Ethics-Based Auditing of Automated Decision-Making Systems

Considerations, Challenges, and Paths Forward

Document submitted for partial fulfilment of the requirements for the degree of DPhil in Information, Communication and the Social Sciences

Oxford Internet Institute, University of Oxford

Jakob Mökander
Keble College
Trinity Term 2023

Supervisor:
Professor Luciano Floridi

Word count: 94,872

ABSTRACT

Decisions impacting human lives and livelihoods are increasingly being automated. While the use of *automated decision-making systems* (ADMS) improves efficiency, it is coupled with ethical risks. Previous research has pointed towards *ethics-based auditing* (EBA) as a promising governance mechanism for managing the ethical risks ADMS pose. However, the affordances and limitations of EBA have yet to be substantiated by empirical research.

This thesis seeks to clarify and resolve fundamental questions surrounding EBA. What are the limitations of EBA? How can feasible and effective EBA procedures be designed? These questions are approached on three levels. The *conceptual level* concerns what EBA is and how it works. The *descriptive level* focuses on the challenges organisations face when implementing EBA. The *applied level* concerns how to design EBA procedures that are feasible and effective in practice.

This is an integrated thesis, in which the substantive chapters (3–7) are based on published journal articles. Chapter 3 provides a theoretical explanation of how EBA contributes to good governance; Chapter 4 presents new empirical data from a case study of a real-world EBA implementation; Chapter 5 analyses the role of auditing in the proposed EU AI Act; Chapter 6 provides guidance on how to demarcate the material scope of EBA; and Chapter 7 outlines a blueprint for how to audit ADMS with highly general capabilities.

My findings suggest that EBA is subject to significant conceptual, technical, and institutional limitations. However, they also indicate that EBA – if properly designed and implemented – helps organisations identify and mitigate some of the ethical risks ADMS pose. I conclude by providing recommendations for how researchers, industry practitioners, auditors, and policymakers can facilitate the emergence of feasible and effective EBA procedures. This thesis thereby serves the purpose of better equipping societies to reap the benefits of ADMS while managing the associated risks.

TABLE OF CONTENTS

ABBREVIATIONS

ADMS	Automated Decision-Making Systems
AI	Artificial Intelligence
AAA	The Algorithmic Accountability Act of 2022 (US)
AIA	The Artificial Intelligence Act (EU)
AIEIG	AI Ethics Impact Group
AI HLEG	The High-Level Expert Group on Artificial Intelligence (EU)
BIKG	Biological Insight Knowledge Graphs
CDEI	Centre for Data Ethics and Innovation (UK)
CNIL	Commission Nationale de l'Informatique et des Libertés (France)
CSR	Corporate Social Responsibility
DEK	Datenethikkommission (Germany)
EBA	Ethics-Based Auditing
ECJ	European Court of Justice
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EIU	The Economist Intelligence Unit
ELI	European Law Institute
EPRS	European Parliamentary Research Service
EU	European Union
FAccT	Fairness, Accountability, and Transparency in ML
FDA	Food and Drug Administration (US)
FTC	Federal Trade Commission (US)
GAO	Government Accountability Office (US)
GDPR	General Data Protection Regulation (EU)
IAF	Information Accountability Foundation
ICO	Information Commissioner’s Office (UK)
IEEE	Institute of Electrical and Electronics Engineers
IIA	Institute of International Auditors
IIF	Institute of International Finance
IP	Intellectual Property
ISACA	Information Systems Audit and Control Association

ISO	International Organization for Standardization
LLM	Large Language Models
LoA	Level of Abstraction
ML	Machine Learning
NGO	Non-Governmental Organisation
NDA	Non-Disclosure Agreement
NIST	National Institute of Standards and Technology (US)
NLP	Natural Language Processing
OECD	The Organisation for Economic Co-operation and Development
OII	Oxford Internet Institute
PDPC	Personal Data Protection Commission (Singapore)
QMS	Quality Management Systems
RQ	Research Question
SQ	Subsidiary Research Question
UK	United Kingdom
US	United States
VDE	Verband der Elektrotechnik, Elektronik und Informationstechnik (Germany)

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

While writing itself can be a solitary exercise, knowledge generation is inevitably collaborative. Throughout my DPhil journey, I have benefitted from the support of many people. It would be impossible to list everyone. Nevertheless, I want to take this opportunity to thank some of the individuals without whom this thesis would not have been possible.

