Bengio et al. 2023).

However, it is not always clear who is responsible for AI risk management: the researchers and engineers? The legal and compliance department? The governance team? The three lines of defense (3LoD) model might offer a solution. It is a risk management framework intended to improve an organization's risk governance (van Asselt 2011; Lundqvist

✉ Jonas Schuett
jonas.schuett@governance.ai

[1] Centre for the Governance of AI, Oxford, UK

[2] Legal Priorities Project, Cambridge, MA, USA

[3] Faculty of Law, Goethe University Frankfurt, Frankfurt a.M., Germany

Published online: 27 November 2023

2015) by assigning and coordinating risk management roles and responsibilities (Institute of Internal Auditors [IIA], 2013, 2020a). It is considered best practice in many industries, such as finance and aviation. In this article, I apply the 3LoD model to an AI context.

To date, there has not been much academic work on the intersection of AI and the 3LoD model. Nunn (2020) suggests using the model to reduce discrimination risks from AI, but the relevant passage is very short. There is also some literature on how companies could use AI to support the three lines (Tammenga 2020; Sekar 2022), but I am mainly interested in how to govern AI companies, not how to use AI to govern non-AI companies. It has also been proposed that governments could use the 3LoD model to manage extreme risks from AI (Ord 2021), but here I focus on the challenges of companies, not government.

While academic scholarship on this topic may be limited, there is some relevant work from practitioners. Most notably, there is a blog post by PwC that seeks to answer questions similar to this article (Rao and Golbin 2021). But since they only dedicate a short section to the 3LoD model, their proposal only scratches the surface. The IIA has also published a three-part series, in which they propose an AI auditing framework (IIA 2017a, 2017c, 2018). Although their proposal contains a reference to the 3LoD model, it does not play a key role. Finally, the 3LoD model is mentioned in a playbook that the National Institute of Standards and Technology (NIST) published alongside the AI Risk Management Framework (NIST 2023a). However, the playbook only suggests implementing the 3LoD model (or a similar mechanism); it does not specify how to do so.

Taken together, there are at least two gaps in the current literature. The first one is practical: there does not seem to be a concrete proposal for how organizations that develop and deploy AI systems could implement the 3LoD model. The few proposals that exist are not detailed enough to provide meaningful guidance. The second one is normative: there does not seem to be a thorough discussion about whether implementing the model is even desirable. Given that the model has been criticized and there is not much empirical evidence for its effectiveness, the answer to this question is not obvious. In light of this, the article seeks to answer two research questions (RQs):

RQ1: How could organizations that develop and deploy AI systems implement the 3LoD model?

RQ2: To what extent would implementing the 3LoD model help to reduce risks from AI?

The article has three areas of focus. First, it focuses on organizations that develop and deploy state-of-the-art AI systems,[1] in particular medium-sized research labs (e.g. Google DeepMind and OpenAI) and big tech companies (e.g. Microsoft and Meta), though the boundaries between the two categories are blurry (e.g. Google DeepMind is a subsidiary of Alphabet and OpenAI has a strategic partnership with Microsoft). In the following, I use the term "AI companies" to refer to all of them. I do not cover other types of companies (e.g. hardware companies), nonprofits, or academic institutions, but they might also benefit from my analysis. Second, the article focuses on the organizational dimension of AI risk management. It is not about how AI companies should identify, assess, and respond to risks from AI. Instead, it is about how they should assign and coordinate risk management roles and responsibilities. Third, the article focuses on the model's ability to prevent societal harm (Smuha 2021). I am less interested in risks to companies themselves (e.g. litigation or reputation risks), though occasionally private and public interests are aligned (e.g. one way to reduce litigation risks is to prevent accidents).

The remainder of this article proceeds as follows. Section 2 gives an overview of the model's basic structure, history, criticisms, and evidence base. Section 3 suggests ways in which AI companies could implement the model. Section 4 discusses how the model could help to reduce risks from AI. Section 5 concludes and suggests questions for further research.

2 The 3LoD model

In this section, I give an overview of the basic structure (Sect. 2.1) and history of the 3LoD model (Sect. 2.2). I also engage with some of the main criticisms, briefly discuss alternative models (Sect. 2.3), and review the empirical evidence for its effectiveness (Sect. 2.4).

2.1 Basic structure

There are different versions of the 3LoD model. Most practitioners and scholars are familiar with the version published by the IIA (2013). After a review process, they published an updated version (IIA 2020a), which increasingly replaces the original version. This article will mainly use the updated

[1] There are many different terms that emphasize different features of such systems: the term "foundation model" highlights a model's role in the supply chain (Bommasani et al. 2021); "general-purpose AI system (GPAIS)" puts more emphasis on the generality of its capabilities (Barrett et al. 2023); "generative AI system" focuses on its output (Cao et al. 2023); and "frontier AI system" defines AI systems relative to existing capabilities (Anderljung et al. 2023; Shevlane et al. 2023). For the purposes of this paper, a precise definition is not necessary.

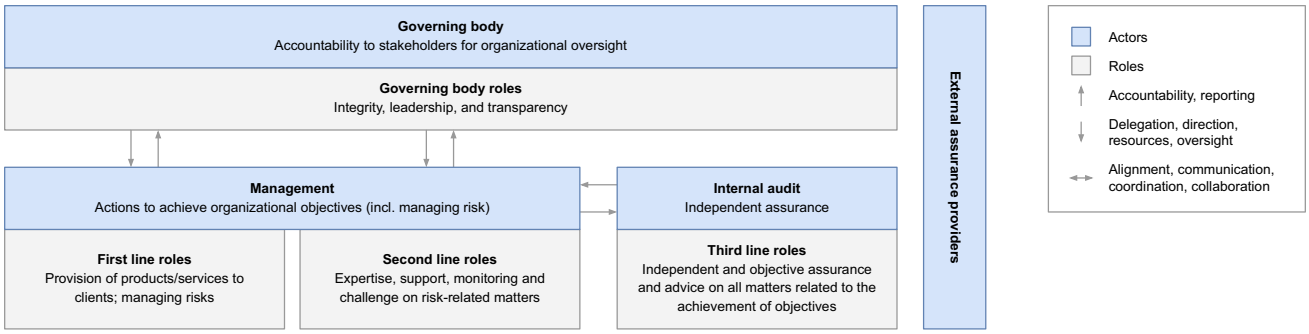

Fig. 1 The 3LoD model as described by the IIA (2020a)

version, as illustrated in Fig. 1. The updated model has three types of elements: actors, roles, and relationships.

The model distinguishes between four actors, represented as blue boxes: the governing body, which is accountable to stakeholders for organizational oversight; management, which takes actions to achieve the organization's objectives; internal audit, which provides independent assurance to the governing body, as do external assurance providers.

The model further distinguishes between four roles, represented as gray boxes. The role of the governing body is to demonstrate integrity, leadership, and transparency. In addition to that, the model contains three roles which it calls "lines of defense". The first line provides products and services to clients and manages the associated risks. The second line assists the first line with regards to risk management. It provides complementary expertise and support, but also monitors and challenges risk management practices. The third line provides independent and objective assurance and advice on all matters related to the achievement of risk objectives. The first two lines are part of management, while the third line is synonymous with internal audit.

Finally, there are three types of relationships between different actors, represented as arrows. There are top-down relationships: the governing body delegates responsibility to management and oversees internal audit. Inversely, there are bottom-up relationships: management and internal audit are accountable and report to the governing body. And lastly, there is a horizontal relationship between actors whose work must be aligned, namely between management and internal audit.

2.2 Brief history

The model's origins are opaque. There are theories suggesting military, sporting, or quality control origins (Davies and Zhivitskaya 2018). It was presumably developed in the late 1990s or early 2000s. In 1999, the Basel Committee on Banking Supervision (BCBS) suggested a similar approach to risk oversight (BCBS 1999), but the first explicit mention of the model was probably in a report by the UK Financial Services Authority (2003) or a paper by Roman Kräussl (2003).

After the financial crisis of 2007–2008, which was partly caused by widespread risk management failures (Boatright 2016), the model's popularity skyrocketed. In response to the crisis, regulators and supervisory authorities paid increasing attention to the chief risk officer (CRO) and the risk committee of the board (Walker 2009; Davies and Zhivitskaya 2018), and started recommending the 3LoD model (BCBS 2012; European Banking Authority [EBA], 2021). Most academic work on the model was also done after the crisis (e.g. Davies and Zhivitskaya 2018; Bantleon et al. 2021) and many risk management professionals only heard about the model in its aftermath (Zhivitskaya 2015).

