AI and Ethics (2023) 3:451–468
https://doi.org/10.1007/s43681-022-00171-7

ORIGINAL RESEARCH

Operationalising AI governance through ethics-based auditing: an industry case study

Jakob Mökander[1] · **Luciano Floridi**[1,2]

Received: 17 March 2022 / Accepted: 29 April 2022 / Published online: 31 May 2022

Abstract
Ethics-based auditing (EBA) is a structured process whereby an entity's past or present behaviour is assessed for consistency with moral principles or norms. Recently, EBA has attracted much attention as a governance mechanism that may help to bridge the gap between principles and practice in AI ethics. However, important aspects of EBA—such as the feasibility and effectiveness of different auditing procedures—have yet to be substantiated by empirical research. In this article, we address this knowledge gap by providing insights from a longitudinal industry case study. Over 12 months, we observed and analysed the internal activities of AstraZeneca, a biopharmaceutical company, as it prepared for and underwent an ethics-based AI audit. While previous literature concerning EBA has focussed on proposing or analysing evaluation metrics or visualisation techniques, our findings suggest that the main difficulties large multinational organisations face when conducting EBA mirror classical governance challenges. These include ensuring harmonised standards across decentralised organisations, demarcating the scope of the audit, driving internal communication and change management, and measuring actual outcomes. The case study presented in this article contributes to the existing literature by providing a detailed description of the organisational context in which EBA procedures must be integrated to be feasible and effective.

Keywords Artificial intelligence · Auditing · Case study · Ethics · Governance · Industry · Practice

1 Introduction

Recent publications have identified ethics-based auditing (EBA) as a governance mechanism with the potential to help bridge the gap between principles and practice in AI ethics [1–6]. EBA is a structured process whereby an entity's present or past behaviour is assessed for consistency with relevant principles or norms.[1] The promise of EBA is underpinned by two ideas. The first is that procedural regularity and transparency contribute to good governance [7, 8]. The second is that proactivity in the design of AI systems help identify risks and prevent harm before it occurs [9].

Software audits is not a new phenomenon and establishing procedures to ensure consistency with predefined requirements is a fundamental aspect of systems engineering [10]. Nevertheless, seminal papers by Sandvig et al. [11] and Diakopoulos [12] helped popularise the idea that AI systems should be audited with regards to not only their technical performance but also their alignment with ethical values. A rich and growing academic literature on EBA has since emerged,[2] and a range of EBA procedures have been developed [13–19].

EBA has also received much attention from policymakers and private companies alike. National regulators like the UK Information Commissioner's Office have provided guidance on how to audit AI systems [18], and professional services firms like PwC, EY, Deloitte and KPMG have all developed auditing (or 'assurance') procedures to help clients ensure that the AI systems they design and

✉ Jakob Mökander
jakob.mokander@oii.ox.ac.uk

Luciano Floridi
luciano.floridi@oii.ox.ac.uk

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

[2] Department of Legal Studies, University of Bologna, Via Zamboni 33, 40126 Bologna, Italy

[1] Different researchers use different terms. [4] call it 'ethical audits'. However, we prefer the term EBA to avoid any confusion: we do not refer to a kind of auditing done ethically, but to auditing procedures for which ethics principles constitute the baseline.

[2] See [153] for a systematised review of previous literature on EBA.

deploy are legal, ethical, and safe [19–23]. In short, a new industry focussing on EBA is already taking shape.

Despite the surge in interest, important aspects of EBA—such as the feasibility and effectiveness of different auditing procedures—are yet to be substantiated by scientific research. For example, Raji and Buolamwini [6] suggest that internal audits *can* help check that the engineering processes involved in designing AI systems meet specific standards. Similarly, Brundage et al. [2] argue that external audits *can* help organisations verify claims about AI systems. These works have articulated important theoretical justifications for EBA. However, the affordances and constraints of EBA procedures can only be investigated and evaluated in applied contexts.

The literature on EBA contains few case studies: Buolamwini and Gebru [24] assessed the efficacy of external audits to address biases in facial recognition systems; Mahajan et al. [25] outlined a procedure to audit AI systems that replicate cognitive tasks in radiology workflows; and Kazim et al. [26] applied a systematic audit to algorithmic recruitment systems. However, there is still little understanding of how organisations implement EBA and what challenges they face in the process. This article addresses that gap by providing insights from a longitudinal industry case study.

Over a period of 12 months, we observed and analysed the activities of AstraZeneca (a biopharmaceutical company) as it prepared for and underwent an ethics-based AI audit. This article describes and discusses the findings from that study to make two contributions to the existing literature. First, it provides a descriptive account of how a large, decentralised, and R&D-driven company like AstraZeneca implements AI governance in practice. Second, by outlining the challenges and tensions involved in conducting a real-world AI audit, it identifies transferable best practices for how to develop EBA procedures. Taken together, these contributions support the objective of outlining the conditions under which EBA is a feasible and effective mechanism for operationalising AI governance.

Our findings suggest that the main difficulties organisations face when conducting AI audits mirror corporate governance challenges. In particular, organisations attempting to implement EBA must consider how to harmonise standards, demarcate the scope of the audit, define key performance indicators, and drive change management. These findings will not come as a surprise to management scholars. Yet efforts to operationalise AI governance are interdisciplinary in nature and the transfer of knowledge from different fields of study will be a key success factor to design and implement EBA procedures. This paper is thus aimed at computer scientists, ethicists, and auditors that develop EBA procedures as well as managers tasked with the implementation of corporate AI ethics principles.

The remainder of this article is structured as follows. Section 2 draws on previous research to establish the need for EBA. Section 3 introduces our case study by giving a descriptive account of AstraZeneca as an organisation as well as of the events leading up to the AI audit, which was conducted in Q4 2021. Section 4 describes AstraZeneca's AI audit in greater detail, situating it relative to previous research on EBA. Section 5 describes our methodology, which is based on participant observation and semi-structured interviews. Section 6 discusses our findings from the case study. Section 7 identifies limitations of our approach. Finally, Sect. 8 highlights current best practices and directions for future research.

2 The need to operationalise AI governance

AI holds great promise to support human development and prosperity [27]. Enabled by advances in machine learning (ML), access to computing power at decreasing costs, the growing availability of data, and the ubiquity of digital devices [28], AI systems can improve efficiency, reduce costs, and help solve complex problems [29].

The gains associated with AI technologies are not only economic but also social in nature. Take healthcare as an example. AI systems aid clinicians in medical diagnostics [30] and enable personalised treatments [31]. AI systems also drive healthcare service improvements through better forecasting [32]. In the pharmaceutical industry, the combination of pattern recognition for molecular structures and laboratory automation promises better and faster drug discovery and delivery processes [33]. In sum, using AI systems in the healthcare sector may allow humans to live more healthy lives while enabling societies to manage the rising costs associated with ageing populations [34].

However, the use of AI systems in the healthcare sector is coupled with ethical challenges [35, 36]. The use of AI systems may leave users vulnerable to discrimination and privacy violations [37]. It may also enable wrongdoing and erode human self-determination [38]. Many of these risks apply to AI systems generally. But how AI systems process health data is particularly delicate [39], since patients may be harmed by reputational damage and suboptimal care [40]. For example, recent studies have found racial biases in medical devices that provide pulse oximetry measurements [41].

While the adoption of AI systems have outpaced the development of governance mechanisms designed to address the associated ethical concerns [42], abstaining from using AI systems in sensitive areas of application is not the way forward [43]. As far as the use of AI in medicine is concerned, a 'precautionary approach' would likely cause significant social opportunity costs due to constraints that undermine the development of promising technologies,

drugs and treatments [44]. Moreover, AI systems are part of larger socio-technical systems that comprise other technical artefacts as well as human operators [45, 46]. No purely technical solution will thus be able to ensure that AI systems operate in ways that are ethically-sound [47, 48].

It is essential that public and private actors seeking to benefit from AI systems understand and address the varied ethical challenges associated with their use. Responding to this need, numerous governments and NGOs have proposed ethical principles that provide normative guidance to organisations designing and deploying AI systems [49, 50].[3] These guidelines tend to converge on five principles: beneficence, non-maleficence, autonomy, justice, and explicability [51].[4] This is encouraging. Yet principles alone cannot guarantee that AI systems are designed and used in ethically-sound ways. The apparent consensus around normative principles hides deep political tensions around interpreting abstract concepts like fairness and justice [52]. Moreover, translating principles into practice often requires trade-offs [53]. Most critically, the industry lacks useful tools to translate abstract principles into verifiable criteria [54].

Due to these constraints, technology-oriented companies have struggled with operationalising AI ethics. Fortunately, companies need not start from scratch: numerous translational mechanisms for AI governance have been proposed and studied [55–57]. These include *impact assessments lists* [58–60], *model cards* [61], *datasheets* [61–63], as well as *human-in-the-loop* protocols [64], *standards* and reporting guidelines for using AI systems [65–67], and the inclusion of broader *impact requirements* in software development processes [68].[5]

All these efforts are complementary and serve the overarching purpose of enabling effective corporate AI governance. That is important because private companies significantly influence regulatory methods and technological developments [69, 70]. This dependency on private actors is a double-edged sword. On the one hand, competing interests can undermine even well-intentioned attempts to translate principles into practice [71]. On the other hand, private companies have strong incentives to implement effective AI governance to improve numerous business metrics like regulatory preparedness, data security, talent acquisition, reputational management, and process optimisation [72, 73].

How AI systems are designed and used is a concern not only for individual organisations but also for society at large [74]. This insight has been reflected in recent regulatory developments. Both the EU Artificial Intelligence Act (AIA) [75] and the US Algorithmic Accountability Act of 2022 [76] constitute attempts to elaborate general legal frameworks for AI. Hard legislation can, if properly designed and enforced, address parts of the gap EBA procedures fill. For example, the AIA requires specific 'high-risk' AI systems to undergo so-called 'conformity assessments by the involvement of an independent third party' [77]. But most AI systems are not classified as 'high-risk' and will thus not subject to the requirements stipulated in the AIA. Moreover, the use of AI systems may be problematic and deserving of scrutiny even when not illegal. In short, there will always be room for more and better, post-compliance, ethical behaviour [78]. The 'ethics-based' approach studied in this article is thus compatible with—and complementary to—hard legislation.

