AI & SOCIETY (2025) 40:1469–1484
https://doi.org/10.1007/s00146-024-01950-y

NETWORK RESEARCH

The case for a broader approach to AI assurance: addressing "hidden" harms in the development of artificial intelligence

Christopher Thomas[1] · Huw Roberts[2] · Jakob Mökander[2,3] · Andreas Tsamados[2] · Mariarosaria Taddeo[1,2] · Luciano Floridi[3,4]

Received: 13 December 2023 / Accepted: 10 April 2024 / Published online: 16 May 2024

Abstract

Artificial intelligence (AI) assurance is an umbrella term describing many approaches—such as impact assessment, audit, and certification procedures—used to provide evidence that an AI system is legal, ethical, and technically robust. AI assurance approaches largely focus on two overlapping categories of harms: deployment harms that emerge at, or after, the point of use, and individual harms that directly impact a person as an individual. Current approaches generally overlook upstream collective and societal harms associated with the development of systems, such as resource extraction and processing, exploitative labour practices and energy intensive model training. Thus, the scope of current AI assurance practice is insufficient for ensuring that AI is ethical in a holistic sense, i.e. in ways that are legally permissible, socially acceptable, economically viable and environmentally sustainable. This article addresses this shortcoming by arguing for a broader approach to AI assurance that is sensitive to the full scope of AI development and deployment harms. To do so, the article maps harms related to AI and highlights three examples of harmful practices that occur upstream in the AI supply chain and impact the environment, labour, and data exploitation. It then reviews assurance mechanisms used in adjacent industries to mitigate similar harms, evaluating their strengths, weaknesses, and how effectively they are being applied to AI. Finally, it provides recommendations as to how a broader approach to AI assurance can be implemented to mitigate harms more effectively across the whole AI supply chain.

Keywords Artificial intelligence · Assurance · Audit · Certification · Compliance · ESG · Ethics · Governance · Impact assessment · Sustainability

✉ Huw Roberts
huw.roberts@oii.ox.ac.uk

Christopher Thomas
c.thomas@turing.ac.uk;

[1] Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK

[2] Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford OX1 3JS, UK

[3] Digital Ethics Center, Yale University, 85 Trumbull St., New Haven, CT 06511, USA

[4] Department of Legal Studies, University of Bologna, Via Zamboni 27/29, 40126 Bologna, Italy

1 Introduction

AI assurance consists of processes that are expected to ensure that the development or use of an AI system is ethical, legally compliant, or at least functioning as claimed (Roberts and Babuta 2023). Typical assurance techniques for AI include algorithm impact assessments (Metcalf et al. 2021), bias audits (Metaxa et al. 2021), and certification schemes (Cihon et al. 2021). Assurance has emerged as a topic of interest in AI governance, for both public and private sectors. The US, EU, China, Singapore, Russia, India and the UK, amongst others, have released policies that require or encourage AI assurance.[1] For example, the UK government's algorithmic transparency recording standard is driving public sector organisations to demonstrate robust due diligence when deploying AI products and services (Kingsman et al. 2022). Similarly, China requires developers of AI systems with "public opinion properties" or "social mobilisation capabilities" to submit key details about their systems to an algorithmic registry, which enables regulators to audit these systems (Sheehan and Du 2022). In the private sector, we have seen an emerging "AI assurance ecosystem", including new AI assurance start-ups, an increasing focus on AI assurance from traditional auditing services, and the creation of open-source toolkits to aid companies in auditing their systems (CDEI 2021; Mökander 2023).[2] Organisations across the globe are now advertising their compliance with emerging standards and regulations, through third-party audits covering areas like privacy, fairness, and explainability.[3]

So far, approaches to AI assurance have predominantly sought to address a narrow set of harms associated with the use and downstream impact (henceforth deployment) of AI systems (Kazim and Koshiyama 2020; Freeman et al. 2022). Particularly in western policy contexts, AI assurance techniques have also focused on the impacts of these systems on individuals.[4] While these approaches mark important steps towards ensuring more trustworthy AI, this narrow focus is problematic as harms occur across the whole AI supply chain and can impact groups and the environment. In the early stages of development, the hardware implementing AI systems requires raw materials, which are often mined and processed in ethically questionable ways by relying on both environmental and labour exploitation (Crawford 2021). Further along the supply chain, data labelling and model training may rely on exploitative labour and high energy costs, respectively. Finally, deployment can lead to biased decisions and the entrenching of societal inequalities. Without also considering the externalities that arise upstream during the development of AI systems and the wider collective and societal dimensions of these harms, assurance will only prove to be partially effective. For example, improving the degree of transparency of an algorithm, or its governance and use is often cited as a rationale for developing AI assurance tools[5]; however, transparency over AI supply chains, including resource extraction, labour, and energy usage, is largely ignored (Hagendorff 2021). This can lead to harms from systems that may, based on existing assurance techniques alone, be labelled "ethical". To avoid this outcome and ensure that AI is more holistically ethical—that is to say, legally permissible, socially acceptable, economically viable and environmentally sustainable along the whole supply chain—it is crucial to develop tools that assess the full spectrum of harms an AI system can produce.

The task of this article is to consider how practical assurance methods can be applied to address the upstream harms associated with the development of AI. The aim is to highlight the gaps in current AI assurance work, drawing on assurance practices in adjacent industries which could be adapted and applied to address these gaps. We centre our analysis on how industry-led assurance mechanisms can be applied to AI. While we recognise that industry-led approaches alone are insufficient to mitigate AI harms, industry will need play a significant role if these harms are to be managed effectively. The task is to draw attention to, and rectify, the incompleteness and narrowness of existing industry led approaches. Considering the full range of potential assurance mechanisms that can be drawn from public sector, civil society, and research and community contexts will be crucial to complement the analysis developed in this article, but we leave this task to future research.

This paper focuses on four types of practices: ex ante impact assessment, ex post reporting and auditing, sustainable investing, and third-party certification.[6] This is due to

[1] See the UK's AI Regulation White Paper (DSIT 2023); The EU AI Act (European Commission 2021) and; China's Internet Information Service Algorithmic Recommendation Management Provisions (Creemers et al. 2022); Singapore's AI Strategy 2.0 (including the AI Verify programme) (Smart Nation and Digital Government Office 2023); National Strategy for the development of artificial intelligence for the period until 2030 [updated 15 February 2024] (Government of the Russian Federation 2019); India's advisory on the approval and labelling of AI tools (Digital Watch Observatory 2024).

[2] See IBM AI Fairness 360: (Cihon et al. 2021).

[3] See Holistic AI, assurance case studies: https://www.holisticai.com/case-study.

[4] China's approach to assurance differs to a degree in that more emphasis is placed on assessing models to ensure social stability and political alignment (Roberts et al. 2023; Triolo 2023).

[5] Transparency is cited as an objective for 261 tools across the OECD's tools for trustworthy AI database: https://oecd.ai/en/catalogue/tools?objectiveIds=11&page=1.

[6] Conformity assessment is a prominent assurance mechanism which is not included directly in our analysis. Conformity assessments include the use of impact assessments, audits, and certification, to fulfil the requirements of a standard, often for the purpose of regulatory

their relative maturity in addressing upstream supply chain harms in adjacent industries (Kolk 2004; Matus 2010; Bond et al. 2012; Boffo and Pantalano 2020), as well as their broad reference across AI assurance policy and practice (focused on deployment harms) (Kazim and Koshiyama 2020; Batarseh et al. 2021; CDEI 2021, 2023; Freeman et al. 2022). Sustainable investing is the least well referenced to date in AI assurance literature, however, significant work has recently emerged (Morris and Rosenburg 2023). This overlap suggests that these tools could be adapted and integrated into AI supply chains effectively, to offer a nested approach in which AI assurance is situated within, rather than separate from broader corporate supply chain governance. We are not the first to consider the application of assurance mechanisms to supply chain harms. For example, Matus and Veale (2022), consider how standards and certification programmes from the field of sustainability can be applied along the AI supply chain, and Sætra (2022, 2023) develops an "action oriented" AI ESG reporting framework. In our analysis, we build on their work, considering how a range of assurance practices focused on mitigating supply chain harms in adjacent industries can be adapted and applied to mitigate upstream harms in the AI supply chain, to complement existing AI assurance approaches and ensure more holistically ethical AI.