Today, most listed companies have implemented the 3LoD model. In a 2015 survey of internal audit professionals in 166 countries (n = 14,518), the majority of respondents (75%) reported that their organization follows the 3LoD model as articulated by the IIA (Huibers 2015).[2] Another survey, conducted in 2021 among chief audit executives (CAEs) in Austria, Germany, and Switzerland (n = 415), supports their findings (Bantleon et al. 2021). The majority of respondents (88%) reported that they had implemented the model, with particularly high adoption rates among financial institutions (96%).

In contrast, big tech companies do not seem to have implemented the 3LoD model. It is not mentioned in any of their filings to the US Securities and Exchange Commission (SEC) or other publications. The model is also not explicitly mentioned in the corporate governance requirements by Nasdaq (2022), where all big tech companies are listed. It is worth noting, however, that the risk oversight practices at big tech companies do have some similarities with the 3LoD

[2] Note that respondents who said they were not familiar with the model were excluded.

model. For example, they all seem to have an internal audit function (e.g. Microsoft 2022; Alphabet 2022). Based on public information, medium-sized AI research labs do not seem to have implemented the model either.

2.3 Criticisms and alternative models

Despite the model's popularity in many industries, it has also been criticized (Arndorfer and Minto 2015; Zhivitskaya 2015; Davies and Zhivitskaya 2018; Hoefer et al. 2020; Vousinas 2021). Arndorfer and Minto (2015) identify four weaknesses and past failures of the 3LoD model. First, they argue, the incentives for risk-takers in the first line are often misaligned. When facing a tradeoff between generating profits and reducing risks, they have historically been incentivized to prioritize the former. Second, there is often a lack of organizational independence for second line functions. They are too close to profit-seekers, which can lead to the adoption of more risk-taking attitudes. Third, second line functions often lack the necessary skills and expertise to challenge practices and controls in the first line. And fourth, the effectiveness of internal audit depends on the knowledge, skills, and experience of individuals, which might be inadequate. Another common criticism is that the model provides a false sense of security. Put simply, "when there are several people in charge—no one really is" (Davies and Zhivitskaya 2018). Another criticism is that the model is too bureaucratic and costly. Additional layers of oversight might reduce risk, but they come at the cost of efficiency (Zhivitskaya 2015). A final criticism is that the model depends on information flow between the lines, but there are many barriers to this. For example, the second line might not recognize that they only see what the first line chooses to show them (Zhivitskaya 2015). While these criticisms identify relevant shortcomings and should be taken seriously, they do not put into question the model as a whole. Moreover, the 3LoD model has been improved over the years. Today, the focus is on increasing the model's effectiveness and responding to criticisms (Davies and Zhivitskaya 2018).

In view of these criticisms, several alternative models have been suggested. For example, Arndorfer and Minto (2015) proposed the four lines of defense (4LoD) model to better meet the needs of financial institutions. The fourth line consists of supervisory authorities and external audit, who are supposed to work closely with internal audit. Another example is the five lines of assurance (5LoA) model, which was gradually developed by several scholars and organizations (Leech and Hanlon 2016). However, the proposed changes do not necessarily improve the model. It has been argued that adding more lines would over-complicate the model, and that firms and regulators currently do not want structural changes (Davies and Zhivitskaya 2018). It is also worth noting that the alternative models are far less popular than the original model. Compared to these alternative models, the 3LoD model remains "the most carefully articulated risk management system that has so far been developed" (Davies and Zhivitskaya 2018). But what empirical evidence do we have for its effectiveness?

2.4 Empirical evidence

By "effectiveness", I mean the degree to which the model helps organizations to achieve their objectives. For the purpose of this article, I am mostly interested in the achievement of risk objectives. This may include: (1) reducing relevant risks to an acceptable level, (2) ensuring that management and the board of directors are aware of the nature and scale of key risks, and (3) compliance with relevant risk regulations. I am less interested in other objectives (e.g. improving financial performance), though there might be overlaps (e.g. reducing the risk of harm to individuals might also reduce the risk of financial losses from litigation cases). For an overview of different ways to measure the effectiveness of internal audit, see Rupšys and Boguslauskas (2007), Savčuk (2007), and Boţa-Avram and Palfi (2009).

There do not seem to be any (high-quality) studies on the effectiveness of the 3LoD model in the above-mentioned sense.[3] There only seems to be evidence for the effectiveness of internal audit (Lenz and Hahn 2015; Eulerich and Eulerich 2020). For example, a survey of CAEs at multinational companies in Germany (n = 37) compared audited and non-audited business units within the same company (Carcello et al. 2020). They found that managers of audited units perceive a greater decline in risk compared to managers of non-audited units. Other studies find that internal audit helps to strengthen internal control systems (Lin et al. 2011; Oussii and Taktak 2018) and has a positive influence on the prevention and identification of fraud (Coram et al. 2008; Ma'ayan and Carmeli 2016; Drogalas et al. 2017). The fact that the 3LoD model was not able to prevent past scandals and crises seems to provide weak evidence against its effectiveness (though another explanation could be that the model was poorly implemented in these cases), while the model's ongoing popularity seems to provide weak evidence in favor of its effectiveness (though the model's popularity

[3] There is also not much evidence on the model's effectiveness based on other interpretations of effectiveness. The only exception seems to be a recent study of the 500 largest companies in Denmark, which finds that a higher degree of adherence to first and second line practices is positively associated with financial performance (Andersen et al. 2022). Besides that, there are only studies on the effects of internal audit (Lenz and Hahn 2015; Eulerich and Eulerich 2020; Jiang et al. 2020), none of which mentions the 3LoD model.

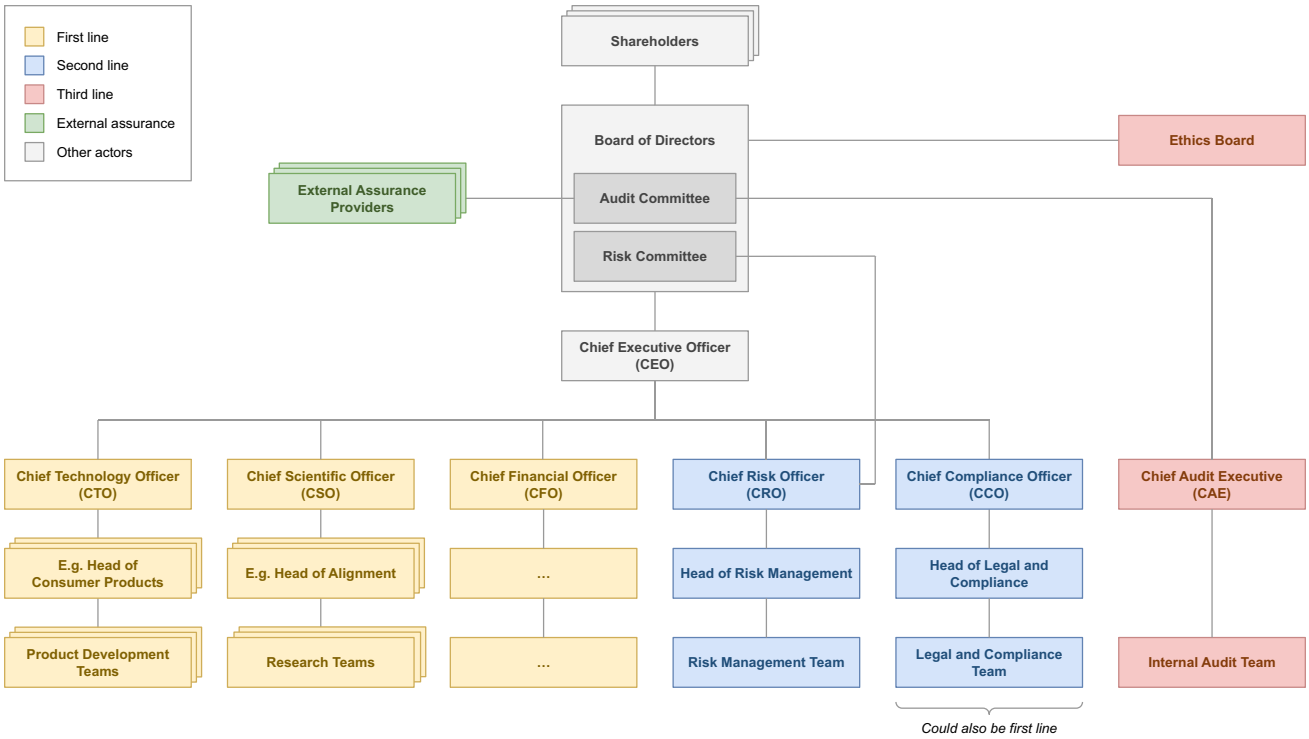

Fig. 2 Sample org chart of an AI company with equivalent responsibilities for each of the three lines

could also be explained by path dependencies). Finally, there is anecdotal evidence in both directions (Zhivitskaya 2015).