3 AstraZeneca and AI governance

AstraZeneca is a multinational biopharmaceutical company headquartered in Cambridge, UK. It has an annual turnover of $26bn and employs over 76,000 people [79]. As an R&D-driven organisation, AstraZeneca discovers and supplies innovative medicines worldwide. Its core business is using science and innovation to improve health outcomes through more effective treatment and prevention of complex diseases.[6]

The biopharmaceutical industry has always been data-driven [80]. To develop new treatments, researchers follow the scientific method by building and testing hypotheses about the safety and efficacy of various treatments.[7] For example, AstraZeneca relies heavily on statistical analysis to probe the efficacy of candidate drugs in the research pipeline. Hence, AstraZeneca has long-established processes for data, quality, and safety management. However, how data can be collected, analysed, and utilised keeps changing [81]. By harnessing the power of AI systems, researchers can find new correlations and draw useful inferences from the growing availability of data.

Examples of use cases of AI systems within AstraZeneca are abundant. For example, the company uses biological insight knowledge graphs (BIKG) to improve drug discovery and development processes [82]. Using BIKG helps

[3] Recent and influential contributions include [154], [155], and [156].

[4] Healthcare practitioners will note the overlap here with the classical principles of bioethics [157].

[5] See [115] for an overview of available tools and methods to translate high-level AI ethics principles into practice.

[6] AstraZeneca is divided into three main therapy areas: Oncology; Cardiovascular, Renal and Metabolism; and Respiratory and Immunology diseases [158].

[7] The process of discovering and developing new drugs is long and complex: only a small proportion of molecules that are identified as a candidate drug are approved [159].

Table 1 AstraZeneca's principles for ethical data and AI usage

Principle	Operationalisation
Private and secure	We respect privacy and act in a manner compatible with intended data use
	We employ Data & AI Systems that are designed to be secure
Explainable and transparent	We are open about the use, strengths, and limitations of our Data & AI systems
	We ensure assumptions are clear, algorithms are appropriately documented, decisions are explainable, and processes to manage unanticipated consequences
Fair	We endeavour to use robust, inclusive datasets in our Data & AI systems
	We treat people and communities fairly and equitably in the design, process, and outcome distribution of our AI systems
Accountable	We apply governance proportional to the impact and risk of Data & AI systems
	We anticipate and mitigate the impact of potential unfavourable consequences of AI through testing, governance, and procedures
Human-centric and socially beneficial	Where Data & AI is involved, humans oversee the system and are accountable for driving clear, expected benefits to people and society
	We employ human-led governance over our AI systems. We respect human dignity and autonomy and strive to reflect this in our AI systems

synthesise, integrate, and leverage prior knowledge to gain new insights into disease characteristics and design smarter clinical trials [83]. AI systems are also used for fast and accurate medical image analysis. Using AI systems based on ML and image recognition cuts analysis time by over 30% and improves accuracy [84]. Moreover, AI systems help automate various tasks. For example, AstraZeneca use natural language processing to prioritise adverse event reports [85, 86]. Here, AI systems help classify events, separate outcomes by severity to enable appropriate action, thus leading to quicker response times and better patient experiences.

Despite excitement about these opportunities, AstraZeneca is conscious about risks associated with AI systems. As discussed in Sect. 2, these include concerns related to privacy, fairness, transparency and safety. In November 2020, AstraZeneca's board moved towards addressing these risks by publishing a set of Principles for Ethical Data and AI (henceforth, *ethics principles*. See Table 1 above). These stipulate that the use of data and AI systems should be private and secure; explainable and transparent; fair; accountable; as well as human-centric and socially beneficial [87].[8]

The primary aim of these *ethics principles* is to help employees and partners safely and effectively navigate the risks associated with AI systems.[9] However, for AstraZeneca, AI governance serves numerous additional purposes. To use AI systems in line with the overall company strategy helps realise synergies and maximise value creation. Moreover, the voluntary adoption and publication of corporate AI ethics principles strengthens AstraZeneca's brand.[10] Finally, the same internal processes that allow AstraZeneca to demonstrate adherence to its *ethics principles* also help it manage legal risks by ensuring compliance with existing legislation and anticipating forthcoming legislation.

These advantages are potential and not guaranteed. Principles alone cannot ensure that AI systems are designed and used in ways that are ethical [52]. Hence, AstraZeneca followed its commitment to its *ethics principles* by focussing on their implementation. However, doing so was not straightforward. AstraZeneca already had several related governance structures in place, e.g., with regard to quality and data management, corporate social responsibility (CSR), sustainability, and product safety. Furthermore, AstraZeneca is a decentralised organisation in which different business areas operate independently. This structure provides flexibility but complicates the agreement and enforcement of common standards and procedures.

Taking those considerations into account, AstraZeneca allowed each business area to develop their own AI governance structures to reflect local variations in objectives, digital maturity, and ways of working—as long as these align with the externally published *ethics principles*. To support local activities, however, four enterprise-wide initiatives were launched:

(1) The creation of an overarching *compliance document*
(2) The development of a *Responsible AI playbook*

[8] The process of formulating AstraZeneca's *ethics principles* involved numerous internal workshops and consultations with external experts and stakeholders.

[9] The *ethics principles* are thus to be seen as an extension of AstraZeneca's overarching organisational values.

[10] As noted by Slee [160], creating auditable algorithms and datasets is a promising avenue for organisations to bridge the presentation gab between brands and the AI systems they design and deploy.

(3) The establishment of an *AI resolution Board* and an internal *Responsible AI Consultancy Service*
(4) The commissioning of an *AI audit* conducted in collaboration with an independent party

First, a compliance document was created, breaking down each high-level principle into more tangible and actionable formulations. Table 1 illustrates how that document attempts to bridge the gap between principles and practice in AI ethics.

Second, a Responsible AI Playbook was developed to provide more detailed, end-to-end guidance on developing, testing, and deploying AI systems within AstraZeneca.[11] The Playbook is a continuously updated online repository directing AstraZeneca employees to relevant resources, guidelines and best practices. The Playbook also summarises the specific regulations applicable to different AI use cases.

Third, new organisational functions were established. Specifically, an AI resolution board was created to review 'high-risk' AI use cases and an internal Responsible AI Consultancy Service was launched to facilitate the sharing of best practices and to educate staff about the risks of using AI systems in different contexts. The Responsible AI Consultancy Service serves three objectives: providing ethical guidance; supporting the practical embedding of the *ethics principles*; and monitoring the governance of AI projects.

Fourth, and most relevant for our purposes, AstraZeneca underwent an AI audit. This audit constituted the research material for our case study, and framing it is the focus of the next section.

4 An 'ethics-based' AI audit

In Q4 2021, AstraZeneca underwent an AI audit. However, because the term 'AI audit' has been used in many different ways, some clarifications are needed to specify what we refer to in this case. The AI audit conducted within AstraZeneca was an ethics-based, process-oriented audit conducted in collaboration with an independent third party. The remainder of this section unpacks what this means in practice.

The audit was 'ethics-based' insofar as AstraZeneca's *ethics principles* constituted the baseline against which organisational practices were evaluated. In short, the audit concerned what ought to be done over and above existing legislation. Of course, AI audits can be employed by different stakeholders and for different purposes. For example, Brown et al. [1] distinguish between AI audits used (i) by regulators to assess whether a specific system meets legal standards; (ii) by providers looking to mitigate risks; and (iii) by other stakeholders wishing to make informed decisions about how they engage with specific companies. The AI audit conducted within AstraZeneca corresponds to (ii) since it was directed towards demonstrating adherence to voluntary codes of conduct.[12]

Further, AstraZeneca's audit was 'external' because it involved the commissioning of an independent third-party auditor. Specifically, the audit was coordinated by AstraZeneca's internal audit function and conducted by an external service provider.[13] In the literature, a distinction is often made between internal audits, based on self-assessment, and external audits conducted by expert organisations [88]. The latter tend to be limited by reduced access to internal processes [89]. However, involving external experts can address the confirmation bias that may prevent internal audits from recognising critical flaws [90]. By subjecting itself to external review, AstraZeneca thus got valuable feedback on how to improve its existing and emerging AI governance structures.[14]

A central idea underpinning EBA is that procedural regularity and transparency contribute to good governance. Hence, one aim of EBA is to create traceable documentation.[15] However, transparency must always be understood in context, i.e., with regards to a specific audience and intended purpose [91]. In AstraZeneca's case, the audit's audience was internal decision-makers, and its most obvious purpose was assessing the extent to which the *ethics principles* had been adopted.

Operationally, the audit conducted within AstraZeneca consisted of two types of activities: a high-level *governance audit* of organisational structures and processes and *in-depth audits* of specific projects that either develop or use AI systems. Here, it is worth mentioning that the subject matter of EBA can either be a process, an organisational unit, or a technical system.[16] These approaches are

[11] The Playbook was developed by AstraZeneca's R&D Data Office yet is accessible to everyone in the organisation.

[12] Oversight is critical to operationalise AI governance. In practice, this implies establishing evidence of how the AI systems were created and how they are operating [161].

[13] The company that conducted the AI audit is a leading professional services firm. In line with the non-disclosure agreement (NDA) for this research, its name is not disclosed. Instead, it is referred to as 'the external auditor'.

[14] Note that all the other three enterprise-wide activities conducted by AstraZeneca to operationalise AI governance (see Sect. 3) were internal in nature.

[15] As noted by Kroll [161], public documentation serves its function when, and largely because, its creation forces organisations to consider how to develop systems that can be presented in the best possible light.

[16] A consequence of viewing AI systems as parts of larger sociotechnical systems is that AI governance concern not only technical artifacts but also the organisations that develop or operate these [162].

not mutually exclusive but rather crucially complementary [92]. AstraZeneca's AI audit focussed on processes and people, i.e., on assessing (i) the soundness and completeness of organisational processes and (ii) the extent to which different organisational entities adhered to these processes. During technical audits, in contrast, AI systems' source codes can be reviewed [93] or, alternatively, the behaviour (i.e., outputs) of such systems can be tested for a wide range of different input values [94]. However, no technical audits of individual AI systems were conducted during AstraZeneca's AI audit.

In Sect. 6, we discuss lessons learned from this audit. However, qualitative findings are best interpreted in the context in which they emerged. Hence, before exploring our findings, the next section outlines how we collected and analysed our data.