By focusing on the extension of ethical duties across the supply chain, our argument in this article is framed within the prevailing principles-based approach to AI ethics focused on harm-mitigation, which underpins the majority of AI assurance practices (Hagendorff 2020; Morley et al. 2020; Ryan and Stahl 2020).[7] Our article addresses a deficiency in the scope of current principles-based AI ethics and associated assurance practices, whereby the duties and obligations associated with ethical principles—such as transparency, fairness, and safety—are limited to mitigating the downstream development and deployment harms of AI systems. A focus on the narrowness of existing ethical principles and corresponding assurance obligations anchors our analysis and our concrete recommendations for broadening the scope of AI assurance. Similarly, focusing on the negative externalities or indirect costs of AI development invites conceptualising ethics in terms of harm mitigation, rather than maximising benefit, or increasing flourishing. Conceptually, it makes more sense to talk about reducing the environmental cost of AI development, rather than maximising the environmental benefit of AI development or flourishing through the indirect consequences of upstream AI development. Similarly with labour and data exploitation, the immediate obligation and most achievable step to improving individual and societal conditions is to mitigate the concrete harms caused to individuals and society, rather than to consider what maximally beneficial labour or data conditions might be. This is not to dismiss the importance of undertaking a wider evaluation of the ethical framing underpinning current AI assurance practices, but it is beyond the scope of this policy-oriented paper.

The remainder of this article is structured as follows. Section two maps AI harms, highlighting three significant areas: environmental, labour, and data exploitation. Section three considers impact assessment, reporting and auditing, sustainable investing, and certification approaches, as they are used in other fields, and how these can be applied across the AI supply chain. The conclusion is that most approaches are currently immature or poorly applied when it comes to the upstream, collective, and societal harms associated with AI. Section four discusses the limitations of these approaches, and how they can be overcome. Section five concludes the article by providing a set of recommendations for adapting and integrating these assurance mechanisms to mitigate the environmental, social, and economic harms that arise throughout the entire AI supply chain.

2 Mapping AI harms

Harms occur across the whole AI supply chain. The types of harms that AI can bring about can be mapped in different ways.[8] Here, building on existing AI harms taxonomies (Future of Privacy Forum 2017; Bommasani et al. 2022; Shelby et al. 2022; Weidinger et al. 2022) we present a non-exhaustive conceptual map of AI harms along two intersecting dimensions: individual-collective-societal harms and development-deployment harms, as shown in Table 1. The table provides a simple heuristic for thinking about the different ways harms occur, and it should be noted that in practice, these categories are not clearly distinct but overlapping.

When considering discussions on AI ethics and tools that can be developed to assure ethical outcomes (Morley et al. 2020), it becomes apparent that some types of harms receive far more attention than others (Tsamados et al. 2021). Specifically, most attention has been paid to deployment harms that impact the individual and, to a lesser degree, collective and societal harms that result from development. This is problematic, as these harms also cause significant ethical concerns but often require coordination between multiple parties to be brought to light. It should also be stressed that

Footnote 6 (continued)

compliance. Recommendations made in this paper regarding impact assessments, audits, and certification can be extended to their use in conformity assessment procedures. See Mökander et al. (2021) for more information on conformity assessment procedures.

[7] Other scholars have advocated virtue ethics as a normative foundation for AI ethics for example, see Farina et al. (2022).

[8] For instance, one mapping that we have not employed here is between intentional and unintentional AI harms.

Table 1 Mapping intersecting dimensions of harm across the AI supply chain

	Resource extraction	Resource processing	Deployment
Individual	Exploitative labour Data theft	Exploitative labour Non-consensual data processing	Behavioural manipulation Physical harm Unfair surveillance
Collective	Exploitation of marginalised or indigenous groups Group privacy harms from data collection Exploitation of community resources	Exploitation of marginalised or indigenous groups Group privacy harms from data processing	Unfair surveillance of groups Exploitation of gig workers Algorithmic group discrimination
Societal	Institutionalised labour exploitation Depletion of natural resources Biodiversity loss Injury to animals Extraction emmissions	Institutionalised labour exploitation Contamination of natural resources Model training emissions Chemical waste Water use	Institutionalised labour exploitation Widening structural inequalities Model use emissions Increasing digital divide Eroding trust in social/political institutions

harms across these stages are often interlinked. For example, the overextraction of data at the development stage leads to heightened risks of downstream predictive privacy harms.

Environmental harms are deeply embedded in the development of AI systems. Cobalt, silicon, lithium, and other raw materials need to be extracted, processed, assembled, and then packaged as various electronic products, in this case especially semiconductor chips, which are a crucial physical component of AI systems (Crawford 2021). The extraction of these resources in low governance areas is often associated with harms to the natural ecosystem, such as various forms of land, air, and water pollution. Further down the supply chain, data collection and storage, as well as model training require energy intensive computational power (Kouhizadeh et al. 2019; Taddeo et al. 2021). These are not negative externalities that can be offset only through direct investments in carbon credits, which is an industry favourite instrument to reach climate goals (Gabbatiss and Pearson 2023). Notable recent trends in "green computing" are exploring approaches to reducing the energy consumption of computational systems (Vale et al. 2022; Tornede et al. 2023), but given the growing compute requirements of cutting-edge AI systems, it is unclear how effective these mitigations will prove.

Following the same patterns, global AI supply chains externalise labour costs to low-income countries and/or low governance areas. For instance, in the Democratic Republic of Congo, where over 70% of the world's cobalt mining takes place, it is estimated that over 40,000 of the workers are children, some as young as six (Lawson 2021). Digital labour exploitation has also increased through the outsourcing of "data labelling", and Reinforcement Learning from Human Feedback (RLHF) tasks to low income countries (notably Kenya), which are essential to the maintenance and "alignment" of already deployed AI systems (Farina and Lavazza 2023; Perrigo 2023). Data are also exploited in the development of AI. For example, the advent of generative AI has led to major concerns around the theft of copyrighted, and creative labour, through the unlicenced scraping of datasets. Two authors have filed a class action lawsuit against OpenAI, claiming that the organisation has breached copyright law by training its models on their novels without permission (Creamer 2023). Personal data such as biometric data are also captured from unwitting individuals or purchased at a discount to diversify and improve AI datasets (Harvey and LaPlace 2021; Grant and Hill 2023).

Different types of harm cut across this supply chain. Individual-level harms directly impact individuals. For example, harm to a miner as part of the resource extraction process. Collective or group harms which occur when a group—either aligning with a traditional category or an ad hoc group—experiences a harm in their capacity as a member of that group, e.g. a group of workers, local or indigenous community (Floridi et al. 2016). In this article, *collective* harm refers to smaller, often informal groupings of people, *societal* harms refer to harms affecting larger-scale human groups bounded by persistent interactions, normally sharing the same spatial territory, typically subject to the same political authority and dominant cultural expectations, interests, and norms, e.g. a society harmed by the cumulative impact of AI model training on the environment (Floridi 2002, 2007; Smuha 2021).

A potential challenge to this framing is that many of the harms outlined here are not unique to AI, but common across industries, dependent on both pre-existing impacts of technology automation on supply chain labour (Acemoglu and Restrepo 2019), and broader geopolitical considerations affecting global supply chains. Many of the same harms occur in the production of laptops, or the creation and maintenance of search engines. Therefore, supply chain assurance

is an important but separate discussion to AI assurance.[9] Such objections are not incorrect but misguided. The fact that many AI supply chain harms are shared with broader digital supply chains is not an objection to the importance of the topic, but an invitation to see AI supply chain harms within this broader context. AI supply chain assurance cannot and should not be developed in isolation. Rather, it needs to rely on existing supply chain assurance, learn from it, adapt it, and be applied as part of this broader context (Brown 2023). For example, addressing the environmental impacts of deep learning systems will require addressing the novel, AI-specific impacts of model training, alongside the broader impacts of the data, hardware, and labour that these systems depend on. Integrating and adapting existing assurance approaches will be crucial. AI development has exacerbated many existing supply chain harms as well as raising unique problems. For example, AI foundation models which can be operated and adapted across contexts have drastically increased the complexity of supply chains in novel ways (Cobbe, Veale and Singh 2023), and created the potential for the exploitation of personal data and theft of creative labour on new levels (Gebru et al. 2023). Latonero and Agarwal (2021) have highlighted the dangers of poorly applied assurance in novel technology contexts, noting the failures of Facebook's (now Meta's) human rights impact assessment of their algorithm in Myanmar. Thus, identifying the gaps and challenges created by AI development will be crucial to adapting supply chain assurance mechanisms to AI.

3 Existing industry assurance practices and their application to AI

So far, measures to hold companies to account for their practices upstream in the AI supply chain have been missing from AI assurance policy and practice (Batarseh et al. 2021; CDEI 2023). Table 2 summarises key assurance mechanisms being used in other industries, highlighting their potential applications throughout the AI supply chain.