Overall, despite the model's popularity, "its effectiveness [remains] untested" (Davies and Zhivitskaya 2018) and "not based on any clear evidence" (Power et al. 2013). To be clear, it is not the case that we have robust evidence that the model is ineffective. It is still very plausible that the model can be effective, but there have not been (high-quality) studies providing empirical evidence for its effectiveness in the above-mentioned sense.

This surprising lack of evidence could potentially be explained by the following reasons. First, since it is not feasible to run randomized controlled trials on organizational interventions, it is inherently difficult to collect robust evidence. Second, the model is designed to be flexible and adaptable, which means that there is not a single, standardized way to implement it. This lack of standardization can make it difficult to compare different implementations of the model and to assess their effectiveness. Third, since most practitioners mainly care about financial performance, scholars might be incentivized to focus on economic measures of effectiveness to justify the relevance of their work (though there is not much evidence on that either).

Even if we had more empirical evidence from other industries, its informative value might still be limited. One reason is that findings might not generalize to an AI context. AI companies are structurally different from other companies because they have a special focus on research, and, since AI is a general-purpose technology (Crafts 2021; Garfinkel 2022), risks from AI are broader than risks from other products and services. Another reason is that the biggest driver of the model's ability to reduce risks is likely the concrete way in which it is implemented. So instead of asking "is the 3LoD model effective?", AI companies should ask "how can we implement the model in an effective way?".

3 Applying the 3LoD model to an AI context

This section suggests ways in which AI companies could implement the 3LoD model. For each of the three lines, I suggest equivalent roles and responsibilities. First, I describe the content of their responsibilities, then I discuss which team or individual would be responsible, as illustrated in Fig. 2.

3.1 First line

The first line has two main responsibilities: providing products and services to clients, which corresponds to AI research and product development, and managing the associated risks. Below, I focus on the latter.

The first line is responsible for establishing and maintaining appropriate structures and processes for the management of risk. This involves measures along all steps of the risk management process (NIST 2023b; International Organization for Standardization [ISO] 2018). To identify risks, the first line could use risk taxonomies (Weidinger et al. 2021, 2023; Raji et al. 2022a, b; Shelby et al. 2022), incident databases (McGregor 2021; Organisation for Economic Co-operation and Development [OECD], 2023), or scenario analyses (International Electrotechnical Commission [IEC], 2019; Koessler and Schuett 2023). To estimate the likelihood and impact of the identified risks, they might conduct probabilistic risk assessments, Delphi studies, or use risk matrices (IEC 2019; Koessler and Schuett 2023). These estimates will typically be informed by model evaluations (Chen et al. 2021; Perez et al. 2022b; Liang et al. 2022; Gehrmann et al. 2022), potentially with a focus on dangerous model capabilities (Shevlane et al. 2023; Kinniment et al. 2023; Alaga and Schuett 2023), and an assessment of the company's safeguards (O'Brien et al. 2023; Koessler and Schuett 2023). To mitigate risks, the first line could fine-tune the model on a curated dataset (Solaiman and Dennison 2021), via reinforcement learning from human feedback (RLHF) (Christiano et al. 2017; Ziegler et al. 2019; Lampert et al. 2022), or reinforcement learning from AI feedback (RLAIF), more commonly known as "constitutional AI" (Bai et al. 2022). To prevent leakage or theft of the model weights, the first line might take measures to strengthen the company's information security (Anthropic 2023c; Schuett et al. 2023a, b). And to prevent misuse, they could introduce a policy for the publication of potentially harmful research (Partnership on AI [PAI], 2021; Solaiman et al. 2019), or only grant access to models via an application programming interface (API) (Shevelane 2022; Solaiman 2023; Seger et al. 2023).

It might also make sense to take a more holistic approach and implement an AI-specific risk management framework (e.g. NIST 2023b; ISO and IEC 2023) or customize a more general enterprise risk management (ERM) framework (e.g. ISO 2018; Committee of Sponsoring Organizations of the Treadway Commission [COSO], 2017). Several organizations provide guidance on how to apply those frameworks to the specific needs of frontier AI developers (NIST, 2023c; PAI 2023; Barrett et al. 2022, 2023). In recent months, it has also become common to create specific policies for the responsible development and deployment of frontier AI systems, known as "responsible scaling policies" (ARC Evals 2023; Anthropic 2023a) or "risk-informed deployment policies" (OpenAI 2023a). For most of the above-mentioned measures, the first line needs support from the second line (see below).

The first line is also responsible for ensuring compliance with legal, regulatory, and ethical expectations. Legal obligations might stem from anti-discrimination law (Hacker 2018; Wachter et al. 2021), data protection law (Hamon et al. 2022), or antitrust law (Petit 2017; Hua and Belfied 2021). A notable example of AI regulation is the proposed EU AI Act (European Commission 2021), which requires providers of high-risk AI systems to implement a risk management system (Schuett 2023a). Ethical expectations might stem from AI ethics principles that organizations have adopted on a voluntary basis (Jobin et al. 2019; Hagendorff 2020). To ensure compliance, the first line relies on support from the second line (see below).

Finally, the first line is responsible for informing the governing body about the outcomes of the above-mentioned measures, the degree to which risk objectives are met, and the overall level of risk. This should take the form of a continuous dialogue, including reporting about expected and actual outcomes. Reports will typically include risk registers and risk matrices (IEC 2019), but they could also involve information about specific models, in the form of (preliminary) model cards (Mitchell et al. 2019), data sheets (Gebru et al. 2021), and system cards (Green et al. 2022). Note that there should also be a reporting line from the CRO to the chief executive officer (CEO) and the risk committee of the board (see below).

Responsible are operational managers, often in a cascading responsibility structure. At big tech companies, the lowest level of responsibility would lie with those managers who are in charge of the development of individual AI products. If there is no stand-alone AI product and AI systems make up only part of a product (e.g. WaveNet as a part of Google Assistant), then the lowest level of responsibility would lie with those managers who lead the development of the AI part of the product (e.g. the research lead for WaveNet). At medium-sized research labs, the lowest level of responsibility for risk management would lie with research leads, i.e. senior researchers who are in charge of individual research projects.

There will usually be one or more intermediate levels of responsibility. This might include a number of mid-level managers responsible for broader product areas (e.g. gaming) or research areas (e.g. reinforcement learning), though the details depend on the particular organizational structures. The ultimate responsibility for AI risk management lies with those C-suite executives who are responsible for product development (e.g. the chief technology officer [CTO]) or research (e.g. the chief scientific officer [CSO]). While it is possible to split responsibilities between two or more executives, this is often not advisable, mainly because it can dilute responsibilities.

3.2 Second line

The second line is responsible for assisting the first line with regards to risk management. It provides complementary expertise and support, but also monitors and challenges risk management practices.

Some risk management activities require special expertise that the first line does not have. This might include legal expertise [e.g. how to comply with the risk management requirements set out in the proposed EU AI Act (Schuett 2023a, b)], technical expertise [e.g. how to evaluate dangerous model capabilities (Shevlane et al. 2023; Kinniment et al. 2023) or develop more truthful language models (Evans et al. 2021)], or ethical expertise [e.g. how to define normative thresholds for fairness (Kleinberg, et al. 2016)]. It might also include risk-specific expertise [e.g. what risks language models pose (Weidinger et al. 2021)] or risk management-specific expertise [e.g. best practices for red teaming safety filters (Rando et al. 2022)]. The second line could support the first line by drafting policies, processes, and procedures, as well as frameworks, templates, and taxonomies. It might also advise on specific issues [e.g. how to customize a risk management framework to better meet the specific needs of the company (Barrett et al. 2022)], provide general guidance (e.g. how to ensure compliance with safety-related policies among researchers and engineers), or offer training (e.g. how to process training data in a GDPR compliant way).