5 Methodology: an industry case study

To investigate the feasibility and effectiveness of EBA, we adopted that 'pragmatic stance'[17] and conducted a *industry case study* [95, 96]. Specifically, we observed and analysed AstraZeneca's internal activities as it prepared for and underwent an AI audit. The case study was *longitudinal* [97] insofar as it lasted 12 months.[18] Three research questions (RQs) guided our research:

- *RQ1: How do industry firms integrate EBA with existing governance structures?*
- *RQ2: What challenges do industry firms face when attempting to implement EBA in practice?*
- *RQ3: What are best practices for how to prepare for and conduct EBA?*

With respect to these RQs, AstraZeneca's AI audit constituted what Merton [98] calls 'strategic research material' for at least two reasons. First, as an organisation that regularly harnesses data and AI systems for process automation and R&D purposes, AstraZeneca had practical concerns that overlapped with the theoretical problems we sought to address. Second, the timing was advantageous, in that we could follow the entire journey, from the publication of AstraZeneca's *ethics principles* (in November 2020) to the evaluation of the AI audit during Q4 2021.

Methodologically, the present study leveraged two qualitative research methods: *participant observation* and *semi-structured interviews*. The former, in which research is carried out through the researcher's direct participation, has a long history in organisational research [99]. It is particularly well-suited to making sense of organisational practices [100]. In this study, participant observation meant embedding ourselves in the organisation to observe the activities associated with preparing for and conducting the AI audit. This involved joining weekly meetings, reviewing working documents, and taking notes—not only of the audit's eventual findings but also of the points raised and decisions made along the way.

Specifically, we observed two types of meetings: *internal meetings*, in which AstraZeneca employees prepared for (or evaluated the results from) the AI audit, and *audit meetings*, in which external auditors asked questions to, and reviewed documentation provided by, AstraZeneca employees. Because AstraZeneca's employees are distributed internationally—and because of Covid-19-related travel restrictions—all meetings took place online.

In addition, we conducted semi-structured interviews [101] with different stakeholders involved in the audit. This allowed us to follow up on themes emerging from regular audit meetings and explore different actors' motivations and perspectives.[19] In total, we interviewed 18 people—some on several occasions. Rather than interviewing a predefined list of people, we used a snowballing technique [102] to recruit new interviewees. Nevertheless, we tried to interview representatives in different roles and strove for a balance between different genders, ethnicities, and educational backgrounds amongst our interviewees. Each interview lasted 1–2 h. To make the participants feel comfortable and avoid disturbing the flow of meetings, we did not record interviews, taking notes instead.

NVivo was used to import, code,[20] and analyse meeting notes. A *parallel research design* [103] was used, in which the participant observation and the semi-structured interviews were conducted simultaneously. This enabled an iterative process through which research findings could be triangulated [104], thereby minimising the risk of 'losing context' that is associated with qualitative coding [105].

Ethical approval for this research was granted by the Oxford Internet Institute's departmental research ethics committee. Access to AstraZeneca's internal processes and stakeholders was obtained through an agreement that leveraged an existing institutional relationship: JM's doctoral research is funded through a studentship provided by

[17] According to the American pragmatists (notably C.S. Peirce, William James, and John Dewey) theories should be judged by their success when applied practically to real-world situations [163, 164].

[18] Longitudinal case studies have long been used to observe how different governance mechanisms impact organisational practices. See e.g., [148].

[19] Another advantage of semi-structured interviews is that the conversation is directed to the problem under investigation rather than the researcher's preconceived interests [165].

[20] Here, 'code' refers to a word that assigns an essence-capturing attribute for a portion of language [166].

AstraZeneca.[21] A separate NDA was signed with the external auditor, allowing us to join relevant meetings to study the process. In all meetings, participants were informed about our presence and our research's purpose. No personal details were collected or stored during the research.

6 Lessons learned from AstraZeneca's 2021 AI audit

When analysing the data, we found that the answers to questions about how to best design and implement EBA procedures often hinge on decisions made earlier in the process of operationalising corporate AI governance. Hence, when presenting our findings, we start with high-level observations and proceed with increasing levels of specificity.

6.1 Balancing legitimate yet competing interests

A fundamental tension exists between the need for risk management, on the one hand, and incentives for innovation on the other.[22] This tension is particularly acute for R&D-driven organisations like AstraZeneca—both from an ethical and a financial point of view. For example, when developing new treatments, it is essential to monitor patient responses from a safety perspective. Hence, AstraZeneca trains AI systems to detect treatment response patterns and associate biophysical reactions with the safety risks of specific drugs [106]. Excessive red tape could hamper the development and adoption of such, potentially lifesaving, procedures. This shows that it is often not possible to 'err on the safe side'. Both the pharmaceutical industry and society at large have an obligation to put patients' care and safety first—and this means using innovative technologies to develop new drugs as well as to diagnose and intervene as early as possible in the course of a disease.

Similarly, from a financial perspective, R&D-oriented activities always carry risks since they involve trying new ideas—which often fail to progress.[23] But even 'failed' R&D projects inform pharmaceutical innovation [107]. Hence, risk per se is not undesirable from an organisational point of view. Rather, the priority is to define and control the risk appetite in different projects. From an auditing perspective, this dynamic has two direct implications. First, EBA procedures that duplicate existing governance structures, or are perceived as unnecessary, are unlikely to be feasible and effective. Second, post hoc EBA procedures that only highlight the risks associated with specific AI use cases are less likely to be adopted than continuous EBA procedures that help technology providers define and regulate technology-related risks.

6.2 Demarcating the material scope for AI governance

Another high-level observation concerns the difficulty to define the material scope of AI governance in general and EBA in particular. As is well known, there is no universally accepted definition of AI [108].[24] Nevertheless, every policy needs to define its material scope [109].[25] Consequently, when attempting to operationalise its *ethics principles*, AstraZeneca struggled to define the systems and processes to which they ought to apply. That is partly because both human decision-makers and AI systems have their own strengths and weaknesses [110] and partly because ethical tensions can sometimes be intrinsic to the decision-making tasks at hand [111].[26]

Within AstraZeneca, representatives from the internal audit function stressed that underinclusive definitions of AI may lead to potential risks going unnoticed and unmitigated. Other stakeholders, including some managers and statisticians from the IT and R&D departments, warned that overinclusive AI definitions risk adding unnecessary layers of governance to very well-established systems and processes. As one manager objected:

> *"We are not doing any AI projects. We are, of course, doing large scale analytics, but only using statistical techniques that have long been standard practice in the industry."* (P5)

[21] The studentship is funded by AstraZeneca but administered and paid out by the University. There have been no direct financial transactions between AstraZeneca and the researcher. The research is academically independent, and all views expressed in the article are those of the authors.

[22] Critically-oriented researchers often highlight AI systems' failures to stress the need for more regulation [167]. In contrast, techno-optimists point towards the gains such systems bring and caution against red tape [168].

[23] Pammolli et al. [169] analysed R&D activities related to drug development and found that over 70% of projects initiated between 2000 and 2009 had been terminated within one year.

[24] Some researchers use the term AI to refer to a type of agents that display some levels of autonomy, adaptability, and problem-solving capacity [170]. Others take AI to demarcate the set of computational techniques designed to approximate cognitive tasks [171]. Yet others use the term to describe the science and engineering of making specific machines [172].

[25] See [173] for a comparison of methods for classifying AI systems for governance purposes.

[26] As Bryson [174] notes, many problems associated with 'AI' have not so much been created as exposed by it.

To solve this tension, AstraZeneca did not try to define what AI *is*.[27] Instead, AstraZeneca's Responsible AI Playbook lists and exemplifies the functional capabilities of the systems to which their AI governance framework *applies*. For each functional capability (such as the ability to emulate cognitive tasks), the Playbook provides concrete examples. Amongst others, the Playbook states that statistical tests (e.g., a T-test) conducted during data analysis are outside AstraZeneca's AI governance framework's scope. In contrast, automated statistical tests informing decisions that impact humans (e.g., stratifying patients into different arms of a clinical trial) are within scope. A list of examples does not constitute a definition of AI, nor does it provide a sufficient basis on which to create an exhaustive inventory of an organisation's AI systems. Nevertheless, listing examples of use cases that are in (or out) of scope informs attempts to operationalise AI governance.

Furthermore, AstraZeneca adopted a risk-based approach, whereby the level of governance required for a specific system is proportionate to its risk level.[28] This means that systems within scope are classified as either low-, medium- or high-risk, depending on (i) the types of risk the system poses to humans and the organisation and (ii) the extent to which it makes autonomous decisions without human judgement. The approach taken is pragmatic[29] since it enables managers and developers to determine whether the *ethics principles* apply to specific systems. At the same time, the approach makes it difficult to assemble an inventory of an organisation's various AI systems. Without such an inventory, AstraZeneca's AI auditors depended on the business to identify and select relevant projects and systems for the in-depth audits.

The main takeaway here is that designing and implementing EBA procedures is intrinsically linked to the question of material scope. Until the material scope of AI governance is accepted throughout the organisation, any EBA procedure would struggle to produce verifiable claims.

6.3 Harmonising standards across decentralised organisations

A further challenge faced during the AI audit is rooted in the problem of ensuring harmonised standards across decentralised organisations. As mentioned, each business area within AstraZeneca operates independently. From an AI governance perspective, this implies that the business areas face different realities in terms of digital maturity, the type of AI systems employed, economic pressures, and employees' levels of training.

Consider the contrast between two functions within AstraZeneca: R&D and Commercial. First, there are operational differences. R&D routinely creates AI systems in-house to aid drug discovery and testing. An understanding of the statistical models underpinning different AI systems is therefore closely linked to R&D's core business. Within Commercial, sales representatives typically rely on data analytics software (like CSR systems or predictive modelling) as a means to an end. Second, there are structural differences. R&D relies on a centralised Data Office to manage and curate data. In contrast, analytics within Commercial is largely decentralised. Collaborating with external partners has many advantages, including the possibility to leverage local market knowledge and health data (see Sect. 6.5 below).