Impact assessments are currently limited in their applications upstream in the AI supply chain. Existing human rights impact assessment (HRIA) methodologies have largely been focused on supply chains with clear 'physical footprints' (Kernell et al. 2020), and recent human rights work in the AI assurance field has focused on algorithm impacts (Reisman et al. 2018; IEEE 2020; Leslie et al. 2022; Mantelero 2022). Between these areas, there is a lack of guidance for assessing the particular kinds of impacts that are associated with AI and other digital projects, products and services, across the supply chain (Kernell et al. 2020). Guidance is emerging on the application of HRIA in digital contexts (Kernell et al. 2020); however, the failings of Facebook's (now Meta's) HRIA in Myanmar, highlights the need for appropriate methods and criteria when applying HRIA in AI-specific contexts (Latonero and Agarwal 2021). The environmental impacts of AI have been explored further; yet, this has focused heavily on energy consumption and carbon footprint, over areas such as biodiversity, water consumption, rare earth mining and transport (OECD 2022). Furthermore, AI-specific methodologies for the indirect environmental impacts of AI compute (such as unsustainable changes in consumption patterns) are rare, and those that exist remain high-level and qualitative (OECD 2022). There is little evidence of Social Impact Assessment (SIA) applications specific to the AI supply chain. However, SIA holds important relevance for addressing harms produced through externalised AI supply chain labour. For example, SIA has been used effectively to build multistakeholder initiatives with rural and indigenous communities affected by resource extraction and infrastructural projects (Esteves et al. 2012). Such initiatives are crucial to addressing the harmful core/periphery inequality within AI supply chains (Cobbe, Veale and Singh 2023).

Sustainability reporting is common among firms that develop and deploy AI systems. Google alone published at least 17 reports related to their sustainable practices in 2021.[10] Despite this, the potential development harms associated with AI currently receive little attention (Sætra 2021). Existing ESG frameworks do not capture the nature of the sustainability-related impacts of AI adequately. As a result, companies are not incentivised to analyse and evaluate these impacts (Sætra 2023). Minkkinen et al. (2022) describes this lack of tools for tracing AI impacts across supply chains as a "critical bottleneck" of AI-related ESG evaluations. The field of corporate digital responsibility (CDR) provides a model for how this gap in ESG tools and evaluations may be filled. CDR has emerged as a subfield within CSR and has adapted the traditional CSR framework to support technology specific evaluation. CDR criteria guide organisations' operations with respect to the creation and operation of digital technology and data (Lobschat et al. 2021) and addresses the ways that digital technology has reshaped and extended traditional corporate responsibilities, for example through its pervasiveness, malleability, and scalability (Mihale-Wilson et al. 2022). A reporting and auditing framework focused on AI harms will need to build on this approach, highlighting and addressing impacts that are unique to, or exacerbated in AI supply chains (complexity, data exploitation, model

[9] To provide an example, the cobalt that is being extracted could be used for AI systems, but it could also be used for a variety of other, unrelated products.

[10] See Google's sustainability reporting catalogue here: https://sustainability.google/reports/.

Table 2 Existing assurance practices and their potential applications to the AI supply chain

Assurance mechanism	Description	Potential applications to AI supply chain harms
Impact assessment	Human rights impact assessment is a tool to evaluate the potential or actual impact of an organisation's strategy, practices, or products on people's human rights, throughout the supply chain	Assessing human rights impacts including exploitation of workers, marginalised or indigenous communities, and child labour in resource extraction and processing, engaging with affected stakeholders, and supporting risk mitigation through ongoing human rights due diligence
	Environmental impact assessment considers the likely implications of policies and projects on all aspects of the environment to develop appropriate responses to the issues identified	Assessing the environmental impact of an organisation's AI supply chain activities, from rare earth mining and harms to biodiversity, water use, chemical waste, transportation, to assessing carbon emissions from model training and operation
	Social impact assessment evaluates the ways in which people and communities interact with their social, cultural, economic, and physical surroundings	Assessing the impact of an organisations AI supply chain activities such as resource extraction and supply chain labour on affected communities, involving affected communities in decision-making, and establishing social value criteria for AI supply chain activities
Sustainability reporting and auditing	Reporting and auditing frameworks support the ongoing assessment and communication of impacts to stakeholders. Environmental, Social, Governance (ESG) frameworks are the most prominent ways in which corporations assess ethical risks related to their practices. Corporate social responsibility (CSR) has largely given way to ESG due to its broader scope and explicit inclusion of environmental impact	Providing an organising framework for industry to report on and audit their AI supply chain activities and impacts. These could include any environmental, societal, or governance related impacts, through reporting on more specific practices including human rights due diligence or the greenhouse gas protocol, under the ESG umbrella
Investor led assurance	In ESG investing, ESG issues are integrated into a company's investment analysis, with the aim of generating sustainable long term financial returns	ESG style investing criteria could be adapted to assess harms across the AI supply chain, including for AI specific harms such as model emissions, and the provenance of copyrighted data, driving more sustainable AI investment decisions, to incentivise sustainable AI business practices
Third-party certification schemes	Certification schemes are built on top of impact assessment, and reporting and auditing practices as a way of providing third-party endorsement, and communicating trustworthy practices throughout the supply chain	Certification schemes such as Fair Trade which focuses on labour rights, and Rainforest Alliance which more specifically environmental-oriented, could be developed or adapted to account for AI supply chain harms, to provide third party oversight of corporations' AI supply chain practices and their assurance

See Latonero and Agarwal (2021) for human rights impact assessment; Morgan (1999) for environmental impact assessment; Burdge & Vanclay (1996) for social impact assessment; Esty and Cort (2020) for sustainability reporting and Auditing; Boffo & Pantalano (2020) for sustainable investing; and Matus (2010) for certification

emissions), while ensuring parsimony with existing corporate, IT and digital governance initiatives (Mäntymäki et al. 2022).

Much like the gap in corporate reporting for AI-related risks, there is currently a significant gap in the application of ESG criteria to AI investment decisions. Recent hype around generative AI has led to "an investor wake-up call on artificial intelligence" (Bryan and Tett 2023). However, with the exception of Radical Ventures' recent Responsible AI for Startups (RAIS) framework (Morris and Rosenburg 2023), AI governance literature pays little attention to shaping the frameworks through which investors influence the development of AI systems, which are primarily focused on profit over pursuing societal or moral goals (Brynjolfsson and McAfee 2014; Minkkinen et al. 2022). One of the reasons for this gap is the different levels at which ESG criteria and AI auditing criteria are specified. While several metrics have been developed for AI assurance,[11] they focus on the technical assessment and verification of algorithmic systems, whereas ESG investment analysis is conducted at the organisational level. ESG investing focuses on company level processes, including organisational governance, supply chain due diligence and environmental sustainability (Boffo and Pantalano 2020). This focus positions ESG investing well for addressing salient AI supply chain harms. However, there is currently a lack of usable metrics to assess AI-related harms at the organisational level. ESG investing for AI sits in a gap between existing ESG frameworks which lack AI-specific guidance, and AI audit frameworks, which are more narrowly focused on algorithmic outputs (Minkkinen et al. 2022). Developing a robust approach to ESG investing in AI companies will require bridging the gap between approaches in these two areas.

Researchers and standard-setting bodies have also begun to consider how certification schemes can be applied to AI systems (Cihon 2019). Internationally, ISO and the IEC have formed a joint technical (JTC 1/SC 42) committee to develop standards specifically for AI,[12] and the IEEE is working on a certification scheme for Autonomous Intelligent Systems.[13] Standards are also set to play a key role in national level AI governance approaches, such as in the UK, EU, and China.[14] However, these initiatives are primarily focused on downstream development and deployment. One of the few certification schemes focused on AI upstream development is the Fairwork Foundation's Cloudwork Principles, used to provide third-party scoring of cloud work platforms (Fairwork 2021). A notable collaboration launched in May 2022 between the Responsible AI institute, the Standards Council of Canada (SCC), and EY, piloting a certification for ISO/IEC's AI management system standard (ISO 42001), which specifies the requirements and guidance for establishing and implementing an AI management system within the context of an organization (Shankar 2022). This pilot constitutes a potentially significant step forward in developing an AI certification approach that addresses a range of impacts upstream in the AI supply chain. Management system standards—including environmental, health and safety, risk, and quality management systems—are designed to work together, as part of a "harmonised structure".[15] Within this structure, ISO 42001 could provide an AI-specific lens through which organisations can access and interpret the whole suite of ISO management system standards to ensure ethical practices throughout the entire supply chain.

4 Towards a broader concept of, and approach to, AI assurance

Efforts to adapt and apply existing business practices to AI are a good starting point for addressing the various societal and development harms associated with AI. A nested approach to AI assurance is required to address AI supply chain harms within the context of existing IT, digital, and other supply chain assurance practices. However, the existing initiatives we have surveyed are either immature in their application to AI or unsuitable in their current form. Where assurance methods are applied to AI, they are disjointed, meaning that assurance is only partially provided, with some harms either side-lined or given little serious attention (Prunkl et al. 2021). Beyond this immaturity, the complexity of AI supply chains poses a serious coordination challenge for implementing a joined-up approach to AI assurance across development and deployment stages (Engler and Renda 2022). If these problems are not overcome, the effectiveness of industry-led assurance for addressing AI supply chain harms can be called into question. In the rest of this section, we consider these limitations, to provide

[11] The OECD catalogue of tools for trustworthy AI contains 101 metrics and methodologies for measuring and evaluating AI trustworthiness and AI risks: https://oecd.ai/en/catalogue/metrics.