The second line is also responsible for monitoring and challenging the adequacy and effectiveness of risk management practices. Risk management practices are ineffective if risk objectives are not met (e.g. the company fails to comply with relevant laws and regulations, or it is unable to reduce risks to an acceptable level). They are inadequate if the same results could have been achieved with fewer resources. The second line will typically use a number of key performance indicators (KPIs) to evaluate various dimensions of the adequacy and effectiveness of risk management (e.g. number of identified risks, number of incidents, or percentage of personnel trained on specific matters).

Second line responsibilities are split across multiple teams. This typically includes the risk management team as well as the legal and compliance team. Although most big tech companies already have a risk management team, these teams are mostly concerned with business risks (e.g. litigation or reputation risk). Risks from AI, especially societal risks, are usually not a major concern (Smuha 2021). If big tech companies want to change this, they could expand the responsibilities of existing teams. Setting up a new AI-specific risk management team seems less desirable, as it could lead to a diffusion of responsibilities. There would likely be a cascading responsibility structure where the CRO acts as the single point of accountability for the risk management process. Medium-sized research labs usually do not have a dedicated risk management team. A notable exception is OpenAI's new Preparedness team (OpenAI 2023b). They could either set up a new team or task one or more people in other teams with risk management-related support functions.

All AI companies beyond the early startup phase have a legal and compliance team. The team lead, and ultimately the chief compliance officer (CCO) or chief legal officer (CLO), would be responsible for risk-related legal and compliance support. It is worth noting that the legal and compliance team can also be part of the first line if they are actually responsible for ensuring compliance. They are part of the second line if they do not have any decision power and only support the first line (e.g. by writing legal opinions). The legal and compliance team can also seek support from external law firms.

Many organizations that develop and deploy AI systems have other teams that could take on second line responsibilities. This might include technical safety, ethics, policy, or governance teams. However, in practice, these teams rarely consider themselves as being responsible for risk management. This needs to be taken into account when implementing the 3LoD model (e.g. by running workshops to sensitize them to their widened responsibility). In general, AI companies should arguably avoid assigning second line responsibilities to them.

3.3 Third line

The third line is responsible for providing independent assurance. It assesses the work of the first two lines and reports any shortcomings to the governing body.

While the second line already monitors and challenges the adequacy and effectiveness of the risk management practices, the third line independently assesses their work—they supervise the supervisors, so to speak. They could do this by conducting interviews (e.g. with research leads) and attending meetings (e.g. regular meetings of development teams) (Schuett 2023b). They could also conduct internal audits (Raji et al 2020) or commission external audits (Buolamwini and Gebru 2018; Mökander and Floridi 2022; Raji et al. 2022a, b). Such audits could have different purposes and scopes (Mökander et al. 2023). They could evaluate compliance with laws, standards, or ethics principles ("compliance audit") or seek to identify new risks in a more open-ended fashion ("risk audit"). They could also assess the model itself, including the dataset it was trained on ("model audit"), the model's impact ("impact audit"), or the company's governance ("governance audit"). Similarly, the third line could engage a red team before or after a model is deployed to assess if the first two lines were

able to identify all relevant risks (Ganguli et al. 2022; Perez et al. 2022a). In addition to that, the third line could review key policies and processes to find flaws and vulnerabilities (e.g. a company's responsible scaling policy [ARC Evals 2023; Anthropic 2023a] or their deployment protocol). Note that this should also include a meta-assessment of the company's implementation of the 3LoD model itself.

The third line also supports the governing body, typically the board of directors, by providing independent and objective information about the company's risk management practices (IIA 2020b; Schuett 2023b). Their main audience is usually the audit committee, which is mainly composed of non-executive directors. But since non-executive directors only work part-time and heavily depend on the information provided to them by the executives, they need an independent ally in the company to effectively oversee the executives (Davies & Zhivitskaya 2018). The third line serves this function by maintaining a high degree of independence from management and reporting directly to the governing body following best practices. It is often described as their "eyes and ears" (IIA 2020a).

The third line has a well-defined organizational home: internal audit. Note that, in this context, internal audit refers to a specific organizational unit (Schuett 2023b). It does not merely mean an audit that is done internally (Raji et al 2020). Instead, it means "those individuals operating independently from management to provide assurance and insight on the adequacy and effectiveness of governance and the management of risk (including internal control)" (IIA 2020a).

Typically, companies have a dedicated internal audit team, led by the CAE or Head of Internal Audit. Most big tech companies have such a team, but similar to the risk management team, they often neglect the societal risks from AI. Instead of creating a separate AI-specific internal audit team, they should create a sub-team within their existing internal audit team, or simply task one or more team members to focus on AI-specific risk management activities. Medium-sized research labs usually do not have an internal audit team. They would have to create a new team or task at least one person with third line responsibilities. In short, big tech companies need to "bring AI to internal audit", while research labs need to "bring internal audit to AI". It is worth noting that, although there are promising developments (IIA 2017a, 2017c), the profession of AI-specific internal auditors is still in its infancy.

Some AI companies have an ethics board (e.g. Microsoft's Aether Committee and Meta's Oversight Board) which could also take on third line responsibilities, typically in addition to internal audit (Schuett et al. 2023b; Schuett 2023b). It would have to be organizationally independent from management, but still be part of the organization (in contrast to external assurance providers). If organizations already have an independent ethics board (e.g. consisting of representatives from academia and civil society), they could form a working group that takes on third line responsibilities.

4 How the 3LoD model could help to reduce risks from AI

While there are many reasons why AI companies may want to implement the 3LoD model, this section focuses on three arguments about the model's ability to prevent individual, collective, and societal harm: the model could help to reduce risks from AI by identifying and closing gaps in risk coverage (Sect. 4.1), increasing the effectiveness of risk management practices (Sect. 4.2), and enabling the governing body to oversee management more effectively (Sect. 4.3). I also give an overview of other benefits (Sect. 4.4). It is worth noting that, in the absence of robust empirical evidence (see above), the following discussion remains theoretical and often relies on abstract plausibility considerations.

4.1 Identifying and closing gaps in risk coverage

AI risk management involves different people from different teams with different responsibilities (Baquero et al. 2020). If these responsibilities are not coordinated adequately, gaps in risk coverage can occur (Bantleon et al. 2021). Such gaps may have different causes. For example, it might be the case that no one is responsible for managing a specific risk (e.g. there could be a blind spot for diffuse risks), or it might be unclear who is responsible (e.g. two teams might incorrectly assume that the other team already takes care of a risk). Gaps could also occur if the responsible person is not able to manage the risk effectively (e.g. because they do not have the necessary expertise, information, or time). If a specific risk is not sufficiently covered by the risk management system, it cannot be identified, which might result in an incorrect risk assessment (e.g. the total risk of an unsafe AI system is judged acceptable) and an inadequate risk response (e.g. an unsafe AI system is deployed without sufficient safety precautions).

The 3LoD model could prevent this by identifying and closing gaps in risk coverage. It could do this by offering a systematic way to assign and coordinate risk management-related roles and responsibilities. It ensures that people who are closest to the risk are responsible for risk management (first line) and get the support they need (second line). Another way the 3LoD model can help identify blind spots is through the internal audit function (third line). They are responsible for assessing the adequacy and effectiveness of the entire risk management regime, which includes potential gaps in risk coverage.

One might object that, in practice, gaps in risk coverage are rare, and even if they occur, they only concern minor risks (e.g. because AI companies have found other ways to address the biggest risks). However, the AI Incident Database (McGregor 2021) contains numerous entries, including several cases classified as "moderate" or "severe", which indicates that incidents are not that uncommon. While these incidents had many different causes, it seems plausible that at least some of them were related to gaps in risk coverage. But since there does not seem to be any public data on this, the issue remains speculative.