These operational and structural differences between business areas are reflected in their capacities to manage AI-related risks. Hence, different EBA procedures may be needed to assess each business area's governance structure.[30] For example, AstraZeneca's AI audit showed that business areas understood risk differently. Within R&D, many employees work directly with patients and patient data. Hence, they typically see patient-centric risks. As one of our interviewees stated:

> *"Some colleagues have been working with data protection for years. When they hear 'AI ethics', they immediately think of privacy breaches. I often have to remind them that AI ethics is more than just compliance with data protection laws."* (P2)

In contrast, employees working within the Commercial function typically understood risk in financial or contractual terms. Both perspectives are of course valid, and the only purpose of this example is to highlight the difficulty of harmonising a 'risk-based' approach across an organisation that encompasses different understandings of 'risk'.

However, this problem need not be insurmountable. A distinction is often made between *compliance assurance,* which aims at comparing a system to existing laws and regulations, and *risk assurance*, which corresponds to asking open-ended questions about how a system works [112]. Using this distinction, current best practice would demand

[27] Within AstraZeneca a 'high-level' definition of AI exists. However, this definition is flexible enough to allow each business area to further refine the material scope of its AI governance activities.

[28] A parallel can be made to the EU AIA, which takes an explicitly risk-based approach to AI governance [175, 176].

[29] Pragmatic problem-solving demands that things should be sorted so that their grouping will promote successful actions for some specific end [177].

[30] Note that different EBA *procedures* does not imply different *objectives*. In AstraZeneca's case, the control objective of the audit was the same across all business areas whereas the method of verification varied due to the decentralised nature of the organisation.

harmonising EBA procedures that aim to provide compliance assurance across business areas. In contrast, EBA procedures that aim at risk assurance should be adapted locally to reflect how respective business areas understand risk.

6.4 Internal communication as a key to operationalising AI governance

Our observations of AstraZeneca's AI audit suggest that internal communication and training efforts are central to operationalising corporate AI governance. These communication efforts were continuous and happened on several different levels in parallel. For example, the *ethics principles* were agreed upon through a bottom-up process that included extensive consultations with employees and external experts. Importantly, this process was not just about agreeing on a set of principles. It also aimed to anchor the proposed policy with key stakeholders internally. If, for example, managers and software developers do not understand or agree with a policy, they will not prioritise it. However, if they can see how it helps in their daily activities, they will likely adopt it even without top-down directives. As one AstraZeneca employee stated:

> *"Working with the AI Ethics and Governance team was beneficial as it pushed me to think about my project in different ways and gave me new points to consider when developing an AI solution."* (P17)

Moreover, corporate AI governance is about change management. Having formulated the *ethics principles*, AstraZeneca proceeded to the implementation phase. To some extent, that required a time-consuming, top-down roll-out of value statements and compliance documents. This was not a straightforward task: employees have limited attention spans and are frequently bombarded with information about different governance initiatives. It took AstraZeneca over six months to formulate the principles and another year to embed them across the business. Even as the AI audit took place, pockets of the organisation remained unaware of the compliance document.

Previous academic literature has given much attention to (i) the principles that should guide the design and deployment of AI systems [113, 114] and (ii) the tools enabling managers and software developers to translate these principles into practice [55, 115]. While these aspects remain important, our observations suggest that internal communication's role in corporate AI governance deserve more attention. After all, ensuring that AI systems are designed and used legally, ethically, and safely requires organisations to not only have the right values and tools in place but also to make their employees aware of them.

In terms of raising awareness, our findings suggest three best practices. First, communication concerning AI governance is most effective when supported by senior executives.[31] Second, communication efforts around specific EBA procedures work best when stressing how these are relevant to employees' daily tasks. Third, communication around EBA procedures should make explicit why these are needed, thus assuring staff that existing governance procedures are not being duplicated.

6.5 Upholding organisational values in procurement and external collaborations

The full cycle of designing and deploying AI systems seldom takes place within one organisation. Typically, AI systems result from a complex and extended supply chain spanning a plurality of actors and different geographic regions [116]. For example, in 2019, AstraZeneca entered a strategic collaboration with the British start-up BenevolentAI to combine the former's scientific expertise and rich datasets with the latter's biomedical knowledge graph to better understand the mechanisms underlying chronic kidney disease and identify more efficacious treatments [117]. Similarly, in 2021, AstraZeneca launched a collaboration with American healthcare company GRAIL to evaluate the effectiveness of early cancer detection technologies [118].

From a business perspective, external R&D collaborations offer numerous advantages.[32] Yet such collaborations are also coupled with several governance challenges. For example, AstraZeneca's compliance document stipulates that robust, inclusive datasets should be used to train AI systems. During the AI audit, the external auditors explored that by asking how datasets had been collected, cleaned, and processed. However, such EBA procedures are only effective in evaluating AI systems trained in-house. For systems procured from external vendors, neither AstraZeneca nor the independent auditors had full visibility of the internal processes of, or the data used by, suppliers and vendors when training these systems. When discussing the training data for a particular AI system, one participant in an audit meeting stated:

> *"I don't know to be honest. We don't have access to that data. I have tried to get access to the same data but without success. You will have to ask [the external partner]."* (P14)

[31] This finding is supported by previous research. For example, Gasser and Schmitt [178] have shown that the effectiveness of corporate governance mechanism depend on issues related to leadership, values and culture.

[32] External R&D collaborations benefit innovation by increasing efficiency, reducing costs, and granting access to valuable resources not available internally [179].

This has several direct implications for EBA. First, to be effective, the same requirements must apply to all AI systems used by an organisation. Without harmonised requirements, there is a risk that potentially sensitive development projects will only be outsourced to external partners. Second, to be feasible, EBA procedures must encompass a review of corporate procurement processes. However, that may not necessarily require the creation of additional layers of governance. Rather, organisations should undertake a gap finding and filling exercise, adding ethics-based evaluation criteria to existing procurement processes.

6.6 Ethics-based auditing as a catalyst for internal change

There are many reasons why organisations subject themselves to EBA. For example, audits can help to control technology-related risks and inform AI design choices. However, our findings suggest that there are also other motivations for conducting EBA. These include facilitating agenda setting, serving as a catalyst for internal change, and expanding organisational units' mandates.

First, AstraZeneca aims to leverage AI and other data-driven technologies to transform how research is conducted. Digitalisation has thus been put on top of the corporate agenda. Yet as an organisation's technological resources evolve, old governance structures risk becoming ineffective. Hence, AstraZeneca has strong incentives to understand how its internal governance structures need to change to keep up with operational practices.

Second, while organisational change is often incremental, distinct events—such as an audit—can catalyse activities that increase the rate of change. Within AstraZeneca, the upcoming AI audit motivated managers to communicate with their teams about the *ethics principles* and incentivised business areas to develop appropriate governance mechanisms to demonstrate their adherence to those principles. Several interviewees even expressed concerns about how much focus was put on preparing for the audit as a discrete event:

> *"Whenever the upcoming audit took up too much of our internal focus, I felt the need to remind myself and the team that we are not trying to operationalise AI governance because of the audit but to do the right thing."* (P2)

Third, any governance initiative can expand the operational and budgetary mandates of specific organisational units. For example, depending on how AI governance initiatives are framed, they might extend the reach of central functions such as IT or increase the resources allocated to specific CSR initiatives. In AstraZeneca's case, the sustainability team drove the formulation of the *ethics principles*. During the roll-out, a more decentralised structure emerged, with each business area responsible for practically implementing the principles.

The point we seek to stress here is that identifying or mitigating harm resulting from AI failures is not the only reason to implement EBA. EBA procedures can—and often do—serve other important functions, e.g., catalysing organisational change.

6.7 Making verifiable claims on the basis of ethics-based audits

The subject matter of AI audits can be a person, an organisation, a process, a system, or any combination thereof. The AI audit conducted within AstraZeneca took a process approach in which the assessment was based on management representation, e.g., through interviews with key decision-makers and a review of sample documentation. In line with this approach, no detailed reviews of source codes, data sets, or model outputs were performed. Some interviewees expressed surprise regarding this:

> *"We are only talking about basic assumptions and the completeness of our documentation. I don't see what this has to do with AI?"* (P14)

Despite some individuals' misgivings, the procedure followed during AstraZeneca's AI audit is well-supported by previous research. While AI systems may appear opaque, technologies can always be understood in terms of their designs and intended operational goals [119]. Similarly, third-party auditors can make verifiable claims about AI systems without accessing the underlying data and computational models by analysing publicly available information [120].

In fact, EBA procedures that focus on organisational processes have several advantages. They are less demanding than code audits in terms of access to proprietary data. Since proprietary protection is one of the main drivers of AI systems' opacity [121, 122], that facilitates the process of conducting AI audits. Moreover, EBA procedures focussing on organisational processes are explicitly forward-looking. Rather than conducting post hoc evaluations, the auditor and the technology provider collaborate to assess and improve the processes that shape future AI systems' properties and safeguards. This helps distinguishing accountability from blame [123].[33]

Nevertheless, it is important to remain realistic about what EBA procedures focussing on organisational processes

[33] According to Diakopoulos [180], what is needed to operationalise AI governance in an organisation is a map that models the assignment of responsibility based on the ethical expectations of different actors.

can be expected to achieve. Such procedures can verify claims about technology providers' quality management systems but are fundamentally unable to produce verifiable claims about the impacts that autonomous, self-learning AI systems that co-evolve with complex environments may have over time.

6.8 Measuring progress and demonstrating success

Social phenomena are increasingly measured, described, and influenced by numbers,[34] and the corporate governance field is no exception. Sine Taylor, management scholars have refined metrics to measure and control workers' productivity as well as the societal impact and environmental footprint of products and services [124, 125].[35] Such metrics are relevant for EBA for two reasons. First, organisations investing in corporate AI governance want to demonstrate success by pointing towards tangible improvements. Second, ethical decision-making requires a frame of reference, i.e., a baseline against which normative judgements can be made. EBA producers should, therefore, include metrics that quantify the behaviour of technology providers and the AI systems they design and deploy.

Recently, much literature has focussed on measuring and assessing the performance of different AI systems along normative dimensions such as fairness, transparency, and accountability [126]. For example, Wachter et al. [127] compiled a list containing over 20 different fairness metrics, accompanied by a guide for choosing the most appropriate one for different use cases. These metrics can, in turn, be leveraged by conceptual tools or software that measure, evaluate, or visualise one or more properties of AI systems during EBA (see e.g., [128, 129]).