[12] See here for more detail on ISO/IEC JTC1/SC 42: https://www.iso.org/committee/6794475.html.

[13] See here for more detail on the IEEE CertifAIed programme: https://engagestandards.ieee.org/ieeecertifaied.html.

[14] See here for more detail on standards in the UK AI white paper: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper; The EU AIA Act draft standardisation request https://ec.europa.eu/docsroom/documents/52376; and China's draft guidelines on standardising the AI industry: https://www.reuters.com/technology/china-issues-draft-guidelines-standardising-ai-industry-2024-01-17/.

[15] See here for more detail on management systems standards: https://www.iso.org/management-system-standards.html.

the grounds for our recommendations to policymakers and industry, where we propose how these limitations can be addressed to develop a broader approach to AI assurance.

4.1 The immaturity of current approaches

The immaturity of current assurance approaches highlights two, connected problems. Many existing assurance mechanisms lack the appropriated *specificity* to address AI harms satisfactorily. However, where approaches are more specific, there is a danger that assurance will become *disjointed*, leading to gaps. Considering the lack of specificity, evidence suggests that, so far, the application of assurance mechanisms to the AI supply chain has been ineffective, if not detrimental in some cases. Sætra's (2021) examination of Microsoft's sustainability reporting led to the conclusion that current sustainability reporting practices are not fit for addressing AI harms. Sætra notes that while their CSR reports seemingly acknowledge that there are issues related to AI, there is very little discussion about the specifics of the potential harms originating in Microsoft AI systems. Significant translation work will be required to ensure that existing frameworks satisfactorily capture harms which are distinctive to the AI supply chain.

The above example shows that increased specificity will be crucial to ensuring that existing assurance methods can mitigate harms effectively in the AI context. However, when it comes to specifying assurance mechanisms for the AI context, a disjointed approach could prove problematic. Focusing narrowly on one specific area of assurance could give the false pretence of ethical development, when in reality, several unethical practices are taking place. For instance, a company may be scored highly by the Fairwork Foundation for their ethical labour practices, while being environmentally unfriendly. At the same time, if a company chooses to focus on addressing a specific development harm, then they could inadvertently end up exacerbating problems in other areas. For instance, using more sustainable energy sources for model training would likely be reliant, at least in some part, on lithium-ion batteries. This could increase demand for exploitative and environmentally damaging cobalt mining, which is necessary for lithium-ion batteries (Mancini et al. 2021). These examples highlight the need for an organising framework for a broader AI assurance approach, enabling organisations to consider trade-offs between competing goals and values. Options for such a framework are set out in the recommendations in the following section.

4.2 Issues with supply chain complexity

Due to the complexity of AI supply chains, the task of translating assurance practices into the AI context and coordinating a more holistic approach to AI assurance will not be easy. Considering digital hardware as a related example, it took Intel over 4 years to understand its supply chain consisting of over 19,000 suppliers in 100 countries (Crawford and Joler 2018). AI supply chains add further layers of complexity. The AI supply chain can present itself in a wide variety of configurations that challenge more linear understandings of supply chains, with the European Centre for Policy Studies identifying seven basic configurations of AI supply chains of varying degrees of complexity (Engler and Renda 2022). In particular, the emergence of general-purpose AI has challenged understandings of accountability in ways that have not been widely addressed by the research community, exacerbating the 'many hands' problem of assigning accountability (Cobbe et al. 2023). The complex dependencies between companies that develop and deploy general purpose AI models, their functionality across a variety of sectors and risk contexts, as well as how open-source software communities contribute to model development, pose a serious challenge for assigning accountability (Küspert et al. 2023).

Similarly, due to the complexity of the impacts in the AI supply chain, it is often difficult to measure what 'good' looks like (Roberts et al. 2022). There are multiple different standards that can be followed for sustainability reporting, and there are currently no agreed upon measures for systematically and accurately measuring the carbon footprint of AI (Henderson et al. 2020; Cowls et al. 2021). Uncertainty over who has the skills, powers and access needed to measure and report on such risks throughout the AI supply chain provides an added layer of difficulty (Brown 2023). In the face of this complexity, organisations could be disincentivised from implementing assurance practices if there is a lack of clear standards of best practice, if assurance requirements are seen as being too complex or burdensome, and if there are not intermediary bodies tasked with checking and monitoring legal compliance and ethical alignment. The complex and interdependent nature of AI harms across the supply chain will require coordinated approach, which recognises the interdependencies between accountable supply chain actors.

4.3 The limitations of industry-led assurance

If the above challenges of immaturity and complexity can be dealt with, the limitations of industry-led assurance must still be considered. Considering these limitations is crucial to addressing the current narrow scope of AI assurance which is the focus of this article. In line with the ethical approach established in the introduction to this paper, the ability to extend the principle-based duties and obligations for AI assurance across the whole supply chain will rely on effectively incentivising and enforcing industry assurance practice. The role of policymakers in creating incentives and overcoming enforcement gaps will be crucial to extending

ethical duties and assurance obligations effectively throughout the AI supply chain. While existing initiatives such as ESG reporting and auditing do offer a more joined-up approach to assuring the trustworthiness of systems, products, or processes, they can also suffer from limitations of scope, due to their focus on issues which are relevant to the financial performance of a company (Minkkinen et al. 2022). Traditional ESG investing focuses on issues that are *financially material* rather than ethical values or social impact goals. In the traditional framework, ESG issues are considered to the extent that they are relevant to the financial performance of an investment (Sandberg et al. 2009). Policymakers will need to play a significant role in incentivising a stronger impact focus, if ESG investing is to drive more ethical outcomes. In the following section, we offer recommendations as to how this deficiency can be overcome in ESG reporting and investing.

Policymakers will also need to take a further role in incentivising the use of assurance mechanisms which are currently, largely voluntary. Without enforcement, they will lack the incentives to change the behaviour of organisations that benefit from exploitative practices. For example, the UK's Financial Conduct Authority (FCA) have recently warned banks about "greenwashing" and "conflicts of interest" in sustainability linked loans, which have included weak sustainability-related terms and low-ambition sustainability targets (Bryan 2023). Policy-driven incentives and enforcement, including trusted third-party certification will be necessary to strengthen organisations' commitments, by creating a competitive market for demonstrating verifiable sustainable practices. Policymakers will also have to recognise the limitations of industry assurance mechanisms, and where market forces could create conflicts of interest leading to unethical practices (Floridi 2019). To remedy this, policymakers should define where non-market assurance mechanisms will be required to supplement gaps or potential conflicts of interest in an industry-led approach.

5 Recommendations for a broader approach to AI assurance

Based on the arguments and evidence presented above, we offer a set of recommendations for policymakers and industry actors, which address the limitations described and provide the building blocks for a broader approach to AI assurance. We consider stakeholder responsibilities for implementing these recommendations, noting where unintended consequences could arise, and how they can be mitigated.

5.1 For policymakers

5.1.1 Policymakers should incentivise the development of an AI sustainability impact assessment framework to coordinate ex ante risk mitigation across the AI supply chain.

Human rights, environmental and social impact assessment each have distinctive strengths, which can help to extend current deployment focused approaches to AI impact assessment, to address issues throughout the AI supply chain. The human rights impact assessment and due diligence process offers a focus on supply chain human rights harms, and a firm normative basis in human rights duties to extend current narrow AI assurance approaches. Building on deployment-focused human rights impact assessments for AI (Leslie et al. 2022; Mantelero 2022; Reisman et al. 2018), policymakers should incentivise the inclusion of existing supply chain focused human rights impact assessment approaches. Policymakers should consider extending the duties under existing AI regulations further upstream and provide guidance on how to link existing assurance requirements with supply chain due diligence. Environmental impact assessment (EIA), as well as bringing environmental issues into focus, can help to address two crucial blind spots within AI supply chains. Requirements for public disclosure and participation,[16] which are mandated in EIA legislation can increase the visibility and attribution of environmental harms and enable broader stakeholder participation, which can help to address the limited "accountability horizon" of supply chain actors (Metcalf et al. 2021). Social impact assessment can further supplement these approaches, bringing societal and collective harms into focus through engagement with project affected communities in the AI supply chain (Burdge and Vanclay 1996; Esteves et al. 2012).

Successful implementation of these approaches will require balancing comprehensiveness against potential administrative burden which if too high, could unintentionally inhibit the ability of organisations—particularly SMEs and organisations with fewer resources—to effectively mitigate harms. To achieve this balance, these approaches should be harmonised. Sustainability impact assessment, which was developed partly to remedy some of the insufficiencies of narrower impact assessment approaches, could provide the framework for this harmonisation. Sustainability impact assessments are designed to act as a soft policy instrument for developing integrated policies which take full account of the three sustainable development dimensions (Bond et al. 2012). Despite relatively few companies currently using

[16] Public participation is not limited to those considered at risk of individual harm and is aimed at protecting the interests of society more broadly.

sustainability impact assessments, they are commonly used by some government bodies. For instance, the European Commission's Directorate-General for Trade uses sustainability impact assessments to examine the impacts of major trade agreements.[17] Policymakers should build on these use cases to explore how sustainability impact assessment could integrate the diverse strengths of existing ex-ante impact assessments across AI development and deployment. Additionally, policymakers will need to determine appropriate standards and enforceable oversight mechanisms for sustainability impact assessments, to mitigate risks of them being used as an ethical façade or purely "tick-box" exercise.