Even if one thinks that gaps in risk coverage are a common problem among AI companies, one might question the model's ability to identify and close them. One might suspect that the people involved and their ability and willingness to identify gaps play a much bigger role. While it is certainly true that implementing the model alone is not sufficient, neither is having able and willing personnel. Both are necessary and only together can they be sufficient (though other factors, such as information sharing between different organizational units, might also play a role).

Overall, it seems plausible that implementing the 3LoD model would help uncover some gaps in risk coverage that would otherwise remain unnoticed.

4.2 Increasing the effectiveness of risk management practices

Some risk management practices are ineffective—they might look good on paper, but do not work in practice. AI companies might fail to identify relevant risks, misjudge their likelihood or impact, or be unable to reduce them to an acceptable level. Ineffective risk management practices can have many different causes, such as reliance on a single measure (e.g. using a single taxonomy to identify a wide range of risks), a failure to anticipate deliberate attempts to circumvent measures (e.g. stealing an unreleased model), a failure to anticipate relevant changes in the risk landscape [e.g. the emergence of systemic risks due to the increasing reliance on foundation models (Bommasani et al. 2021)], cognitive biases of risk managers [e.g. the availability bias, i.e. the tendency to "assess the frequency of a class or the probability of an event by the ease with which instances or occurrences can be brought to mind" (Tversky and Kahneman 1974)], and other human errors (e.g. a person filling out a risk register slips a line), among other things.

The 3LoD model can increase the effectiveness of risk management practices by identifying such shortcomings. As mentioned above, internal auditors assess the effectiveness of risk management practices and report any shortcomings to the governing body, which can engage with management to improve these practices (Schuett 2023b).

One might object that most shortcomings only occur in low-stakes situations. In high-stakes situations, existing risk management practices are already more effective. For example, AI companies often conduct extensive risk assessments before deploying state-of-the-art models (Brundage et al. 2022; Kavukcuoglu et al. 2022). While this might be true in obvious cases, there are less obvious cases where practices might not be as effective as intended (e.g. because they are insensitive to human errors or deliberate attempts to circumvent them). For example, Anthropic (2023b) recently published a blog post in which they outline some of the challenges they have encountered while evaluating their models. Against this background, I would certainly not want to rely on the counterargument that the effectiveness of risk management practices already scales sufficiently with the stakes at hand.

Some AI companies might further object that they already have the equivalent of an internal audit function, so implementing the 3LoD would only be a marginal improvement. While it might be true that some people at some companies perform some tasks that are similar to what internal auditors do, to the best of my knowledge, assessing the effectiveness of risk management practices is not their main responsibility and they do not follow best practices from the internal audit profession, such as being organizationally independent from management (IIA 2017b), which can lead to noticeable differences.

Overall, I think this is one of the best arguments for implementing the 3LoD model. Without a serious attempt to identify ineffective risk management practices, I expect at least some shortcomings to remain unnoticed. The degree to which this is true mainly depends on internal audit's ability and willingness to serve this function.

4.3 Enabling the governing body to oversee management more effectively

The governing body, typically the board of directors, is responsible for overseeing management. To do this, they need independent and objective information about the company's risk management practices. However, they heavily rely on information provided to them by the executives. To effectively oversee the executives, they need an independent ally in the company.

Internal audit serves this function by maintaining a high degree of independence from management and reporting directly to the audit committee of the board. This can be important because, compared to other actors, the board has significant influence over management. For example, they can replace the CEO (e.g. if they repeatedly prioritize profits over safety), make strategic decisions (e.g. blocking a strategic partnership with the military), and make changes to the company's risk governance (e.g. setting up an ethics

board). Note that there is a complementary reporting line from the CRO to the risk committee of the board.

One might object that this function could also be served by other actors. For example, third-party auditors could also provide the board with independent and objective information. While external audits can certainly play an important role, they have several disadvantages compared to internal audits: they might lack important context, companies might not want to share sensitive information with them (e.g. about ongoing research projects), and audits are typically only snapshots in time. AI companies should therefore see external audit as a complement to internal audit, not a substitution. There is a reason why the 3LoD model distinguishes between internal audit and external assurance providers.

One might further point out that in other industries, internal audit is often perceived to intervene too late (Davies and Zhivitskaya 2018) and to team up with management, instead of monitoring them (Roussy and Rodrigue 2018). This would indeed be problematic. However, as discussed above, this does not seem to be an inherent property of internal audit. Instead, it seems to be mainly driven by the particular way it is set up and the people involved. Having said that, AI companies should take this concern seriously and take measures to address it.

Overall, I think that implementing the 3LoD model can significantly increase the board's information base. This effect will be more noticeable at medium-sized research labs, as most big tech companies already have an internal audit function, albeit not an AI-specific one (see above).

4.4 Other benefits

Implementing the 3LoD model has many benefits other than reducing risks to individuals, groups, or society. Although these other benefits are beyond the scope of this article, it seems warranted to at least give an overview. Below, I briefly discuss four of them.

First, implementing the 3LoD model can avoid unnecessary duplications of risk coverage. Different people in different teams could be doing the same or very similar risk management work. This is often desirable because it can prevent gaps in risk coverage (see above). But if such duplications are not necessary, they can waste resources, such as labor, that could be used more productively elsewhere. AI companies therefore face an effectiveness-efficiency-tradeoff. How this tradeoff ought to be resolved, depends on the particular context. For example, when dealing with catastrophic risks, effectiveness (preventing gaps in risk coverage) seems more important than efficiency (avoiding unnecessary duplications of coverage). In this case, AI companies should strictly err on the side of too much coverage rather than risk gaps in important areas. Overall, this benefit seems to be overstated and less relevant if one is mainly concerned with risk reduction.

Second, AI companies that have implemented the 3LoD model might be perceived as being more responsible. In general, risk management practices at AI companies seem less advanced compared to many other industries (e.g. aviation or banking). By adapting existing best practices from other industries, they would signal that they aim to further professionalize their risk management practices, which could be perceived as being more responsible. This perception might have a number of benefits. For example, it could make it easier to attract and retain talent that cares about ethics and safety. It could also help avoid overly burdensome measures from regulators. It might even be beneficial in litigation cases for the question of whether or not an organization has fulfilled its duty of care. However, it seems questionable whether implementing the 3LoD model affects perception that much, especially compared to other governance measures (e.g. publishing AI ethics principles or setting up an AI ethics board), mainly because most stakeholders, including most employees, do not know the model and cannot assess its relevance. An exception might be regulators and courts who care more about the details of risk management practices. My best guess is that implementing the model will have noticeable effects on the perception of a few stakeholders, while most other stakeholders will not care.

Third, implementing the 3LoD model can make it easier to hire risk management talent. The profession of AI risk management is in its infancy. I assume that AI companies find it challenging to hire people with AI *and* risk management expertise. In most cases, they can either hire AI experts and train them in risk management, or hire risk management experts from other industries and train them in AI. Implementing the 3LoD model could make it easier to hire risk management experts from other industries because they would already be familiar with the model. This might become more important if one assumes that AI companies will want to hire more risk management talent as systems get more capable and are used in more safety–critical situations (e.g. Degrave et al. 2022). However, I do not find this argument very convincing. I doubt that implementing the 3LoD model would make a meaningful difference on relevant hiring decisions (e.g. on a candidate's decision to apply or accept an offer). Since the model is about the organizational dimension of risk management, it does not have significant effects on the day-to-day risk management work. Having said that, there might be smaller benefits (e.g. making the onboarding process easier). My best guess is that the counterfactual impact of 3LoD implementation on hiring is low.

Fourth, implementing the 3LoD model might reduce financing costs. Rating agencies tend to give better ratings

to companies that have implemented an ERM framework (because doing so is considered best practice), and companies with better ratings tend to have lower financing costs (because they get better credit conditions) (see Bohnert et al. 2019). There might be an analogous effect with regards to the implementation of the 3LoD model. Lower financing costs are particularly important if one assumes that the costs for developing state-of-the-art AI systems will increase because of increasing demand for compute (Sevilla et al. 2022), for example. In scenarios where commercial pressure is much higher than today, lower financing costs could also be important to continue safety research that does not contribute to product development. That said, I am uncertain to what extent the findings for ERM frameworks generalize to the 3LoD model. My best guess is that implementing the 3LoD would not have meaningful effects on the financing costs of medium-sized research labs today. But I expect this to change as labs become more profitable and increasingly make use of other funding sources (e.g. credits or bonds).