However, the use of metrics during AI audits is not unproblematic. Goodheart's Law reminds us that when a measure becomes a target, it ceases to be a good metric [130]. Moreover, as Lee et al. [131] argue, reductionist representations of normative values (like fairness) often bear little resemblance to how these notions are experienced in real-life. In practice, different principles often conflict and require trade-offs [52]. Similarly, different definitions of fairness—like individual fairness and demographic parity—are mutually exclusive [132, 133].

How suitable different metrics are for specific EBA procedures depends on the nature of the audit. For AstraZeneca's process audit, the metrics employed aimed at capturing the extent to which best practices within software development were followed and appropriate safeguards were in place. One way to do so would have been to record 'Yes'/'No' answers to simple checklists. Such an approach has some support; by formalising ad-hoc processes and empowering individual advocates, checklists help organisations identify risks and tensions [134]. Yet simply having a checklist is insufficient to ensure that AI systems are designed and used legally, ethically, and safely [135] and previous research has found that checklists risk reducing auditing to a box-ticking exercise [89].

Rather than using binary checklists, the auditors in AstraZeneca's case made use of open-ended questions that allowed managers and developers to articulate how (and why) specific AI systems were built.[36] Indeed, the most fruitful moments happened when AstraZeneca's in-house experts and the external auditors jointly discussed the merits of different ways of measuring the properties of specific AI models—thereby challenging the assumptions that underpin concepts like fairness or transparency. For example, AstraZeneca staff were asked to consider questions like: do we have rules about when and how we use AI systems? and what evidence can we use to determine whether an AI systems we design is 'fair' or 'robust'? As one of the external auditors put it:

> *"The really rich information comes not from asking a pre-curated list of questions, but from listening to the answers and asking relevant follow-up questions."* (P9)

Taken together, our observations before, during, and after AstraZeneca's AI audit suggest that the primary purpose of metrics in the process of operationalising AI governance is not to decide whether a specific system is 'ethical' or not, but rather to spark ethical deliberation, inform design choices, and help visualise the normative values embedded in that system. This observation is compatible with the claim that multi-dimensional Pareto frontiers can be used to strike publicly justifiable tradeoffs between competing criteria [136]. Thus, a fruitful avenue for future research would be to develop a guide on when and how to use different metrics not only in the software development lifecycle but also as part of holistic EBA procedures.

6.9 The costs associated with ethics-based audits

Efforts to operationalise AI governance inevitably incur both financial and administrative costs. In the case of EBA, that

[34] See Mau [181] for an excellent account of the growing tendency to quantify the social world and how that process changes our assignment of worth.

[35] Note that organisational performance metrics need not be based on financial measures alone. The perhaps most famous example of this is 'the balanced scorecard' [182].

[36] Here, a parallel can be made to 'Ethical Foresight Analysis', a method based on Failure Modes and Effects Analysis (FMEA), which is standard practice in safety engineering [183].

includes *initial costs* (e.g., time and resources invested in preparing for the audit as well as the procurement of audit services, software systems, and test data) and *variable costs* (e.g., the costs of implementing and adhering to an audit's recommendations, such as additional steps in the development process or continuous human oversight).[37]

To start with, formulating organisational values bottom-up is a time-consuming activity. In AstraZeneca's case, the process of drafting and agreeing on the *ethics principles* included multiple consultations with executive leaders on strategy, with senior developers to understand AI-related risks, with heads of different business areas to compare the agreed-upon principles with existing codes of conduct, as well as with academic researchers and industry experts to receive external feedback. Subsequently, the *ethics principles* had to be communicated, anchored, and implemented across the organisation (another labour-intensive activity). Beyond the time invested by senior leaders and individual employees, approximately four full-time staff worked on driving and coordinating the implementation of AI governance within AstraZeneca during 2020 and 2021.

During the actual audit in Q4 2021, the demands on manual resources increased. A team of auditors were contracted to evaluate AstraZeneca's overarching AI governance structure and conduct in-depth reviews of selected AI development projects and use cases. The AI audit took 14 weeks to conduct. Throughout, AstraZeneca employees allocated time to provide the auditors with relevant documents and answer detailed questions during interviews. Taken together, around 2000 person-hours were invested in the audit, even though it was relatively light-touch and did not involve any technical tests of individual AI models.

These numbers only give a ballpark indication of the costs associated with EBA. Indeed, quantifying the costs associated with any governance mechanism is difficult. Take the ongoing debate concerning the costs of complying with the EU AIA as an example. According to the European Commission, obtaining certification for an AI system in line with the AIA will cost on average EUR 16,800–23,000, corresponding to approximately 10–14% of the development cost [137]. While those numbers have been supported by independent researchers [138], the critics claim that the official estimates are too low and fail to incorporate the long-term effects of the legislation such as reduced investments in AI research [139].

The discussion around the cost of complying with the EU AIA illustrates that a governance mechanism's financial viability does not hinge on its direct costs alone but also on long-term opportunity costs and transformative effects. After all, one of the main reasons why technology providers engage with auditors is that it is cheaper and easier to address system vulnerabilities early in the development process. For example, it can costs up to 15 times more to fix a software bug found during the testing phase than fixing the same bug found in the design phase [140]. This suggests that—despite the associated costs—businesses have clear incentives to design and implement effective EBA procedures.

7 Limitations

Conducting qualitative research is challenging and bound to result in methodological shortcomings [141]. Here, we discuss important limitations with regards to the validity, independence, and generalisability of our findings.

Consider validity first. Since our study relied on descriptive methods, it is most relevant to consider construct validity, i.e., the ability to link research observations to their intended theoretical constructs [142]. For example, it is difficult to assess the ethical risks posed by specific AI systems. Therefore, we exclusively focussed on *observing and describing* the challenges organisations face when implementing EBA procedures, rather than *identifying or measuring* the effects such procedures have on the behaviour of AI systems.

A further risk related to validity concerns the possibility of replicating findings from previous research due to confirmation bias [143]. While difficult to eliminate, this risk was managed through an iterative process, with findings from the literature and the case study continuously informing each other. In fact, including longitudinal case studies helps strengthen the validity of nonexperimental research designs [144].

Another limitation concerns the independence of the research. As mentioned, JM's research is funded through an Oxford-AstraZeneca studentship. When such dependencies exist, researchers may feel pressured to produce 'positive' results, i.e., findings that the industry partner wants to hear [145]. To manage this risk, we communicated clear boundaries regarding our roles as independent researchers. We also followed best practices in research ethics, e.g., informing all parties about the constructively critical nature of our work.

A final set of limitations concerns the generalisability of the case study's findings. Inevitably, the input provided by the industry partner can be biased or contextually limited [146]. Moreover, data controllers (like AstraZeneca) have an interest in not disclosing trade secrets [147]. We sought to reduce the risk that biased input distorts the analysis by triangulating the information provided by AstraZeneca

[37] This is nothing new. Already in 1980, Weiss [184] published an article titled *Auditability of Software: A Survey of Techniques and Costs*.

employees with other sources. Still, the findings from the case study should not be treated as neutral, but rather as context-specific knowledge [148].

These limitations do not mean that the findings cannot be generalised. Indeed, AstraZeneca's efforts to operationalise AI governance are highly representative of the many large firms that have recently adopted ethics principles for designing and deploying AI systems. Notable examples include Google, Microsoft, IBM and BMW [149–152]. In short, many large organisations will face similar challenges when developing and implementing EBA procedures. The findings presented in Sect. 6 will thus be relevant to other large corporations attempting to integrate EBA procedures with existing governance structures.

8 Conclusions

A new industry that focuses on auditing AI systems is emerging. The proposed European legislation on AI, which sketches the contours of a professionalised AI auditing ecosystem [77], is likely to accelerate this trend. In such a fast-moving and high-stakes environment, it is of increasing importance for both regulators and business executives to understand the conditions under which EBA is a feasible and effective mechanism for operationalising AI governance. The findings from the AstraZeneca case study helps further such an understanding.

Different EBA procedures serve different purposes. Process audits—such as that undertaken by AstraZeneca—are well suited to verifying claims about the quality management systems a particular technology provider has in place as well as to identifying, assessing, and mitigating risks throughout the AI life cycle. Compared to code audits, they are also less demanding in terms of access to proprietary information and sensitive data. However, it is important to remain realistic about process audits' capabilities and limitations. EBA procedures that do not include any technical elements are fundamentally unable to produce verifiable claims about the effects autonomous and self-learning AI systems may have over time.

In terms of implementation, our observations suggest that EBA procedures are most likely to be effective when integrated into existing governance structures. That is because EBA procedures that duplicate existing structures (or operate in silos) may be perceived as unnecessary by the managers and developers expected to implement them. Similarly, efforts to operationalise AI governance through EBA are most effective when internal communication is centred around how this would help employees with their daily tasks. In contrast, EBA procedures that are perceived as filling only abstract functions are easily reduced to box-ticking exercises—and thereby fail to positively influence the design and deployment of AI systems. Best practice thus demands that AI auditors—whether internal or external—collaborate with managers and software developers to counteract problems related to unethical uses of, and unforeseen risks posed by, AI systems.

Large multinational organisations attempting to operationalise AI governance through EBA will inevitably face at least three critical challenges. First, EBA's feasibility as a governance mechanism is undermined by the difficulty of harmonising standards across decentralised organisations. AI audits require a pre-defined baseline against which organisational units, processes, or systems can be evaluated. However, like AstraZeneca, large organisations often comprise distinct business areas operating independently. Mandating uniform AI governance structures top-down thus poses challenges to the entire way such organisations are structured and run.

Second, the lack of a well-defined material scope for AI governance constitutes an obstacle to EBA. In short, questions as to which systems and processes AI governance frameworks ought to apply to remain unanswered. AstraZeneca's difficulties when attempting to establish a material scope for their AI audit highlight the three-way trade-off between how precise a scope is, how easy it is to apply, and how generalisable it is. Nevertheless, pragmatic problem-solving demands that things should be sorted so that their grouping will promote successful actions for some specific end. As a result, it will remain difficult for any EBA procedure to produce verifiable claims until the material scope of AI governance is accepted throughout an organisation.

Third, unresolved tensions related to procurement and external R&D collaborations risk undermining AI audits' effectiveness. For example, to successfully operationalise AI governance, EBA procedures must treat AI systems developed in-house and those procured from third-party vendors equally. If not, new internal governance structures may cause unethical (or simply 'risky') development projects to be outsourced. This is akin to what Floridi [71] has labelled 'ethics dumping', i.e., the malpractice of exporting unethical activities to countries (or organisations) where there are weaker legal and ethical frameworks and enforcement mechanisms. The solution here would be for organisations to include alignment with internal AI governance policies as a criterion in future procurement processes and contractual agreements with external R&D collaborators.