5.1.2 Policymakers should incentivise the development of an AI-specific framework for ESG reporting, auditing, and investing, to coordinate a coherent approach to AI assurance throughout the supply chain

Due to ESG's integration of different aspects of sustainability, and the embeddedness of ESG practices throughout existing organisational processes, policymakers should incentivise the development of an AI adapted ESG framework to provide the overarching framework within which to organise and coordinate a broader approach to AI assurance. This framework would situate AI supply chain harms within existing digital supply chain assurance, while enabling organisations to consider specific areas where AI development and deployment exacerbate existing risks or generate new types of risks. Policymakers will need to promote the development of an AI-specific ESG framework, for example, by incentivising key players in setting ESG reporting guidelines such as the Global Reporting Initiative (GRI)[18] and the Sustainability Accounting Standards Board (SASB) as part of a multistakeholder collaboration.[19] Sætra's (2023) AI ESG Protocol could provide a basis for these developments. The protocol splits AI impacts into three scopes: (1) impacts related directly related to a company's core activities and governance; (2) the upstream consequences directly related to the entities supply chain; (3) the broader upstream and downstream impacts of the company's AI, data-base capabilities, assets, and activities. The AI ESG protocol is designed to be flexible and link to other standards and frameworks (Sætra 2022). Direction could also be taken from Esty and Cort (2020), who propose a two-tiered approach to AI reporting where tier one contains mandatory disclosure elements and tier two contains more industry (AI)-specific indicators.

However, if an ESG framework is to work at all, a far stronger normative basis will need to be established. Attention to *financially material* risks alone will be far too narrow to address the environmental, labour and data harms that emerge across the AI supply chain. Recognition of *double materiality* in the EU's Sustainable Finance Disclosure Regulation (The European Parliament and The Council of the European Union 2019), and Corporate Sustainability Reporting Directive (The European Parliament and The Council of the European Union 2022) offers a step in the right direction. Beyond this, incentivising alignment with human rights based due diligence, and environmental regulations which mandate broad public participation, will help provide a stronger foundation for mitigating harms. Sætra's (2023) work on an AI ESG reporting framework, based upon the UN sustainable development goals, offers an example for developing a stronger human rights focus within ESG for AI. It is imperative that policymakers recognise the important influence of investors in driving corporate sustainability reporting and auditing. While recognising the potential for industry pushback without appropriate levers or market buy-in, policymakers should consider legislative mechanisms to incentivise ESG investing in AI and consider financial incentives for investors taking AI-related ESG factors into consideration. Mandating a strong normative basis for investment beyond only financially material considerations will be crucial to ensuring AI adapted ESG guidance is suitable to enable impactful investment decisions.

5.1.3 Policymakers should incentivise a market for third party audit and certification, to promote trustworthy impact assessment, reporting and audit approaches

Certification bodies act as "regulatory intermediaries" providing third-party endorsement of an organisation's assurance practices and communicating this to other stakeholders throughout the supply chain (Abbott et al. 2017). This role is formalised in the "three lines of defence" model put forward by the Institute of Internal Auditors, which sets out a framework of consisting of three internal assurance layers: operational management, risk and compliance, and internal audit, which are supported by external audit and certification, and finally regulation.[20] In the approach we suggest, impact assessments, monitoring, and internal auditing offer three internal lines of mitigation for harms across the AI

[17] See here for more detail on the European Commission's sustainability impact assessments used in trade negotiations: https://ec.europa.eu/trade/policy/policy-making/analysis/policy-evaluation/sustainability-impact-assessments/#_methodology.

[18] See here for more information on the Global Reporting Initiative (GRI): https://www.globalreporting.org/.

[19] See here for more information on the Sustainability Accounting Standards Board (SASB) https://www.sasb.org/about/.

[20] See here for more information on the "three lines of defence" model for risk management: https://theiia.fi/wp-content/uploads/2017/01/pp-the-three-lines-of-defense-in-effective-risk-management-and-control.pdf.

supply chain. In addition, policymakers should incentivise third-party audits (Falco et al. 2021; Mökander et al. 2021a, b) and a certification ecosystem (Cihon et al. 2021; Genovesi and Mönig 2022) to provide independent checks on organisations' assurance practices and to create a competitive market for trust, whereby organisations are incentivised to be certified to demonstrate best practice. However, an important consideration when implementing a third-party AI assurance ecosystem will be ensuring that costs for certification are not prohibitive for SMEs and non-profits. Over reliance on costly third-party mechanisms, could reduce competition and choice in the market. The choice to require third party rather than self-certification should be determined proportionately considering the consequences of non-compliance.

Certification against technical standards such as ISO, IEC and IEEE standards offer promising route for developing a broader approach to AI assurance because these standards are particularly influential in global trade and procurement.[21] While technical standards are voluntary, they often serve as de facto mandatory requirements due to their inclusion in trade agreements to signal that an organisation is following a "gold standard". Technical standards and certification have been a major focus for AI amongst policymakers in China,[22] the EU,[23] the US,[24] and the UK[25]; however, these efforts have mostly been confined to deployment harms. Policymakers must incentivise standards development organisation such as ISO, IEC, and IEEE to develop a more joined up approach to AI standardisation across the global AI supply chain. Certification approaches will need to account for the breadth of potential harms throughout the supply chain and be sensitive to a changing risk landscape. To address this challenge, use of the ISO 42001 AI management system standard certification should be incentivised in AI regulatory documents,[26] to enable organisations to coordinate existing environment, health, and safety management systems in the context of their AI-related operations.

6 Policymakers should consider where industry assurance approaches are insufficient for mitigating AI harms across the supply chain and develop or incentivise policy solutions to remedy these insufficiencies.

Independent monitoring mechanisms will play an important role in supplementing industry-led approaches to protect societal interests. Policymakers will need to learn from adjacent domains to develop public sector monitoring mechanisms to remedy gaps in industry approaches and help to provide reliable information about harms across the supply chain to policymakers and the public. For example, the European Environmental Agency (EAA)[27] is an agency of the EU tasked with providing sound, independent information on the environment. The EEA aims to help to achieve significant and measurable improvements in Europe's environment through the provision of reliable information to policymakers and the public. This form of public monitoring helps to ensure that environmental harms are kept track of, narrowing the knowledge gap and providing information for citizens that they couldn't attain themselves. In the AI context, the European Commission has set up AI Watch as the "Commission Knowledge Service to Monitor the Development, Uptake and Impact of Artificial Intelligence for Europe".[28] AI Watch produces an annual AI Watch Index to assess indicators on the performance of the EU across various dimensions of AI relevant to policymaking (Joint Research Centre (European Commission), Nativi and De Nigris 2021). Policymakers need to build on these approaches, committing to funding the development of more integrated monitoring frameworks to make harms visible across the whole of the AI supply chain.

In addition to public sector monitoring tools, policymakers should also incentivise and promote the development of civil society toolkits,[29] and support local NGOs and communities to report on harms and resist unfair practices. Building on crucial work done so far by organisations such as ProPublica, who led the investigation of the harmful COMPAS recidivism algorithm (Larson et al. 2016), policymakers should also engage with independent reporters and investigative journalists to uncover and highlight harms across the AI supply chain. These external approaches should complement concrete policy action on addressing harms, rather than being relied on as an alternative.

[21] ISO standards are developed in accordance with the principles in the WTO's Technical Barriers to Trade Agreement, enabling their use in international trade agreements, see here for more detail: https://www.wto.org/english/tratop_e/tbt_e/tbt_info_e.htm.

[22] See China's Standards Strategy: https://www.gov.cn/zhengce/2021-10/10/content_5641727.htm.

[23] See The EU AI Act: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[24] See the NIST AI Risk Management Framework: https://www.nist.gov/itl/ai-risk-management-framework.

[25] See the AI Standards Hub: https://aistandardshub.org/.

[26] The GB framework enables a Secretary of state to designate standards for regulatory purposes, see here: https://www.bsigroup.com/en-GB/about-bsi/uk-national-standards-body/standards-and-regulation/.

[27] See here for more detail on the European Environmental Agency: https://www.eea.europa.eu/about-us.

[28] See here for more detail on AI Watch: https://ai-watch.ec.europa.eu/about_en.

[29] For instance, see the University of Washington's Algorithmic Equity Toolkit https://uwescience.github.io/AEKit-website/.

6.1 For industry

6.1.1 The development of industry-led assurance mechanisms should be shaped by analysis of how AI harms manifest across the supply chain, on the collective and societal level, as well as the individual level.