5 Conclusion

This article has applied the 3LoD model to an AI context. It has suggested concrete ways in which frontier AI developers like OpenAI, Google DeepMind, and Anthropic could implement the model to reduce risks from AI. It has argued that implementing the model could prevent individual, collective, or societal harm by identifying and closing gaps in risk coverage, increasing the effectiveness of risk management practices, and enabling the governing body to oversee management more effectively. It concluded that, while there are some limitations and the effects should not be overstated, the model can plausibly contribute to a reduction of risks from AI.

Based on the findings of this article, I suggest the following questions for further research. First, my discussion of the model's ability to reduce risks from AI was mostly theoretical and relied on abstract plausibility considerations. I encourage other scholars to assess these claims empirically. An industry case study similar to the one that Mökander and Floridi (2022) conducted for ethics-based auditing could be a first step. Second, although AI companies do not seem to have implemented the 3LoD model, they already perform many of the above-mentioned activities. To better target future work, it would be helpful to review existing risk management practices at these companies and conduct a gap analysis. Since public data is scarce, scholars would have to conduct interviews or surveys (e.g. an "AI risk management benchmark survey"), though I expect confidentiality to be a major obstacle. Such a survey could be similar to the one conducted by Schuett et al. (2023a, b) on best practices in AI safety and governance. Third, the article has focused on the voluntary adoption of the 3LoD model. It would be important to know if existing or future regulations might even require AI companies to implement the model (Anderljung et al. 2023). For example, while Article 9 of the proposed EU AI Act does not mention the 3LoD model, it has been suggested that future harmonized standards or common specifications should include the model (Schuett 2023a). The 3LoD model is also mentioned in the playbook that accompanies the NIST AI Risk Management Framework (NIST 2023a, 2023b). It is conceivable that this framework will be translated into US law, similar to the NIST Framework for Improving Critical Infrastructure Cybersecurity (NIST 2018). Finally, the article has investigated the 3LoD in isolation. It has excluded contextual factors, such as the risk culture at AI companies, which might also affect the model's effectiveness. A better understanding of these factors would further improve the information base for decision-makers at AI companies and beyond.

As famously put by George Box (1976), "all models are wrong, but some are useful". In the same spirit, one might say that the 3LoD model is not a silver bullet against the risks from AI, but it can still play an important role. AI companies should see it as one of many governance tools they can use to tackle today's and tomorrow's threats from AI.

Acknowledgements I am grateful for valuable comments and feedback from Leonie Koessler, James Ginns, Markus Anderljung, Andre Barbe, Noemie Dreksler, Toby Shevelane, Anne le Roux, Alexis Carlier, Emma Bluemke, Christoph Winter, Renan Araújo, José Jaime Villalobos, Suzanne Van Arsdale, Alfredo Parra, and Nick Hollman.

Data availability Not applicable.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Alaga J, Schuett J (2023) Coordinated pausing: an evaluation-based coordination scheme for frontier AI developers. arXiv. http://arxiv.org/abs/2310.00374

Alphabet (2022) Notice of 2022 annual meeting of stockholders and proxy statement. SEC. https://perma.cc/Q23E-WQWP

Anderljung M, Barnhart J, Korinek A, Leung J, O'Keefe C, Whittlestone J et al (2023) Frontier AI regulation: managing emerging risks to public safety. arXiv. http://arxiv.org/abs/2307.03718

Andersen TJ, Sax J, Giannozzi A (2022) Conjoint effects of interacting strategy-making processes and lines of defense practices in strategic risk management: an empirical study. Long Range Plan 55(6):102164. https://doi.org/10.1016/j.lrp.2021.102164

Anthropic (2023a) Anthropic's responsible scaling policy. Anthropic. https://perma.cc/S393-UCHE

Anthropic (2023b) Challenges in evaluating AI systems. Anthropic. https://perma.cc/69ZX-RTGY

Anthropic (2023c) Frontier model security. Anthropic. https://perma.cc/6HQ4-XV73

ARC Evals (2023) Responsible scaling policies (RSPs). ARC Evals. https://perma.cc/Z3QC-GFZ4

Arndorfer I, Minto A (2015) The "four lines of defence model" for financial institutions. Financial Stability Institute, Bank for International Settlements. https://perma.cc/UP35-KEYJ

Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Kaplan J (2022) Constitutional AI: harmlessness from AI feedback. arXiv. http://arxiv.org/abs/2212.08073

Bantleon U, d'Arcy A, Eulerich M, Hucke A, Pedell B, Ratzinger-Sakel NVS (2021) Coordination challenges in implementing the three lines of defense model. Int J Audit 25(1):59–74. https://doi.org/10.1111/ijau.12201

Baquero JA, Burkhardt R, Govindarajan A, Wallace T (2020) Derisking AI by design: how to build risk management into AI development. McKinsey. https://perma.cc/2WPN-A6CW

Barrett AM, Hendrycks D, Newman J, Nonnecke B (2022) Actionable guidance for high-consequence AI risk management: towards standards addressing AI catastrophic risks. arXiv. http://arxiv.org/abs/2206.08966

Barrett AM, Newman J, Nonnecke B, Hendrycks D, Murphy ER, Jackson K (2023) AI risk-management standards profile for general-purpose AI systems (GPAIS) and foundation models. Center for Long-Term Cybersecurity, UC Berkeley. https://perma.cc/8W6P-2UUK

BCBS (1999) Enhancing corporate governance for banking organisations. Bank for International Settlements. https://perma.cc/G2QP-7K5B

BCBS (2012) The internal audit function in banks. Bank for International Settlements. https://perma.cc/A57Q-8LZ6

Bengio Y, Hinton G, Yao A, Song D, Abbeel P, Harari YN et al (2023) Managing AI risks in an era of rapid progress. arXiv. http://arxiv.org/abs/2310.17688

Boatright J (2016) Why risk management failed: ethical and behavioral aspects. In: Malliaris AG, Shaw L, Shefrin H (eds) The global financial crisis and its aftermath: hidden factors in the meltdown. Oxford University Press, Oxford, pp 384–386. https://doi.org/10.1093/acprof:oso/9780199386222.003.0017

Bohnert A, Gatzert N, Hoyt RE, Lechner P (2019) The drivers and value of enterprise risk management: evidence from ERM ratings. Eur J Finance 25(3):234–255. https://doi.org/10.1080/1351847X.2018.1514314

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S et al (2021) On the opportunities and risks of foundation models. arXiv. https://arxiv.org/abs/2108.07258

Boţa-Avram C, Palfi C (2009) Measuring and assessment of internal audit's effectiveness. Ann Faculty Econ Univ Oradea 3(1):784–790

Box GEP (1976) Science and statistics. J Am Stat Assoc 71(356):791–799. https://doi.org/10.1080/01621459.1976.10480949

Brundage M, Mayer K, Eloundou T, Agarwal S, Adler S, Krueger G, Leike J, Mishkin P (2022) Lessons learned on language model safety and misuse. OpenAI. https://perma.cc/8RKR-QJZY

Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, p 77–91. https://perma.cc/976J-AR93

Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu PS, Sun L (2023) A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. arXiv. http://arxiv.org/abs/2303.04226

Carcello JV, Eulerich M, Masli A, Wood DA (2020) Are internal audits associated with reductions in perceived risk? Auditing J Pract Theor 39(3):55–73. https://doi.org/10.2308/ajpt-19-036

Cheatham B, Javanmardian K, Samandari H (2019) Confronting the risks of artificial intelligence. McKinsey. https://perma.cc/T2CX-HYZF

Chen M, Tworek J, Jun H, Yuan Q, de Pinto HPO, Kaplan J et al (2021) Evaluating large language models trained on code. arXiv. http://arxiv.org/abs/2107.03374

Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning from human preferences. arXiv. http://arxiv.org/abs/1706.03741

Coram P, Ferguson C, Moroney R (2008) Internal audit, alternative internal audit structures and the level of misappropriation of assets fraud. Account Finance 48(4):543–559. https://doi.org/10.1111/j.1467-629X.2007.00247.x

COSO (2017) Enterprise risk management—integrating with strategy and performance. https://perma.cc/5Z3G-KD6R

Crafts N (2021) Artificial intelligence as a general-purpose technology: an historical perspective. Oxf Rev Econ Policy 37(3):521–536. https://doi.org/10.1093/oxrep/grab012