To start with, I wish to thank Luciano Floridi, my supervisor, mentor, and friend. Luciano's input has shaped this thesis on multiple levels, ranging from high-level feedback on the scope and aim of my research to detailed line-by-line comments on draft manuscripts. While I am indebted to Luciano for my academic success, his biggest contribution has been on a personal level. Ever since we first met, Luciano has believed in me, guided my scholarly work, and encouraged my curiosity. For this, I am and will always remain incredibly grateful.

I would also like to thank my co-authors. Several of the chapters included in this thesis are based on published journal articles written in collaboration with other researchers who brought with them their own backgrounds and skills. In addition to Luciano, Maria Axente, Federico Casolari, Hannah Kirk, Jessica Morley, Margi Sheth, Jonas Schuett, Mariarosaria Taddeo, and David Watson have co-authored different parts of this thesis. Their contributions have ensured not only the scholarly rigour but also the practical relevance of my work.

Further, I wish to thank AstraZeneca for supporting my research through a fully funded studentship. I am especially grateful to Peder Blomgren, Mimmi Sundler, Margi Sheth, and Wale Olamwini at AstraZeneca's R&D Data Office in Cambridge. Over a period of three years, they have provided me with unique industry insights and access to strategic research material for my observational case study. As a result, AstraZeneca's support has not only afforded me the opportunity to conduct this research but also provided me with the resources required to make novel empirical contributions. In addition to AstraZeneca, I would also like to thank Keble College and The Society of Swedish Engineers for their financial support.

As a DPhil student at the Oxford Internet Institute, I have benefited from both formal and informal feedback. I am especially grateful to those who have volunteered to assess my work at each milestone. For this, I would like to thank Victoria Nash, Ugo Pagallo, Adam Madhi, Viktor Mayer-Schoenberger, and Mariarosaria Taddeo. Their feedback has helped me sharpen the focus of my research and highlight its originality. So too has the guidance provided by Matthias Hollweg at the Said Business School. Further, I wish to thank Ralph Schroeder for his mentorship. Some of my best memories from Oxford include us discussing philosophy and sociology whilst walking around University Park. It would be remiss not to mention my peers

with whom I have often discussed my work, including Jess, Josh, Liam, Marta, Carl, David, Prathm, Andreas, Hannah, Cailean, Felix, Charlie, Chris, Clemie, Thomas, Phillip, Raphael, Leon, Jakob, Utkarsh, Jake, Archit, and Jordan. I have fond memories of meeting up with you at various cafes around Jericho to discuss everything from weekend plans to social theory.

I feel fortunate to have been part of several scholarly communities during my time as a DPhil student. Keble College has served as a second home for me. For this, I thank our director of graduate studies, Ian Archer. In the summer of 2022, I was a Fellow at the Centre for the Governance of AI, where I got to be part of an amazing group of scholars. I am grateful to Jonas Schuett, Emma Bluemke, Markus Anderljung, Ben Garfinkel, Allan Dafoe, and Anders Sandberg and for their support and intellectual sparring. During my final year, I was a Visiting Scholar at Princeton University's Center for Information Technology Policy. Amongst the faculty, I want to thank Arvind Narayanan, Mihir Kshirsagar, Tithi Chattopadhyay, Matthew Salganik, and Eszter Hargittai for creating the ideal environment for me to write up my thesis. I would also like to thank Sayash, Angelina, Varun, Klaudia, Nia, Archana, Shazeda, Christelle, Amna, Monica, Kenia, Sarah, and Jordan for their friendship and for feedback on my work.

So far, I have focused on acknowledging the people who have contributed directly to the shaping of this thesis. However, I would also like to thank the people who enabled me to be here in the first place – my family. I am incredibly grateful to my parents, Helena and Jurgen. Their unconditional love allowed me and my siblings, Daniel and Paula, to grow up in a home where curiosity was nurtured, where practical skills and theoretical knowledge were valued equally, and where we had the freedom to pursue our interests. Most importantly, they taught me and my siblings to treat all humans with respect and be generous when interpreting other peoples' actions and beliefs. I do my best to carry their values forward.

Most of all, I wish to thank Meghna Kulshrestha, my partner and best friend. Without her, I neither could nor would not have written this thesis. When Meghna and I first met, I worked at the Embassy of Sweden in New Delhi, India. I was happy with my job and had no plans to do a PhD. However, Meghna wanted to pursue graduate studies at Oxford and convinced me to apply to the OII so that we could be together in the UK. Following Meghna and her advice proved to be the best decision of my life. While the pandemic delayed some of our plans, we got to share the academic year 2020–2021 in Oxford, during which we lived together at St Antony's College. Meghna has thus shared each step of my DPhil journey: from applying to Oxford, via proofreading my first journal manuscripts, to celebrating holidays and breaks together. She is a living reminder that there are more important things in life than research. Thank you, Meghna, for changing me for the better. This thesis is for you.