Attention to the collective, and societal nature of harms helps to inform the analysis of harms across the AI supply chain. For example, the collective dimension of labour exploitation foregrounds the core/periphery inequalities in AI supply chains which externalise labour to low-income countries and marginalised groups (Crawford 2021). Considering societal harm, the environmental impact of training a single AI model can seem insignificant, making it difficult (a) for any single individual to claim a harm, or (b) to hold any single organisation to account. However, the cumulative, global effect of model training has a significant climate impact which clearly harms broader societal interests in addressing climate change. Understanding the ways in which harms manifest across the supply chain will be crucial to developing appropriate and effective mitigations. In lieu of a full supply chain AI sustainability impact assessment, industry should draw from human rights, environmental, and social impact assessments, to identify and mitigate harms that emerge across the full scope of their AI supply chains.

One important additional resource which industry can look to is The IEEE 7010 *Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems of Human Wellbeing,* which aims to provide a holistic assessment of AI system impacts. The scope of this assessment extends beyond solely individual indicators, with the explicit goal to "increase and help safeguard human well-being at the individual, population, and societal levels." The standard explicitly recognises the relationship between human wellbeing and sustainability, and defines a broad variety of impact indicators, including environment, climate, and biodiversity. Further impact indicators cover culture, community and human settlements which generally receive little attention. Industry should use these indicators and extend consideration to resource extraction and processing stages in the AI supply chain.

Similarly, in the case of large language models, areas of good practice are emerging which are beginning to account for a broader range of harms. Cohere, along with Open AI and AI21 have developed a joint recommendation for language model deployment, which includes treating all labour in the AI supply chain with respect, highlighting the role of workers reviewing model outputs, and providing data labellers with the right to opt out of a given task (Cohere 2022). This example highlights a good start from industry; however, the recommendation does not extend to labour further up the supply chain. While we recognise that voluntary industry approaches alone should not be relied upon as an alternative to action from policymakers, in lieu of a policy framework, in the short-term it is incumbent on industry to build on these starting points, to extend their ethical concern to resource extraction and processing activities, and consider the full scope of harms to individuals, groups and society. It is also worth noting that the implementation of standards and frameworks could place an outsized burden on SMEs with fewer resources. As such, the development of industry frameworks should set out processes and approaches which are adaptable to the resources and needs of different sized organisations.

6.1.2 Industry practitioners should situate their AI assurance approaches within the broader context of their digital supply chain assurance activities, rather than seeing these as separate areas.

This paper has emphasised the need for a nested approach to addressing harms across the AI supply chain and developing a broader approach to AI assurance. Not all the harms that occur in the AI supply chain are unique to AI. Many resource extraction, processing and labour harms are apparent in digital and other supply chains. Building a broader approach to AI assurance should begin from within the context of digital supply chain assurance. Industry practitioners will need to adapt existing approaches to fill gaps where AI supply chains exacerbate existing harms, or novel harms arise. Industry practitioners should look to corporate digital responsibility practices as a model for developing technology specific requirements, within the context of broader corporate assurance structures (Herden et al. 2021).

Practitioners can look to guidance from Lobschat et al. (2021) who set out a CDR framework, embedding the CDR concept in the corporate context to provide accessible, concrete guidance for practitioners. Their framework sets out a lifecycle-based approach digital technology governance, addressing four stakeholder groups, and promotes the realisation of three layers of a CDR culture, which are required to mitigate risks and enable the benefits of digital technologies within corporate governance. Good practice can also be drawn from current industry efforts to embed CDR, such as PwC's CDR Building Bloxx approach, which draws together established standards for sustainability reporting such as the Global Reporting Initiative (GRI) framework and ISO standards, to structure an applicable framework to responsible digitalisation (Pauer and Rübner 2020). Building on these approaches, practitioners should monitor the development of the forthcoming IEEE 7010.1 *Recommended Practice for Environmental Social Governance (ESG) and Social Development Goal (SDG) Action Implementation and Advancing Corporate Sustainability.* This standard builds on IEEE 7010

Human wellbeing impact assessment, and advances existing work in CDR, providing recommendations to align AI focused policies and practices with an assessment of impacts under existing corporate governance approaches.

6.1.3 Investors in AI technologies should recognise their power to shape AI development and deployment, and use investment to incentivise companies to report on and mitigate their AI sustainability-related impacts.

Investors have significant financial power to influence the ways in which companies develop and deploy AI systems, however the AI assurance literature has so far largely ignored them (Minkkinen et al. 2022). The same issue around *financial materiality* and normative strength and scope arises for ESG investing as with ESG reporting. To take AI development harms seriously investors should consider adjacent strategies such as socially responsible investing (SRI) and impact investing to ensure ethical values play a central role in driving investment decisions, to achieve specific social or environmental returns on an investment, rather than purely financial returns (Minkkinen et al. 2022). A stronger social impact focus, or at least a recognition of *double materiality* will be needed to ensure that investments in AI are driving ethical development and deployment. It is also important to recognise that transparency and ultimately robust independent oversight, via policy intervention, will be required to ensure the fidelity of 'ethical' and 'sustainable' investments which have come under significant recent scrutiny.[30]

An example of emerging good practice is the Responsible AI for Start-ups (RAIS) framework, published by Radical Ventures (Morris and Rosenburg 2023). RAIS is an open-source framework, derived from the internal algorithm audit framework developed by Raji et al. (2020), designed to help VCs assess early-stage AI companies and technologies across areas of social and ethical impact, regulatory compliance, and technical risk.[31] The RAIS acts as a risk register for investment organisations to identify risks and potential threats, to help guide investment decisions. The RAIS explicitly requires investors to describe vulnerabilities associated with the company's supply chain relationships and vulnerabilities related to the company's management team. The framework also asks investors to describe the likelihood for large energy consumption and other environmental impacts. Using this framework as a starting point, investors should adapt these considerations to their own portfolios and existing responsible investing principles. This will help to ensure that responsible practices are embedded from the initial stages of AI system development.

7 Conclusion

Embracing a broader approach to AI assurance can enable organisations to operationalise their normative commitments by (a) helping to identify potential harms so that different stakeholders can make more informed decisions and (b) contributing to a transparent and honest public debate around what trade-offs are legally permissible, economically viable, and publicly justifiable.

In this article, we have argued that harms across the whole AI supply chain should be taken seriously as part of AI assurance, and not considered as a separate issue. The fact that many AI supply chain harms are shared with broader digital supply chains demands that AI supply chain harms are addressed within this broader context, relying on existing supply chain assurance, learning from it, adapting it, and being applied as part of this broader context (Brown 2023). We have argued for the integration of four areas of supply chain assurance: ex ante impact assessment, ex post reporting and auditing, responsible investing, and third-party certification. Our assessment of these assurance mechanisms shows that they are currently either immature or disjointed in their application to AI supply chains.

Building on existing work, ESG offers a potential organising framework to develop and apply impact assessment, auditing and reporting, responsible investing, and certification approaches across the AI supply chain, and to balance trade-offs between competing values. Both policymakers and industry practitioners have a responsibility to adapt, streamline, and integrate these assurance practices to ensure that they are made suitable to address harms across the full scope of the AI supply chain.

Data availability Not applicable for this paper.

[30] See here for examples: https://www.ft.com/content/ac10773a-a975-11e9-b6ee-3cdf3174eb89; https://www.ft.com/content/ae78c05a-0481-4774-8f9b-d3f02e4f2c6f.

[31] See here for the Radical Ventures RAIS-Framework github: https://github.com/radicalventures/RAIS-Framework?search=1.

References

Abbott K, Levi-Faur D and Snidal D (2017) Introducing regulatory intermediaries. https://doi.org/10.1177/0002716217695519. Accessed 4 July 2023

Acemoglu D, Restrepo P (2019) Automation and new tasks: how technology displaces and reinstates labor. J Econ Perspect 33(2):3–30. https://doi.org/10.1257/jep.33.2.3

Batarseh FA, Freeman L, Huang C-H (2021) A survey on artificial intelligence assurance. J Big Data 8(1):60. https://doi.org/10.1186/s40537-021-00445-7

Boffo and Pantalano (2020) ESG investing: practices, progress and challenges. https://www.oecd.org/finance/ESG-Investing-Practices-Progress-Challenges.pdf

Bommasani R et al (2022) 'On the opportunities and risks of foundation models'. Preprint at http://arxiv.org/abs/2108.07258. Accessed 8 Dec 2023

Bond A, Morrison-Saunders A, Pope J (2012) Sustainability assessment: the state of the art. Impact Assess Project Appraisal 30(1):53–62. https://doi.org/10.1080/14615517.2012.661974

Brown I (2023) Expert explainer: allocating accountability in AI supply chains. https://www.adalovelaceinstitute.org/resource/ai-supply-chains/. Accessed 4 July 2023