While the conclusions offered above may not be surprising, they nonetheless stand in contrast to what has hitherto been the focus of academic research in this field. Simplified, previous research on EBA fall into one of two categories. The first consists of works that draw on law and political or moral philosophy to justify why EBA is needed. The second consists of works that draw on computer science or systems engineering to specify how EBA ought to be

conducted. However, both the best practices and the challenges highlighted in this article indicate that the main difficulties organisations face when conducting AI audits mirror classical governance challenges. This indicates that not only computer scientists, engineers, philosophers, and lawyers but also management scholars need to be involved in the research on how to design EBA procedures.[38]

Acknowledgements The authors wish to thank Margi Sheth, Olawale Alimi, Mimmi Gersbro-Sundler, Peder Blomgren, Matthias Holweg, and Josh Cowls for helpful comments on earlier versions of this article.

Funding JM's doctoral research at the Oxford Internet Institute is supported through a studentship provided by AstraZeneca.

Declarations

Conflict of interest JM's doctoral research at the Oxford Internet Institute is supported through a studentship provided by AstraZeneca. The studentship is administered and paid out by the University and there have been no direct financial transactions between AstraZeneca and JM. The research was conducted with the approval of AstraZeneca. However, the research was academically independent, and all opinions expressed in the article belongs solely to its authors.

Disclaimer We hereby declare that this article is our original work and has not been submitted to any other journal for publication. Further, we have acknowledged all sources used and cited these in the reference section.

CRediT authorship statement JM: conceptualization, investigation, data curation, formal analysis, writing—original draft, project administration. LF: validation, writing—review and editing, supervision.

References

1. Brown, S., Davidovic, J., Hasan, A.: The algorithm audit: Scoring the algorithms that score us. Big Data Soc. **8**(1), 205395172098386 (2021)
2. Brundage, M., et al.: Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv:2004.07213[cs.CY] (2020)
3. Koshiyama, A., et al.: Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. SSRN Electron. J. 1–31 (2021)
4. LaBrie, R.C and Steinke, G. H.: Towards a framework for ethical audits of AI algorithms. In: 25th Am. Conf. Inf. Syst. AMCIS 2019, pp. 1–5 (2019)
5. Mökander, J., Floridi, L.: Ethics-based auditing to develop trustworthy AI. Minds Mach **31**, 323–327 (2021). https://doi.org/10.1007/s11023-021-09557-8
6. Raji, I.D., and Buolamwini, J.: Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: AIES 2019 - Proc. 2019 AAAI/ACM Conf. AI, Ethics, Soc., pp. 429–435 (2019)
7. Floridi, L.: Infraethics–on the conditions of possibility of morality. Philos. Technol. **30**(4), 391–394 (2017). https://doi.org/10.1007/s13347-017-0291-1
8. Larsson, S.: On the governance of artificial intelligence through ethics guidelines. Asian J. Law Soc. **7**(3), 437–451 (2020)
9. Kazim, E. and Koshiyama, A.: A high-level overview of AI ethics. SSRN Electron. J., no. Lukowicz, pp. 1–18 (2020)
10. Leveson, N.: Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press, Cambridge (2011)
11. Sandvig, C., Hamilton, K., Karahalios, K. and Langbort, C.: Auditing algorithms. In ICA 2014 Data Discrim. Preconference, pp. 1–23 (2014)
12. Diakopoulos, N.: Algorithmic accountability: journalistic investigation of computational power structures. Digit. Journal. **3**(3), 398–415 (2015)
13. Cobbe, J., Lee, M. S. A. and Singh, J.: Reviewable automated decision-making: a framework for accountable algorithmic systems. In FAccT 2021 - Proc. 2021 ACM Conf. Fairness, Accountability, Transpar., pp. 598–609 (2021)
14. ForHumanity: Independent Audit of AI Systems (2021). https://forhumanity.center/independent-audit-of-ai-systems. (**Accessed: 17-Feb-2021**)
15. Zicari, R.V., et al.: Z-Inspection®: a process to assess trustworthy AI. IEEE Trans. Technol. Soc. **2**(2), 83–97 (2021)
16. Kazim, E., and Koshiyama, A.: AI Assurance Processes. SSRN Electron. J., no. September, pp. 1–9, (2020)
17. Keyes, O., Hutson, J. and Durbin, M.: A mulching proposal. no. May 2019, pp. 1–11 (2019)
18. ICO: Guidance on the AI auditing framework: draft guidance for consultation. Inf. Comm. Off. (2020)
19. Floridi, L., Holweg, M., Taddeo, M., Silva, J.A., Mökander, J., Wen, Y.: capAI - A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act (March 23, 2022). Available at SSRN: https://ssrn.com/abstract=4064091 or https://doi.org/10.2139/ssrn.4064091
20. PwC: A practical guide to Responsible Artificial Intelligence (AI) (2019)
21. EY: Assurance in the age of AI Executive summary (2018)
22. Deloitte: Deloitte introduces trustworthy AI framework to guide organizations in ethical application of technology. Press release (2020). https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-introduces-trustworthy-ai-framework.html. (**Accessed: 19-Sep-2020**)
23. KPMG: KPMG offers ethical AI Assurance using CIO Strategy Council standards. Press release (2020) https://home.kpmg/ca/en/home/media/press-releases/2020/11/kpmg-offers-ethical-ai-assurance-using-ciosc-standards.html. (**Accessed: 11-Nov-2021**)
24. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. Conf. Fairness Accountabil. Transparency **1**, 1–15 (2018)
25. Mahajan, V., Venugopal, V.K., Murugavel, M., Mahajan, H.: The algorithmic audit: working with vendors to validate radiology-AI algorithms—how we do it. Acad. Radiol. **27**(1), 132–135 (2020)

[38] This conclusion reiterates findings from previous research. See e.g., Raisch and Krakowski [185].

26. Kazim, E., Koshiyama, A.S., Hilliard, A., Polle, R.: Systematizing audit in algorithmic recruitment. J. Intell. **9**(3), 1–11 (2021)
27. Dignum, V.: Responsibility and artificial intelligence. Oxford Handb. Ethics AI, no. November, pp. 213–231 (2020)
28. Balas, V.E., Kumar, R., Srivastava, R.: Recent Trends and Advances in Artificial Intelligence and Internet of Things. Springer, Cham (2020)
29. Taddeo, M., Floridi, L.: How AI can be a force for good. Science **361**(6404), 751–752 (2018). https://doi.org/10.1126/science.aat5991
30. Grote, T., Berens, P.: On the ethics of algorithmic decision-making in healthcare. J. Med. Ethics **46**(3), 205–211 (2020)
31. Begoli, E., Bhattacharya, T., Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making. Nat. Mach. Intell. **1**(1), 20–23 (2019)
32. Kaushik, S., et al.: AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. Front. Big Data (2020). https://doi.org/10.3389/fdata.2020.00004
33. Schneider, G.: Mind and machine in drug design. Nat. Mach. Intell. **1**(3), 128–130 (2019)
34. Jiang, F., et al.: Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. **2**(4), 230–243 (2017)
35. Morley, J., et al.: The ethics of AI in health care: a mapping review. Soc. Sci. Med. **260**, 113172 (2020)
36. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. **25**(1), 44–56 (2019)
37. Leslie, D.: Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. SSRN J. (2019). https://doi.org/10.2139/ssrn.3403301
38. Tsamados, A., et al.: The ethics of algorithms: key problems and solutions. SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3662302
39. McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., Buyx, A.: Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med. Ethics **23**(1), 1–10 (2022)
40. Laurie, G., Stevens, L., Jones, K.H. and Dobbs, C.: A review of evidence relating to harm resulting from use of health and biomedical data BT—Nuffield Council on Bioethics (2014)
41. Sjoding, M.W., Dickson, R.P., Iwashyna, T.J., Gay, S.E., Valley, T.S.: Racial bias in pulse oximetry measurement. N. Engl. J. Med. **383**(25), 2477–2478 (2020)
42. Taeihagh, A., Ramesh, M., Howlett, M.: Assessing the regulatory challenges of emerging disruptive technologies. Regul. Gov. **15**(4), 1009–1019 (2021)
43. Cookson, C.: Artificial intelligence faces public backlash, warns scientist. Fin. Times (2018)
44. Blasimme, A., Vayena, E.: The ethics of AI in biomedical research, patient care, and public health. In: Dubber, M., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI. Oxford University Press, Oxford (2021)
45. van de Poel, I.: Embedding values in artificial intelligence (AI) systems. Minds Mach. **30**(3), 385–409 (2020)
46. Di Maio, P.: Towards a metamodel to support the joint optimization of socio technical systems. Systems **2**(3), 273–296 (2014)
47. Lauer, D.: You cannot have AI ethics without ethics. AI Ethics **0123456789**, 1–5 (2020)
48. Schneider, J., Abraham, R., and Meske, C.: AI governance for businesses. no. November, 2020
49. Fjeld, J.: Principled Artificial intelligence. IEEE Instrum. Meas. Mag. **23**(3), 27–31 (2020)
50. Jobin, A., Ienca, M., Vayena, E.: Artificial Intelligence: the global landscape of ethics guidelines. Nat. Mach. Intell. **1**(9), 389 (2019)
51. Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. Harvard Data Sci. Rev. **1**, 1–13 (2019). https://doi.org/10.1162/99608f92.8cd550d1
52. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019)
53. Whittlestone, J., Alexandrova, A., Nyrup, R. and Cave, S.: The role and limits of principles in AI ethics: towards a focus on tensions. In: AIES 2019 - Proc. 2019 AAAI/ACM Conf. AI, Ethics, Soc., pp. 195–200 (2019)
54. Vakkuri, V., Kemell, K. K., Kultanen, J., Siponen, M. and Abrahamsson, P.: Ethically aligned design of autonomous systems: industry viewpoint and an empirical study. arXiv, (2019)
55. Ayling, J. and Chapman, A.: Putting AI ethics to work: are the tools fit for purpose?. AI Ethics 0123456789 (2021)
56. Morley, J., et al.: Operationalising AI ethics: barriers, enablers and next steps. AI Soc. (2021). https://doi.org/10.1007/s00146-021-01308-8
57. Morley, J., Elhalal, A., Garcia, F., et al.: Ethics as a service: a pragmatic operationalisation of AI ethics. Minds Mach. **31**, 239–256 (2021). https://doi.org/10.1007/s11023-021-09563-w
58. AI HLEG: Assessment List for Trustworthy AI (ALTAI) (2020)
59. Koshiyama, A.: Algorithmic impact assessment: fairness, robustness and explainability in automated decision-making (2019)
60. Reisman, D., Schultz, J., Crawford, K. and Whittaker, M.: Algorithmic impact assessments: a practical framework for public agency accountability. AI Now Inst. 22 (2018)
61. Mitchell, M., et al.: Model cards for model reporting. In: FAT* 2019—Proc. 2019 Conf. Fairness, Accountability, Transpar., no. Figure 2, pp. 220–229 (2019)
62. Gebru, T., et al.: Datasheets for datasets (2018)
63. Holland, S., Hosny, A., Newman, S., Joseph, J. and Chmielinski, K.: The dataset nutrition label: a framework to drive higher data quality standards (2018)
64. Jotterand, F., Bosco, C.: Keeping the 'Human in the Loop' in the age of artificial intelligence: accompanying commentary for 'correcting the brain?' by Rainey and Erden. Sci. Eng. Ethics **26**(5), 2455–2460 (2020)
65. Cihon, P.: Standards for AI Governance: international standards to enable global coordination in AI research and development. In: Futur. Humanit. Institute, Univ. Oxford, no. April, pp. 1–41 (2019)
66. Cruz Rivera, S., et al.: Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat. Med. **26**(9), 1351–1363 (2020)
67. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M.J., Denniston, A.K.: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat. Med. **26**(9), 1364–1374 (2020)
68. Prunkl, C.E.A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., Dafoe, A.: Institutionalizing ethics in AI through broader impact requirements. Nat. Mach. Intell. **3**(2), 104–110 (2021)
69. Cihon, P., Schuett, J., Baum, S.D.: Corporate governance of artificial intelligence in the public interest. Information **12**(7), 1–30 (2021)
70. Minkkinen, M., Zimmer, M.P., Mäntymäki, M.: Towards ecosystems for responsible AI: expectations, agendas and networks in EU documents. Springer International Publishing, Berlin (2021)
71. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**(2), 185–193 (2019). https://doi.org/10.1007/s13347-019-00354-x
72. EIU. Staying ahead of the curve—The business case for responsible AI. 2020. https://www.eiu.com/n/staying-ahead-of-the-curve-the-business-case-for-responsible-ai/. (**Accessed: 08-Oct-2020**)
73. Holweg, M., Younger, R. and Wen, Y.: The reputational risks of AI