CHAPTER 1

INTRODUCTION

1.1 Synopsis

Decisions that impact humans and the natural environment are increasingly being automated (AlgorithmWatch, 2019). This is understandable; delegating tasks to *automated decision-making systems* (ADMS) can improve efficiency and enable new solutions to complex problems (Taddeo & Floridi, 2018). However, the use of ADMS is coupled with ethical challenges, e.g., related to data privacy, bias and discrimination, and malicious use (Tsamados et al., 2020). The capacity to reap the potential benefits of ADMS while managing the associated risks posed is thus becoming a prerequisite for good governance (Cath et al., 2018).

Against this backdrop, researchers and policymakers have pointed towards *ethics-based auditing* (EBA) – a structured process whereby an entity's past or present behaviour is assessed for consistency with predefined ethics principles – as a promising yet underexplored governance mechanism[1] to manage the ethical risks ADMS pose (see e.g., Brown et al., 2021; Koshiyama et al., 2022; Raji et al., 2020; Sandvig et al., 2014). As I will argue in this thesis, the promise of EBA is underpinned by three ideas: that procedural regularity and transparency contribute to good governance; that proactivity in the design of ADMS helps identify risks and prevent harm before it occurs; and that the operational independence between the auditor and the auditee contributes to the objectivity and professionalism of the evaluation.

However, most research has hitherto referred to EBA as a theoretical proposition. This means that the merits and limitations of EBA as an ADMS governance mechanism have yet to be systematically explored – let alone substantiated by empirical research. This is not only a gap in the academic literature but also a pressing social problem. A new industry is emerging, whereby private companies offer EBA services to help organisations design or use ADMS

[1] I use the term *governance mechanism* to describe the set of activities and controls used by various parties to exert influence and achieve normative ends in society (more on this in Section 1.4).

ethically. However, without a shared understanding of what EBA is, claims that an ADMS has been audited are hard to verify and may even – in the absence of established best practices – do more harm than good by giving a false sense of security.

This thesis seeks to clarify and resolve fundamental issues surrounding EBA. What are the affordances and constraints of EBA as an ADMS governance mechanism? What practical challenges do organisations face when implementing EBA procedures? How can EBA complement other approaches to governing ADMS? And what do feasible and effective EBA procedures look like? Over the course of eight chapters, I critically examine existing work on EBA, conduct a longitudinal industry case study to present new observational data on how EBA procedures are being implemented in applied settings, and develop recommendations for how researchers, industry practitioners, auditors, and policymakers can facilitate the emergence of feasible and effective EBA procedures.

The research underpinning this thesis has both critical and constructive components. On the one hand, it critically investigates the conceptual, technical, and practical limitations associated with EBA. Remaining realistic about what EBA can and cannot do is important to avoid both unnecessary harms resulting from ADMS and corporate malpractices such as making unsubstantiated claims to appear more ethical than one is. On the other hand, it provides practical guidance to organisations seeking to ensure and demonstrate that the ADMS they design or deploy adhere to predefined ethics principles. By exploring not only *whether* but also *how* EBA can be a feasible and effective governance mechanism, the critical and constructive components of this thesis both serve the overarching purpose of better equipping societies to reap the benefits of ADMS while managing the associated risks.

This is an integrated thesis, meaning that the substantive chapters (3–7) are all based on peer-reviewed journal articles. In Section 1.9, I will provide an overview of the different chapters included in this thesis. However, one contribution should be stressed upfront. In Chapter 4, I report the findings from my longitudinal industry case study. Over 12 months, I observed and analysed the internal activities of AstraZeneca – a biopharmaceutical company – as it prepared for and underwent an EBA in collaboration with an external auditor. Drawing on this unique source of data, Chapter 4 provides new qualitative knowledge about how organisations implement EBA and the challenges they face in the process.[2]

[2] I gained access to my observational data through an institutional agreement between AstraZeneca and the University of Oxford, which also involved AstraZeneca funding my DPhil research. In Sections 1.7 and 1.8, I will discuss the methodology used to conduct the case study and the ethical considerations pertaining to it.