Davies H, Zhivitskaya M (2018) Three lines of defence: a robust organising framework, or just lines in the sand? Global Pol 9(S1):34–42. https://doi.org/10.1111/1758-5899.12568

Degrave J, Felici F, Buchli J, Neunert M, Tracey B, Carpanese F et al (2022) Magnetic control of tokamak plasmas through deep reinforcement learning. Nature 602:414–419. https://doi.org/10.1038/s41586-021-04301-9

Drogalas G, Pazarskis M, Anagnostopoulou E, Papachristou A (2017) The effect of internal audit effectiveness, auditor responsibility and training in fraud detection. J Account Manag Inf Syst 16(4):434–454. https://doi.org/10.24818/jamis.2017.04001

EBA (2021) Final report on guidelines on internal governance under Directive 2013/36/EU (EBA/GL/2021/05). https://perma.cc/RCD8-V99V

Eulerich A, Eulerich M (2020) What is the value of internal auditing? A literature review on qualitative and quantitative perspectives. Maandblad Voor Accountancy En Bedrijfseconomie 94(3/4):83–92. https://doi.org/10.5117/mab.94.50375

European Commission (2021) Proposal for a regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) (COM(2021) 206 final). https://perma.cc/4YXM-38U9

Evans O, Cotton-Barratt O, Finnveden L, Bales A, Balwit A, Wills P, Righetti L, Saunders W (2021) Truthful AI: developing and governing AI that does not lie. arXiv. https://arxiv.org/abs/2110.06674

Financial Services Authority (2003) Building a framework for operational risk management: the FSA's observations. https://perma.cc/5AX2-M2LF

Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S et al (2022) Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. arXiv. https://arxiv.org/abs/2209.07858

Garfinkel B (2022) The impact of artificial intelligence. In: Bullock JB, Chen Y-C, Himmelreich J, Hudson VM, Korinek A, Young MM, Zhang B (eds) The Oxford handbook of AI governance. Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780197579329.013.5

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, Crawford K (2021) Datasheets for datasets. arXiv. https://arxiv.org/abs/1803.09010

Gehrmann S, Clark E, Sellam T (2022) Repairing the cracked foundation: a survey of obstacles in evaluation practices for generated text. arXiv. http://arxiv.org/abs/2202.06935

Green N, Procope C, Cheema A, Adediji A (2022) System cards, a new resource for understanding how AI systems work. Meta AI. https://perma.cc/CQZ8-FQ44

Hacker P (2018) Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law. Common Market Law Rev 55(4):1143–1185. https://doi.org/10.54648/cola2018095

Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Mind Mach 30(1):99–120. https://doi.org/10.1007/s11023-020-09517-8

Hagendorff T (2022) A virtue-based framework to support putting AI ethics into practice. Philos Technol. https://doi.org/10.1007/s13347-022-00553-z

Hamon R, Junklewitz H, Sanchez I, Malgieri G, De Hert P (2022) Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. IEEE Comput Intell Mag 17(1):72–85. https://doi.org/10.1109/MCI.2021.3129960

Hoefer E, Cooke M, Curry T (2020) Three lines of defense: failed promises and what comes next. Reuters. https://perma.cc/V35X-VTC5

Hua S-S, Belfield H (2021) AI & antitrust: reconciling tensions between competition law and cooperative AI development. Yale J Law Technol 23(415). https://perma.cc/4VL2-QNNJ

Huibers SCJ (2015). Combined assurance: one language, one voice, one view. IIA Research Foundation, Global Internal Audit Common Body of Knowledge. https://perma.cc/D7YM-9GSY

IEC (2019) Risk management—risk assessment techniques (IEC Standard No. 31010:2019). https://www.iso.org/standard/72140.html

IIA (2013) IIA position paper: the three lines of defense in effective risk management and control. https://perma.cc/NQM2-DD7V

IIA (2017a) Artificial intelligence: considerations for the profession of internal auditing (Part I). https://perma.cc/K8WQ-VNFZ

IIA (2017b) International standards for the professional practice of internal auditing. https://perma.cc/AKU7-8YWZ

IIA (2017c) The IIA's artificial intelligence auditing framework: practical applications (Part A). https://perma.cc/U93U-LN75

IIA (2018) The IIA's artificial intelligence auditing framework: practical applications (Part B). https://perma.cc/826X-Y3L7

IIA (2020a) The IIA's three lines model: an update of the three lines of defense. https://perma.cc/GAB5-DMN3

IIA (2020b) Good practice internal audit reports. https://perma.cc/7BQT-DTRD

ISO and IEC (2023) Information technology—artificial intelligence—guidance on risk management (ISO/IEC Standard No. 23894). https://www.iso.org/standard/77304.html

ISO (2018) Risk management—guidelines (ISO Standard No. 31000:2018). https://www.iso.org/standard/65694.html

Jiang L, Messier WF, Wood DA (2020) The association between internal audit operations-related services and firm operating performance. Auditing J Pract Theor 39(1):101–124. https://doi.org/10.2308/ajpt-52565

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2

Kavukcuoglu K, Kohli P, Ibrahim L, Bloxwich D, Brown S (2022) How our principles helped define AlphaFold's release. Google DeepMind. https://perma.cc/3ARS-XLNV

Kinniment M, Koba Sato LJ, Du H, Goodrich B, Hasin M, Chan L et al (2023) Evaluating language-model agents on realistic autonomous tasks. ARC Evals. https://perma.cc/2V5J-S3M7

Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. arXiv. https://arxiv.org/abs/1609.05807

Koessler L, Schuett J (2023) Risk assessment at AGI companies: a review of popular risk assessment techniques from other safety-critical industries. arXiv. https://arxiv.org/abs/2307.08823

Kräussl R (2003) A critique on the proposed use of external sovereign credit ratings in Basel II. Center for Financial Studies, Goethe University Frankfurt. https://perma.cc/PMB8-WSCA

Lambert N, Castricato L, von Werra L, Havrilla A (2022) Illustrating reinforcement learning from human feedback (RLHF). Hugging Face Blog. https://perma.cc/R9HU-TQ9X

Leech TJ, Hanlon LC (2016) Three lines of defense versus five lines of assurance: elevating the role of the board and CEO in risk governance. In: Leblanc R (ed) The handbook of board governance: a comprehensive guide for public, private and not-for-profit board members. Wiley, Hoboken, pp 335–355. https://doi.org/10.1002/9781119245445.ch17

Lenz R, Hahn U (2015) A synthesis of empirical internal audit effectiveness literature pointing to new research opportunities. Manag Audit J 30(1):5–33. https://doi.org/10.1108/MAJ-08-2014-1072

Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M et al (2022) Holistic evaluation of language models. arXiv. http://arxiv.org/abs/2211.09110

Lin S, Pizzini M, Vargus M, Bardhan IR (2011) The role of the internal audit function in the disclosure of material weaknesses. Account Rev 86(1):287–323. https://doi.org/10.2308/accr.00000016

Lundqvist SA (2015) Why firms implement risk governance: stepping beyond traditional risk management to enterprise risk management. J Account Public Policy 34(5):441–466. https://doi.org/10.1016/j.jaccpubpol.2015.05.002

Maayan Y, Carmeli A (2016) Internal audits as a source of ethical behavior, efficiency, and effectiveness in work units. J Bus Ethics 137(2):347–363. https://doi.org/10.1007/s10551-015-2561-0

McGregor S (2021) Preventing repeated real world AI failures by cataloging incidents: the AI incident database. Proc AAAI Conf Artif Intell 35(17):15458–15463. https://doi.org/10.1609/aaai.v35i17.17817

Microsoft (2022) Notice of annual shareholders meeting and proxy statement 2022. SEC. https://perma.cc/6NYQ-ZTMB

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, p 220–229. https://doi.org/10.1145/3287560.3287596

Mohamed S, Png M-T, Isaac W (2020) Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philos Technol 33(4):659–684. https://doi.org/10.1007/s13347-020-00405-8

Mökander J, Floridi L (2022) Operationalising AI governance through ethics-based auditing: an industry case study. AI and Ethics. https://doi.org/10.1007/s43681-022-00171-7

Mökander J, Schuett J, Kirk HR, Floridi L (2023) Auditing large language models: a three-layered approach. AI Ethics. https://doi.org/10.1007/s43681-023-00289-2