74. Floridi, L.: The end of an era: from self-regulation to hard law for the digital industry. Philos. Technol. **34**(4), 619–622 (2021). https://doi.org/10.1007/s13347-021-00493-0
75. European Commission: Proposal for 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 (2021)
76. Office of U.S. Senator Ron Wyden: Algorithmic accountability act of 2022. In: 117th Congr. 2D Sess., (2022)
77. Mökander, J., Axente, M., Casolari, F., et al.: Conformity assessments and post-market monitoring: a guide to the role ofauditing in the proposed european AI regulation. Minds Mach. (2021). https://doi.org/10.1007/s11023-021-09577-4
78. Floridi, L.: Soft ethics and the governance of the digital. Philos. Technol. **31**(1), (2018). https://doi.org/10.1007/s13347-018-0303-9
79. AstraZeneca: AstraZeneca annual report & form 20-F information 2020. Issues Sci. Technol. **25**(4), 23–30 (2020)
80. Langkafel, P., (ed.) Big Data in Medical Science and Healthcare Management (2015)
81. Ashenden, S. K., Deswal, S., Bulusu, K. C., Bartosik, A. and Shameer, K.: Data types and resources. In: Era Artif. Intell. Mach. Learn. Data Sci. Pharm. Ind., pp. 27–60 (2021)
82. Crowe, D.: Modelling biomedical data for a drug discovery knowledge graph. Towards Data Science (2020). https://towardsdatascience.com/modelling-biomedical-data-for-a-drug-discovery-knowledge-graph-a709be653168. (**Accessed: 22-Nov-2021**)
83. Vasetenkov, A.: AstraZeneca's knowledge graph: drug discovery is a lot about connections. Eckher Insights (2021). https://www.eckher.com/c/21h530pr6z. (**Accessed: 22-Nov-2021**)
84. AstraZeneca: Data science and artificial intelligence: unlocking new science insights. (2021). https://www.astrazeneca.com/r-d/data-science-and-ai.html#UsingAI
85. Lea, H., et al.: Can machine learning augment clinician adjudication of events in cardiovascular trials? A case study of major adverse cardiovascular events (MACE) across CVRM trials. Eur. Heart J. (2021). https://doi.org/10.1093/eurheartj/ehab724.3061
86. Rizk, J.G., Barr, C.E., Rizk, Y., Lewin, J.C.: The next frontier in vaccine safety and VAERS: lessons from COVID-19 and ten recommendations for action. Vaccine **39**(41), 6017 (2021)
87. AstraZeneca: AstraZeneca data and AI ethics. In: Position statement (2020). https://www.astrazeneca.com/sustainability/ethics-and-transparency/data-and-ai-ethics.html. (**Accessed: 09-Mar-2021**)
88. Mantelero, A.: AI and big data: a blueprint for a human rights, social and ethical impact assessment. Comput. Law Secur. Rev. **34**(4), 754–772 (2018)
89. Raji, I. D., et al.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: FAT* 2020 - Proc. 2020 Conf. Fairness, Accountability, Transpar., pp. 33–44 (2020)
90. Bauer, J.: The necessity of auditing artificial intelligence. SSRN J. **577**, 1–16 (2016)
91. Larsson, S., Heintz, F.: Transparency in artificial intelligence. Internet Policy Rev. **9**(2), 1–16 (2020)
92. Mökander, J., Axente, M.: Ethics-based auditing of automated decision-making systems: intervention points and policyimplications. AI & Soc. (2021). https://doi.org/10.1007/s00146-021-01286-x
93. Mittelstadt, B.: Auditing for transparency in content personalization systems. Int. J. Commun. **10**(June), 4991–5002 (2016)
94. Kroll, J. A., et al.: Accountable algorithms. Univ. PA. Law Rev. (633), 66 (2016)
95. Bass, J. M., Lero, S. B. and Noll, J.: Experience of industry case studies: a comparison of multi-case and embedded case study methods. In: Proc. - Int. Conf. Softw. Eng., pp. 13–20 (2018)
96. Yin, R.K.: Case study research: design and methods, 2nd edn. Sage, Thousand Oaks (1994)
97. Thomson, R., Plumridge, L., Holland, J.: Longitudinal qualitative research: a developing methodology. Int. J. Soc. Res. Methodol. Theory Pract. **6**(3), 185–187 (2003)
98. Merton, R.K.: Three fragments from a sociologist's notebooks: establishing the phenomenon, specified ignorance, and strategic research materials. Rev. Lit. Arts Am. **13**(1987), 1–28 (1987)
99. Vinten, G.: Participant observation: a model for organizational investigation? J. Manag. Psychol. **9**(2), 30–38 (1994)
100. Woodside, A.G.: "Participant Observation Research in Organizational Behavior", in Case Study Research, pp. 331–352. Emerald Group Publishing Limited, Boston (2016)
101. Edwards, R., Holland, J.: What is Qualitative Interviewing? Bloomsbury Academic, London (2013)
102. Given, L.M.: The SAGE Encyclopedia of Qualitative Research Methods. SAGE, Los Angeles (2008)
103. Creswell, J., Clark, V.: Designing and Conducting Mixed Methods Research, 3rd edn. SAGE Publications, Berlin (2011)
104. Frey, B.: Document analysis. In: Frey, B.B. (ed.) The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation. SAGE Publications, Inc., Thousand Oaks (2018)
105. Bryman, A.: Social research methods, 5th edn. Oxford (2016)
106. Nadler, E., et al.: Treatment patterns and clinical outcomes in patients with advanced non-small cell lung cancer initiating first-line treatment in the US community oncology setting: a real-world retrospective observational study. J. Cancer Res. Clin. Oncol. **147**(3), 671–690 (2021)
107. Chiou, J., Magazzini, L., Pammolli, F. and Riccaboni, M.: The value of failure in pharmaceutical R & D the value of failure in pharmaceutical R & D. (1), 1–22 (2012)
108. Wang, P.: On defining artificial intelligence. J. Artif. Gen. Intell. **10**(2), 1–37 (2019)
109. Schuett, J. Defining the scope of AI regulations. arXiv:1909.01095 (2019). Accessed 22 Aug 2021
110. Baum, S. D.: Social choice ethics in artificial intelligence. AI Soc., pp. 1–12 (2017)
111. Danks, D. and London, A. J.: Algorithmic bias in autonomous systems. In: IJCAI Int. Jt. Conf. Artif. Intell., vol. 0, no. January, pp. 4691–4697 (2017)
112. CDEI. AI assurance (2021)
113. Alshammari, M. and Simpson, A.: Towards a principled approach for engineering privacy by design. In: Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10518 LNCS, pp. 161–177 (2017)
114. Fjeld, J.: Principled artificial intelligence. IEEE Instrum. Measur. Mag. **23**(3), 27–31 (2020). https://doi.org/10.1109/MIM.2020.9082795
115. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: 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 (2020). https://doi.org/10.1007/s11948-019-00165-5
116. Crawford, K.: The atlas of AI. Yale University Press, New Haven (2021)
117. BenevolentAI: AstraZeneca starts artificial intelligence collaboration to accelerate drug discovery | BenevolentAI. Press release (2019). https://www.benevolent.com/news/astrazeneca-starts-artificial-intelligence-collaboration-to-accelerate-drug-discovery. (**Accessed: 05-Jan-2022**)
118. GRAIL: GRAIL Announces Collaborations with Amgen, AstraZeneca, and Bristol Myers Squibb to Evaluate Cancer Early Detection Technology for Minimal Residual Disease – GRAIL. Press release (2021). https://grail.com/press-releases/grail-announces-collaborations-with-amgen-astrazeneca-and-brist