Bryan K and Tett G (2023) An investor wake-up call on artificial intelligence, Financial Times. https://www.ft.com/content/f0b04f43-8e75-4745-b1db-530959dfab06. Accessed 6 Dec 2023

Bryan K (2023) FCA warns banks over "greenwashing" in sustainable loans, Financial Times. https://www.ft.com/content/10c3e16b-d1c7-4f76-a2f8-b92d54b1e2a7. Accessed 12 Nov 2023

Brynjolfsson E and McAfee A (2014) The second machine age: work, progress, and prosperity in a time of brilliant technologies. New York, NY, US: W W Norton & Co (The second machine age: Work, progress, and prosperity in a time of brilliant technologies), p 306

Burdge RJ, Vanclay F (1996) Social impact assessment: a contribution to the state of the art series. Impact Assess 14(1):59–86. https://doi.org/10.1080/07349165.1996.9725886

CDEI (2021) The roadmap to an effective AI assurance ecosystem. The centre for data ethics and innovation. https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem. Accessed 14 Dec 2021

CDEI (2023) CDEI portfolio of AI assurance techniques, GOV.UK. https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques. Accessed 4 July 2023

Cihon P (2019) Standards for AI governance: international standards to enable global coordination in ai research and development. https://www.fhi.ox.ac.uk/standards-technical-report/. Accessed 6 Dec 2023

Cihon P et al (2021) AI certification: advancing ethical practice by reducing information asymmetries. IEEE Trans Technol Soc 2(4):200–209. https://doi.org/10.1109/TTS.2021.3077595

Cobbe J, Veale M and Singh J (2023) Understanding accountability in algorithmic supply chains. In: *FAccT '23: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. https://doi.org/10.1145/3593013.3594073

Cohere (2022) Best practices for deploying language models, context by cohere. https://txt.cohere.com/best-practices-for-deploying-language-models/. Accessed: 12 Nov 2023

Cowls J et al (2021) The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. AI Soc. https://doi.org/10.1007/s00146-021-01294-x

Crawford K (2021) Atlas of AI. Yale University Press. https://yalebooks.yale.edu/9780300264630/atlas-of-ai. Accessed 6 June 2023

Crawford K and Joler V (2018) Anatomy of an AI system: the amazon echo as an anatomical map of human labor, data and planetary resources, anatomy of an AI system. http://www.anatomyof.ai. Accessed 14 Decem 2021

Creamer E (2023) Authors file a lawsuit against OpenAI for unlawfully "ingesting" their books. *The Guardian*. https://www.theguardian.com/books/2023/jul/05/authors-file-a-lawsuit-against-openai-for-unlawfully-ingesting-their-books. Accessed 10 Aug 2023

Creemers R, Webster G and Toner H (2022) Translation: internet information service algorithmic recommendation management provisions—effective March 1, 2022. *DigiChina*. https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/. Accessed 12 Nov 2023

Digital Watch Observatory (2024) 'India's IT ministry issues advisory on approval and labelling of AI tools | Digital Watch Observatory'. https://dig.watch/updates/indias-it-ministry-issues-advisory-on-approval-and-labelling-of-ai-tools. Accessed 5 Apr 2024

DSIT (2023) A pro-innovation approach to AI regulation, GOV.UK. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper. Accessed 12 Nov 2023

Engler A and Renda A (2022) Reconciling the AI value chain with the EU's artificial intelligence act. https://www.ceps.eu/download/publication/?id=37654&pdf=CEPS-In-depth-analysis-2022-03_Reconciling-the-AI-Value-Chain-with-the-EU-Artificial-Intelligence-Act.pdf. Accessed 16 June 2023

Esteves AM, Franks D, Vanclay F (2012) Social impact assessment: the state of the art. Impact Assess Project Appraisal 30(1):34–42. https://doi.org/10.1080/14615517.2012.660356

Esty DC, Cort T (2020) Values at work: sustainable investing and ESG reporting. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-55613-6

European Commission (2021) *Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206. Accessed 12 Nov 2023

Fairwork (2021) *Cloudwork (Online Work) Principles*. https://fair.work/en/fw/principles/cloudwork-principles/. Accessed 12 Nov 2023

Falco G et al (2021) Governing AI safety through independent audits. Nat Mach Intell 3(7):566–571. https://doi.org/10.1038/s42256-021-00370-7

Farina M, Lavazza A (2023) ChatGPT in society: emerging issues. Front Artif Intell. https://doi.org/10.3389/frai.2023.1130913

Farina M et al (2022) AI and society: a virtue ethics approach. AI Soc. https://doi.org/10.1007/s00146-022-01545-5

Floridi L (2002) Information ethics: an environmental approach to the digital divide. https://philarchive.org/rec/FLOIE. Accessed 16 June 2023

Floridi L (2007) 'Global Information Ethics: The Importance of Being Environmentally Earnest. IJTHI 3:1–11. https://doi.org/10.4018/978-1-59904-813-0.ch015

Floridi L (2019) Translating principles into practices of digital ethics: five risks of being unethical. Philos Technol 32(2):185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi L, Taylor L, van der Sloot B (2016) Group privacy: new challenges of data technologies. Springer, Cham

Freeman L et al (2022) The path to a consensus on artificial intelligence assurance. Computer 55(3):82–86. https://doi.org/10.1109/MC.2021.3129027

Future of Privacy Forum (2017) 'Unfairness by algorithm: distilling the harms of automated decision-making - future of privacy forum. https://fpf.org/blog/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/. Accessed 6 June 2023

Gabbatiss J and Pearson T (2023) *Analysis: how some of the world's largest companies rely on carbon offsets to 'reach net-zero'*, *Carbon Brief*. https://interactive.carbonbrief.org/carbon-offsets-2023/companies.html. Accessed 27 Oct 2023

Gebru T (DAIR), Bender EM (University of Washington), Angelina McMillan-Major (University of Washington), Margaret Mitchell (Hugging Face) (no date). In: *Statement from the listed authors of Stochastic Parrots on the "AI pause" letter*. https://www.dair-institute.org/blog/letter-statement-March2023. Accessed 1 May 2023

Genovesi S, Mönig JM (2022) Acknowledging sustainability in the framework of ethical certification for AI. Sustainability 14(7):4157. https://doi.org/10.3390/su14074157

Government of the Russian Federation (2019) *National strategy for the development of artificial intelligence for the period until 2030*. https://base.garant.ru/72838946/. Accessed 5 Apr 2024

Grant N and Hill K (2023) 'Google's photo app still can't find gorillas. And neither can Apple's. *The New York Times*. https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html. Accessed 27 Oct 2023

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

Hagendorff T (2021) 'Blind spots in AI ethics. AI Ethics. https://doi.org/10.1007/s43681-021-00122-8

Harvey A and LaPlace J (2021) *Exposing.ai: MS-Celeb-1M (MS1M)*, *Exposing.ai*. https://exposing.ai/datasets/msceleb/. Accessed 26 Oct 2023

Henderson P et al (2020) Towards the systematic reporting of the energy and carbon footprints of machine learning. J Mach Learn Res 21:1–43

Herden CJ et al (2021) "Corporate digital responsibility": new corporate responsibilities in the digital age. Sustain Manag Forum Nachhaltigkeits Manag Forum 29(1):13–29. https://doi.org/10.1007/s00550-020-00509-x

IEEE (2020) IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being. IEEE. https://doi.org/10.1109/IEEESTD.2020.9084219

Joint Research Centre (European Commission), Nativi S and De Nigris S (2021) *AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework*. Publications Office of the European Union, LU. https://doi.org/10.2760/376602. Accessed 30 Mar 2022

Kazim E, Koshiyama A (2020) AI assurance processes. SSRN Electron J. https://doi.org/10.2139/ssrn.3685087

Kernell L, Veiberg C and Jacquot C (2020) *Guidance on human rights impact assesssment of digital activities*, *Business & Human Rights Resource Centre*. https://www.business-humanrights.org/en/latest-news/danish-institute-for-human-rights-publishes-guidance-for-businesses-other-actors-in-the-digital-ecosystem-on-how-to-conduct-human-rights-impact-assessment-of-digital-activities/. Accessed 4 July 2023

Kingsman N et al (2022) Public sector AI transparency standard: UK Government seeks to lead by example. Discov Artif Intell 2(1):2. https://doi.org/10.1007/s44163-022-00018-4

Kolk A (2004) A decade of sustainability reporting: developments and significance. Int J Environ Sustain Dev 3(1):51–64. https://doi.org/10.1504/IJESD.2004.004688

Kouhizadeh M, Sarkis J, Zhu Q (2019) At the Nexus of blockchain technology, the circular economy, and product deletion. Appl Sci 9(8):1712. https://doi.org/10.3390/app9081712

Küspert S, Moës N and Dunlop C (2023) *The value chain of general-purpose AI*. https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/. Accessed 16 June 2023

Larson J, Mattu S, Kirchner L, Angwin J (2016) How we analyzed the COMPAS recidivism algorithm. ProPublica 9(1), 3–3