Nasdaq (2022) Nasdaq 5600 series: corporate governance requirements. https://perma.cc/4M7B-U42F
NIST (2018) Framework for improving critical infrastructure cybersecurity (Version 1.1). https://doi.org/10.6028/NIST.CSWP.04162018
NIST (2023a) Artificial intelligence risk management framework: playbook (Govern 4.1). https://perma.cc/LNF7-REPM
NIST (2023b) Artificial intelligence risk management framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1
NIST (2023c) Biden-Harris administration announces new NIST publicworking group on AI. https://perma.cc/FCP7-Z7P3
Nunn R (2020) Discrimination in the age of algorithms. In: Barfield W (ed) The Cambridge handbook of the law of algorithms. Cambridge University Press, Cambridge, pp 182–198. https://doi.org/10.1017/9781108680844.010
O'Brien J, Ee S, Williams Z (2023) Deployment corrections: an incident response framework for frontier AI models. arXiv. http://arxiv.org/abs/2310.00328
OECD (2023) OECD AI incidents monitor. OECD. https://oecd.ai/en/incidents
OpenAI (2023a) OpenAI's approach to frontier risk. OpenAI. https://perma.cc/9YGS-NZVX
OpenAI (2023b) Frontier risk and preparedness. OpenAI. https://perma.cc/5AFJ-JZG4
Ord T (2021) Proposal for a new 'three lines of defence' approach to UK risk management. Future of Humanity Institute, University of Oxford. https://perma.cc/VHH9-L36R
Oussii AA, Boulila Taktak N (2018) The impact of internal audit function characteristics on internal control quality. Manag Audit J 33(5):450–469. https://doi.org/10.1108/MAJ-06-2017-1579
PAI (2021) Managing the risks of AI research: six recommendations for responsible publication. https://perma.cc/BX5A-KE8D
PAI (2023) PAI's Guidance for safe foundation model deployment: a framework for collective action. PAI. https://perma.cc/W9GN-6QY3
Perez E, Huang S, Song F, Cai T, Ring R, Aslanides J et al (2022a) Red teaming language models with language models. arXiv. https://arxiv.org/abs/2202.03286
Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S et al (2022b) Discovering language model behaviors with model-written evaluations. arXiv. http://arxiv.org/abs/2212.09251
Petit N (2017) Antitrust and artificial intelligence: a research agenda. J Eur Compet Law Pract 8(6):361–362. https://doi.org/10.1093/jeclap/lpx033
Power M, Ashby S, Palermo T (2013) Risk culture in financial organisations: a research report. The London School of Economics and Political Science. https://perma.cc/R9YC-AT4Z
Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B et al (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. arXiv. https://arxiv.org/abs/2001.00973
Raji ID, Kumar IE, Horowitz A, Selbst A (2022) The fallacy of AI functionality. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, p 959–972. https://doi.org/10.1145/3531146.3533158
Raji ID, Xu P, Honigsberg C, Ho D (2022) Outsider oversight: designing a third party audit ecosystem for AI governance. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, p 557–571. https://doi.org/10.1145/3514094.3534181
Rando J, Paleka D, Lindner D, Heim L, Tramèr F (2022) Red-teaming the stable diffusion safety filter. arXiv. https://arxiv.org/abs/2210.04610
Rao A, Golbin I (2021) Top-down and end-to-end governance for the responsible use of AI. Towards Data Sci. https://perma.cc/SM8Y-6CUN
Roussy M, Rodrigue M (2018) Internal audit: Is the 'third line of defense' effective as a form of governance? An exploratory study of the impression management techniques chief audit executives use in their annual accountability to the audit committee. J Bus Ethics 151:853–869. https://doi.org/10.1007/s10551-016-3263-y
Rupšys R, Boguslauskas V (2007) Measuring performance of internal auditing: empirical evidence. Eng Econ 55(5):9–15
Savčuk O (2007) Internal audit efficiency evaluation principles. J Bus Econ Manag 8(4):275–284. https://doi.org/10.3846/16111699.2007.9636180
Schuett J (2023) Risk management in the Artificial Intelligence Act. Eur J Risk Regul. https://doi.org/10.1017/err.2023.1
Schuett J, Dreksler N, Anderljung M, McCaffary D, Heim L, Bluemke E, Garfinkel B (2023) Towards best practices in AGI safety and governance: a survey of expert opinion. arXiv. http://arxiv.org/abs/2305.07153
Schuett J, Reuel A, Carlier A (2023) How to design an AI ethics board. arXiv. https://arxiv.org/abs/2304.07249
Schuett J (2023b) AGI labs need an internal audit function. arXiv. https://arxiv.org/abs/2305.17038
Seger E, Dreksler N, Moulange R, Dardaman E, Schuett J, Wei K et al (2023) Open-sourcing highly capable foundation models: an evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. arXiv. https://arxiv.org/abs/2311.09227
Sekar M (2022) Machine learning for auditors: automating fraud investigations through artificial intelligence. Apress. https://doi.org/10.1007/978-1-4842-8051-5
Sevilla J, Heim L, Ho A, Besiroglu T, Hobbhahn M, Villalobos P (2022) Compute trends across three eras of machine learning. arXiv. https://arxiv.org/abs/2202.05924
Shelby R, Rismani S, Henne K, Moon A, Rostamzadeh N, Nicholas P et al (2022) Sociotechnical harms of algorithmic systems: scoping a taxonomy for harm reduction. arXiv. http://arxiv.org/abs/2210.05791
Shevlane T (2022) Structured access: an emerging paradigm for safe AI deployment. In: Bullock JB, Chen Y-C, Himmelreich J, Hudson VM, Korinek A, Young MM, Zhang B (eds) The Oxford handbook of AI governance. Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780197579329.013.39
Shevlane T, Farquhar S, Garfinkel B, Phuong M, Whittlestone J, Leung J et al (2023) Model evaluation for extreme risks. arXiv. http://arxiv.org/abs/2305.15324
Smuha NA (2021) Beyond the individual: governing AI's societal harm. Internet Policy Rev. https://doi.org/10.14763/2021.3.1574
Solaiman I, Dennison C (2021) Process for adapting language models to society (PALMS) with values-targeted datasets. Adv Neural Inf Process Syst 34:5861–5873
Solaiman I, Brundage M, Clark J, Askell A, Herbert-Voss A, Wu J et al (2019) Release strategies and the social impacts of language models. arXiv. https://arxiv.org/abs/1908.09203
Solaiman I (2023) The gradient of generative AI release: methods and considerations. *arXiv*. http://arxiv.org/abs/2302.04844
Tammenga A (2020) The application of artificial intelligence in banks in the context of the three lines of defence model. Maandblad Voor Accountancy En Bedrijfseconomie 94(5/6):219–230. https://doi.org/10.5117/mab.94.47158
Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185(4157):1124–1131. https://doi.org/10.1126/science.185.4157.1124
van Asselt MBA, Renn O (2011) Risk governance. J Risk Res 14(4):431–449. https://doi.org/10.1080/13669877.2011.553730
Vousinas GL (2021) Beyond the three lines of defense: the five lines of defense model for financial institutions. ACRN J Finance Risk Perspect 10(1):95–110. https://doi.org/10.35944/jofrp.2021.10.1.006

Wachter S, Mittelstadt B, Russell C (2021) Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. Comput Law Secur Rev 41:105567. https://doi.org/10.1016/j.clsr.2021.105567

Walker D (2009) A review of corporate governance in UK banks and other financial industry entities: final recommendations. https://perma.cc/2K9C-EMME

Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang P-S et al (2021) Ethical and social risks of harm from language models. arXiv. https://arxiv.org/abs/2112.04359

Weidinger L, Rauh M, Marchal N, Manzini A, Hendricks LA, Mateos-Garcia J et al (2023) Sociotechnical safety evaluation of generative AI systems. arXiv. http://arxiv.org/abs/2310.11986

Zhivitskaya M (2015) The practice of risk oversight since the global financial crisis: closing the stable door? [Doctoral dissertation, The London School of Economics and Political Science]. LSE Theses Online. https://perma.cc/KKA6-QK56

Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, Irving G (2019) Fine-tuning language models from human preferences. arXiv. http://arxiv.org/abs/1909.08593