ol-myers-squibb-to-evaluate-cancer-early-detection-technology-for-minimal-residual-disease/. (**Accessed: 05-Jan-2022**)
119. Kroll, J.A.: The fallacy of inscrutability. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. (2018). https://doi.org/10.1098/rsta.2018.0084
120. Dash, A., Mukherjee, A. and Ghosh, S.: A network-centric framework for auditing recommendation systems. In: Proc. - IEEE INFOCOM, vol. April, pp. 1990–1998 (2019)
121. Burrell, J.: How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. (2016). https://doi.org/10.1177/2053951715622512
122. Pasquale, F.: The Black Box Society: the secret algorithms that control money and information. Inf. Commun. Soc. **19**(12), 1727–1728 (2016)
123. Chopra, A.K. and Singh, M. P.: Sociotechnical systems and ethics in the large. In: AIES 2018 - Proc. 2018 AAAI/ACM Conf. AI, Ethics, Soc., pp. 48–53 (2018)
124. Islam, G., and Greenwood, M.: The metrics of ethics and the ethics of metrics. J. Bus. Ethics 0123456789 (2021)
125. Cugueró-Escofet, N., Rosanas, J.M.: The ethics of metrics: overcoming the dysfunctional effects of performance measurements through justice. J. Bus. Ethics **140**(4), 615–631 (2017)
126. Hoffmann, A.L., Roberts, S.T., Wolf, C.T., Wood, S.: Beyond fairness, accountability, and transparency in the ethics of algorithms: contributions and perspectives from LIS. Proc. Assoc. Inf. Sci. Technol. **55**(1), 694–696 (2018)
127. Wachter, S., Mittelstadt, B., Russell, C.: Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. SSRN Electron. J. (2021). https://doi.org/10.2139/ssrn.3792772
128. Bellamy, R.K.E., et al.: AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. **63**(4–5), 4:1-4:15 (2019)
129. Cabrera, Á.A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J. and Chau, D. H.: FairVis: visual analytics for discovering intersectional bias in machine learning (2019)
130. Greenfield, A.: Radical Technologies : the Design of Everyday Life. London ; New York (2017)
131. Lee, M.S.A., Floridi, L., Singh, J.: Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI Ethics **1**, 529–544 (2021). https://doi.org/10.1007/s43681-021-00067-y
132. Kusner, M., Loftus, J., Russell, C. and Silva, R.: Counterfactual fairness. Adv. Neural Inf. Process. Syst. 4067–4077 (2017)
133. Verma, S., and Rubin, J.: Fairness definitions explained. In: Proc. - Int. Conf. Softw. Eng., pp. 1–7 (2018)
134. Madaio, M.A., Stark, L., Wortman Vaughan, J., and Wallach, H.: Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. *Chi 2020* (2020)
135. McNamara, A., Smith, J. and Murphy-Hill, E.: Does ACM's code of ethics change ethical decision making in software development?. In: ESEC/FSE 2018 - Proc. 2018 26th ACM Jt. Meet. Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng., no. March, pp. 729–733 (2018)
136. Kearns, M.J. and Roth, A.: The Ethical Algorithm : the Science of Socially Aware Algorithm Design. New York (2020)
137. Renda, A., Arroyo, J., Fanni, R., Laurer, M., Maridis, G. and Devenyi, V.: Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe (2021)
138. Haataja, M. and Bryson, J. J.: What costs should we expect from the EU's AI Act?. pp. 1–6 (2021)
139. Mueller, B.: How much will the artificial intelligence act cost Europe? (2021)
140. Dawson, M., Burrell, D.N., Rahim, E., Brewster, S.: Integrating software assurance into the software development life cycle (sdlc) meeting department of defense (dod) demands. J. Inf. Syst. Technol. Plan. **3**(6), 49–53 (2010)
141. Miles, M.B., Huberman, A.M.: Qualitative Data Analysis: An Expanded Sourcebook. Sage, Thousand Oaks (1994)
142. Smith, E.: Research design. In: Reis, H. and Judd, C. (eds) Handbook of Research Methods in Social and Personality Psychology, pp. 27–48 (2014)
143. Pub, F. M. W., et al.: Meta-analysis and synthesizing research (2019)
144. Levendusky, M.: Partisan media exposure and attitudes toward the opposition. Polit. Commun. **30**(4), 565–581 (2013)
145. Maruyama, G., Ryan, C.S.: Research methods in social relations. West Sussex, Chichester (2014)
146. Morgan, C.D.L., Krueger, R.A., and Morgan, E.D.L.: Successful Focus Groups: Advancing the State of the Art When to Use Focus Groups and Why," pp. 3–20 (2016)
147. Flyvbjerg, B.: Making Social Science Matter. Cambridge University Press, Cambridge (2001)
148. Jackall, R.: Moral Mazes: the World of Corporate Managers, 20th anniv. Oxford University Press, New York (2010)
149. Google: Artificial Intelligence at Google: Our principles. Communication (2018). https://ai.google/principles/. (**Accessed: 24-Jan-2019**)
150. Microsoft: Microsoft AI principles. Communication (2019). https://www.microsoft.com/en-us/ai/our-approach-to-ai. (**Accessed: 01-Feb-2019**)
151. Cutler, A., Pribić, M. and Humphrey, L.: Everyday ethics for artificial intelligence. Ibm 48 (2018)
152. BMW Group: Seven Principles for AI: BMW Group Sets Out Code of Ethics for the Use of Artificial Intelligence. Press release (2020). https://www.press.bmwgroup.com/global/article/detail/T0318411EN/seven-principles-for-ai:-bmw-group-sets-out-code-of-ethics-for-the-use-of-artificial-intelligence?language=en. (**Accessed: 09-Mar-2021**)
153. Mökander, J., Morley, J., Taddeo, M., et al.: Ethics-based auditing of automated decision-making systems: nature, scope, andlimitations. Sci. Eng. Ethics **27**, 44 (2021). https://doi.org/10.1007/s11948-021-00319-4
154. AI HLEG: European Commission's Ethics Guidelines for Trustworthy Artificial Intelligence. (2019)
155. IEEE: Ethically aligned design. Intell. Syst. Control Autom. Sci. Eng. **95**, 11–16 (2019)
156. OECD: Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449 (2019)
157. Dunn, M., Hope, R.A.: Medical ethics: a very short introduction, 2nd edn. Oxford (2018)
158. AstraZeneca: Our therapy areas (2021). https://www.astrazeneca.com/our-therapy-areas.html. (**Accessed: 22-Nov-2021**)
159. Ashenden, S. K.: Introduction to drug discovery. In: Era Artif. Intell. Mach. Learn. Data Sci. Pharm. Ind., pp. 1–13 (2021)
160. Slee, T.: The incompatible incentives of private-sector AI. In: Dubber, M., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI. Oxford University Press, Oxford (2021)
161. Kroll, J.A.: "Accountability in Computer Systems", in The Oxford Handbook of Ethics of AI. Oxford University Press, Oxford (2021)
162. Powers, T.M., Ganascia, J.-G.: The ethics of the ethics of AI. In: Dubber, M., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI. Oxford University Press, Oxford (2021)
163. Legg, C. and Hookway, C.: Pragmatism. Stanford Encyclopedia of Philosophy (2020)
164. Salkind, N.J.: Encyclopedia of Research Design. SAGE, Los Angeles (2010)
165. Wang,, J. and Yan, Y.: The interview question. In: The SAGE Handbook of Interview Research: The Complexity of the Craft, SAGE Publications Inc., pp. 231–242 (2012)

166. Saldaña, J.: The Coding Manual for Qualitative Researchers. SAGE, London (2009)
167. Greene, D., Hoffmann, A. L. and Stark, L.: Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: Proc. 52nd Hawaii Int. Conf. Syst. Sci., pp. 2122–2131 (2019)
168. Diamandis, P., Kotler, S.: Abundance: The Future Is Better Than You Think. Free Press, New York (2012)
169. Pammolli, F., Righetto, L., Abrignani, S., Pani, L., Pelicci, P.G., Rabosio, E.: The endless frontier? The recent increase of R&D productivity in pharmaceuticals. J. Transl. Med. **18**(1), 1–14 (2020)
170. Legg, S. and Hutter, M.: A Collection of Definitions of Intelligence, pp. 1–12 (2007)
171. USDOD: US National Defence Authorization Act. Department of Defence: 115th Congress (2018)
172. McCarthy, J.: What is artificial intelligence?. Stanford Univ., (2007)
173. Mökander, J., Sheth, M., Watson, D., Floridi, L.: Models for classifying AI systems: the Switch, the Ladder, and the Matrix. Minds Mach. (2022)
174. Bryson, J.J.: The artificial intelligence of the ethics of artificial intelligence. In: Dubber, M., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI. Oxford University Press, Oxford (2021)
175. European Commission: Proposal for Regulation of the European Parliament and of the Council. Brussels, COM(2021) 206 final (2021)
176. European Commission: White Paper on Artificial Intelligence—A European Approach to Excellence and Trust," p. 27 (2020)
177. Dewey, J.: Reconstruction in philosophy, Enl Beacon Press, Boston (1957)
178. Gasser, U., Schmitt, C.: The role of professional norms in the governance of artificial intelligence. In: Dubber, M., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI. Oxford University Press, Oxford (2021)
179. Grimpe, C., Kaiser, U.: Balancing internal and external knowledge acquisition: the gains and pains from R&D outsourcing. J. Manag. Stud. **47**(8), 1483–1509 (2010)
180. Diakopoulos, N.: Transparency. In: Dubber, M., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI. Oxford University Press, Oxford (2021)
181. Mau, S.: The metric society: On the quantification of the social. UK; Medford, MA, Cambridge (2019)
182. Kaplan, R.S., Norton, D.P.: The Balanced Scorecard: Translating Strategy into Action. Harvard Business School Press, Boston (1996)
183. Floridi, L., Strait, A.: Ethical foresight analysis: what it is and why it is needed? Minds Mach **30**(1), 77–97 (2020). https://doi.org/10.1007/s11023-020-09521-y
184. Weiss, I.R.: Auditability of software: a survey of techniques and costs. MIS Q. Manag. Inf. Syst. **4**(4), 39–50 (1980)
185. Raisch, S., Krakowski, S.: Artificial intelligence and management: the automation–augmentation paradox. Acad. Manag. Rev. **46**(1), 192–210 (2021)