Latonero M and Agarwal A (2021) *Human rights impact assessments for AI: learning from Facebook's failure in Myanmar*. https://carrcenter.hks.harvard.edu/publications/human-rights-impact-assessments-ai-learning-facebook/E2/80/99s-failure-myanmar

Lawson MF (2021) *The DRC Mining Industry: Child Labor and Formalization of Small-Scale Mining*, *Wilson Center*. Available at: https://www.wilsoncenter.org/blog-post/drc-mining-industry-child-labor-and-formalization-small-scale-mining. Accessed 7 Mar 2022

Leslie D et al (2022) Human rights, democracy, and the rule of law assurance framework for AI systems: a proposal. https://doi.org/10.5281/zenodo.5981676

Lobschat L et al (2021) Corporate digital responsibility. J Bus Res 122:875–888. https://doi.org/10.1016/j.jbusres.2019.10.006

Mancini L et al (2021) Assessing impacts of responsible sourcing initiatives for cobalt: insights from a case study. Resour Policy 71:102015. https://doi.org/10.1016/j.resourpol.2021.102015

Mantelero A (2022) Beyond data: human rights. Eth Soc Impact Assess AI. https://doi.org/10.1007/978-94-6265-531-7

Mäntymäki M, Minkkinen M, Birkstedt T et al (2022) Defining organizational AI governance. AI Ethics 2, 603–609. https://doi.org/10.1007/s43681-022-00143-x

Matus KJM (2010) Certifiably sustainable?: the role of third-party certification systems: report of a workshop. The National Academies Press, Washington

Matus KJM, Veale M (2022) Certification systems for machine learning: lessons from sustainability. Regul Gov 16(1):177–196. https://doi.org/10.1111/rego.12417

Metaxa D et al (2021) Auditing algorithms: understanding algorithmic systems from the outside. Found Trends ® Hum Comput Interact 14(4):272–344. https://doi.org/10.1561/1100000083

Metcalf J et al (2021) Algorithmic impact assessments and accountability: the co-construction of impacts. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery (FAccT '21), New York, pp 735–746. https://doi.org/10.1145/3442188.3445935

Mihale-Wilson C, Hinz O, van der Aalst W, Weinhardt C (2022) Corporate digital responsibility: relevance and opportunities for business and information systems engineering. Bus Inf Syst Eng 64(2), 127–132

Minkkinen M, Niukkanen A, Mäntymäki M (2022) 'What about investors? ESG analyses as tools for ethics-based AI auditing. AI Soc. https://doi.org/10.1007/s00146-022-01415-0

Mökander J (2023) Auditing of AI: legal, ethical and technical approaches. Digital Soc 2(3):49. https://doi.org/10.1007/s44206-023-00074-y

Mökander J, Axente M et al (2021a) 'Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. Minds Mach. https://doi.org/10.1007/s11023-021-09577-4

Mökander J, Morley J et al (2021b) Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. Sci Eng Ethics 27(4):44. https://doi.org/10.1007/s11948-021-00319-4

Morgan RK (1999) Environmental impact assessment: a methodological approach. Springer Science & Business Media, Berlin

Morley J et al (2020) From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26(4):2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morris L and Rosenburg A (2023) *Underwriting responsible AI: venture capital needs a framework for AI investing*, *radical ventures*. https://radical.vc/underwriting-responsible-ai-venture-capital-needs-a-framework-for-ai-investing/. Accessed 12 Nov 2023

OECD (2022) Measuring the environmental impacts of artificial intelligence compute and applications: the AI footprint. OECD, Paris. https://doi.org/10.1787/7babf571-en

Pauer A and Rübner K (2020) *CDR building Bloxx*, *Corporate digital responsibility*. https://corporatedigitalresponsibility.net/f/corporate-digital-responsibility-cdr-building-bloxx. Accessed 12 Nov 2023

Perrigo B (2023) *Exclusive: the $2 per hour workers who made ChatGPT Safer*, *time*. https://time.com/6247678/openai-chatgpt-kenya-workers/. Accessed 7 July 2023

Prunkl CEA et al (2021) Institutionalizing ethics in AI through broader impact requirements. Nat Mach Intell 3(2):104–110. https://doi.org/10.1038/s42256-021-00298-y

Raji ID et al (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery (FAT* '20), New York, pp 33–44. https://doi.org/10.1145/3351095.3372873

Reisman D et al (2018) Algorithmic impact assessments report: a practical framework for public agency accountability. In: *AI Now Institute*. https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2. Accessed 6 June 2023

Roberts H et al (2022) Artificial intelligence in support of the circular economy: ethical considerations and a path forward. SSRN Electron J. https://doi.org/10.2139/ssrn.4080909

Roberts H, Cowls J et al (2023) Governing artificial intelligence in China and the European Union: comparing aims and promoting ethical outcomes. Inf Soc 39(2):79–97. https://doi.org/10.1080/01972243.2022.2124565

Roberts H, Babuta A, Morley J, Thomas C, Taddeo M, Floridi L (2023) Artificial intelligence regulation in the United Kingdom: a path to good governance and global leadership?. Inter Policy Rev 12(2). https://doi.org/10.14763/2023.2.1709

Ryan M, Stahl BC (2020) Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc 19(1):61–86. https://doi.org/10.1108/JICES-12-2019-0138

Sætra HS (2021) A framework for evaluating and disclosing the ESG related impacts of AI with the SDGs. Sustainability 13(15):8503. https://doi.org/10.3390/su13158503

Sætra HS (2022) AI for the sustainable development goals. CRC Press

Sætra HS (2023) The AI ESG protocol: evaluating and disclosing the environment, social, and governance implications of artificial intelligence capabilities, assets, and activities. Sustain Dev 31(2):1027–1037. https://doi.org/10.1002/sd.2438

Sandberg J et al (2009) The heterogeneity of socially responsible investment. J Bus Ethics 87(4):519–533. https://doi.org/10.1007/s10551-008-9956-0

Shankar V (2022) *RAI Institute Launches First-of-Its-Kind AI Certification Pilot with Standards Council of Canada*, *RAI Institute*. https://www.responsible.ai/post/raii-launches-first-of-its-kind-ai-certification-pilot-with-standards-council-of-canada-why-this-m. Accessed 3 July 2023

Sheehan M and Du S (2022) *What China's algorithm registry reveals about AI governance*, *Carnegie endowment for international peace*. https://carnegieendowment.org/2022/12/09/what-china-s-algorithm-registry-reveals-about-ai-governance-pub-88606. Accessed 4 Apr 2024

Shelby R et al (2022) Sociotechnical harms: scoping a taxonomy for harm reduction. Preprint at http://arxiv.org/abs/2210.05791. Accessed 29 Nov 2022

Smart Nation and Digital Government Office (2023) *National AI Strategy*. https://www.smartnation.gov.sg/nais/. Accessed 5 Apr 2024

Smuha NA (2021) Beyond the individual: governing AI's societal harm. Internet Policy Rev. https://doi.org/10.14763/2021.3.1574

Taddeo M et al (2021) Artificial intelligence and the climate emergency: opportunities, challenges, and recommendations. One Earth 4(6):776–779. https://doi.org/10.1016/j.oneear.2021.05.018

The European Parliament and The Council of the European Union (2019) *Regulation (EU) 2019/2088 of the European Parliament and of the Council of 27 November 2019 on sustainability-related disclosures in the financial services sector (Text with EEA relevance)*, *OJL*. http://data.europa.eu/eli/reg/2019/2088/oj/eng. Accessed 12 Nov 2023

The European Parliament and The Council of the European Union (2022) *Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting (Text with EEA relevance)*, *OJ L*. http://data.europa.eu/eli/dir/2022/2464/oj/eng. Accessed 12 Nov 2023

Tornede T et al (2023) Towards green automated machine learning: Status quo and future directions. J Artif Intell Res 77:427–457. https://doi.org/10.1613/jair.1.14340

Triolo P (2023) ChatGPT and China: how to think about large language models and the generative AI race. In: *The China Project*. https://thechinaproject.com/2023/04/12/chatgpt-and-china-how-to-think-about-large-language-models-and-the-generative-ai-race/. Accessed 4 Apr 2024

Tsamados A et al (2021) The ethics of algorithms: key problems and solutions. In: Floridi L (ed) Ethics, governance, and policies in artificial intelligence. Springer International Publishing (Philosophical Studies Series), Cham, pp 97–123. https://doi.org/10.1007/978-3-030-81907-1_8

Vale Z et al (2022) Green computing: a realistic evaluation of energy consumption for building load forecasting computation. J Smart Environ Green Comput 2:34–45. https://doi.org/10.20517/jsegc.2022.06

Weidinger L et al (2022) Taxonomy of risks posed by language models. In: 2022 ACM conference on fairness, accountability, and transparency. FAccT '22: 2022 ACM Conference on fairness, accountability, and transparency. ACM, pp 214–229. https://doi.org/10.1145/3531146.3533088.