The remainder of this introductory chapter is divided into two parts. In the first part (Sections 1.2–1.5), I introduce my research topic through real-world examples to establish both the social importance and academic relevance of this thesis. In the second part (Sections 1.6–1.11), I outline my research questions, discuss my methodological approach, portray my target audience, provide an overview of the chapters included in this thesis, and demonstrate how they fit together in a larger narrative. The aim of this introductory chapter is to position the thesis as a whole in relation to previous research and larger societal developments. To do so, it is useful to first take a step back to review the context in which ADMS are being deployed.

1.2 Background: The promise and peril of automated decision-making

ADMS increasingly permeate all sectors of society (AlgorithmWatch, 2019; ELI, 2022). This means that ever more decisions – which were previously made by human experts – are now made by ADMS (Krafft et al., 2020a; Zarsky, 2016).[3] Consider recruitment as an example. Many companies use ADMS to streamline recruitment processes and reduce the time HR staff spend on manual tasks (Gupta & Mishra, 2023). The capabilities of ADMS used for recruitment vary from CV analysis to game-based assessments (Kazim et al., 2021). What they have in common is that they are trained on data from past job applicants and use predictive models to determine the likelihood that a candidate is suitable for a role (Raghavan et al., 2020).

Hiring decisions are among the most consequential that individuals face in their lives. But they are far from the only significant decisions being automated. ADMS are used in many other potentially sensitive areas like medical diagnostics (Grote & Berens, 2020) and the issuing of loans (Lee et al., 2020). For example, banks use ADMS to cluster prospective borrowers into different risk categories that determine whether a loan is granted and its interest rate (Sargeant, 2022).[4] As information societies mature, ADMS will likely be used to make ever more critical decisions with significant implications for individuals and groups.

The widespread use of ADMS is not an isolated phenomenon but goes hand in hand with larger transformations brought about by digitalisation (Zarsky, 2016). ADMS leverage the growing availability of digital data (Wiggins et al., 2023) and recent advances in machine learning (ML) research (Balas et al., 2020) to perform increasingly sophisticated tasks.

[3] See Section 1.3 for a discussion of how I use the term ADMS in this thesis.

[4] A recent survey revealed that 37% of banks have fully automated their credit scoring process (IIF, 2019).

Delegating tasks to ADMS can improve efficiency and enable new solutions to complex problems (Taddeo & Floridi, 2018). The European Commission (2020a) estimates that the use of ADMS will add 13% to the EU-28's cumulative GDP by 2030. While such estimates should be approached with a degree of scepticism, it is important to stress that the benefits of ADMS are not only economic but also social. For example, when used for hiring, ADMS can improve recruiters' working conditions (Leicht-Deobald et al., 2019) and create fairer outcomes by reducing the impact of human biases (Savage & Bales, 2017). In healthcare, the ADMS used in medical image recognition have enhanced diagnostic services and made them more widely accessible (Jiang et al., 2017; Kaushik et al., 2020).

These examples demonstrate two important points: first, that the automation of high-impact decisions is not a hypothetical future scenario but already widespread practice (Richardson, 2022); and second, that the use of ADMS can benefit individuals, organisations, and society (Lomborg et al., 2023). However, this account has so far shed light only on one side of the story. In addition to benefits, ADMS also bring ethical, legal, and social challenges. Understand why this is the case, it is useful to consider some recent failures of ADMS.

In 2021, Mark Rutte, the prime minister of the Netherlands, offered to resign following controversy surrounding a data-driven welfare fraud detection system referred to as 'SyRI.' Simplified, SyRI was an ADMS that used statistical models to flag potential benefit fraud based on data aggregated from government agencies (Meuwese, 2020). Whilst SyRI's purpose was to make state administration more efficient, it proved to cause significant real-world harm (van Bekkum & Borgesius, 2021). To begin with, SyRI's linking of personal data from multiple sources did not comply with the right to privacy under the European Convention on Human Rights.[5] Moreover, SyRI systematically discriminated against minorities (Rachovitsa & Johann, 2022). The scale of the scandal was pronounced. It is estimated that SyRI wrongly accused over 26,000 families of benefit fraud (Amaro, 2021).

Private companies have faced similar controversies. In 2018, Amazon's automated recruitment tool was found to discriminate against female candidates (Dastin, 2018). The ADMS had been trained on past resumes, which were predominantly from men. It had thus learned to prefer male candidates and to downgrade resumes containing words associated with women (Langenkamp et al., 2020). Amazon claims that their recruiters never relied solely on

[5] In 2020, a Dutch court decided that SyRI violated the right to privacy as it collected too much data without specifying clearly enough the reasons why (Tweede Kamer, 2020).

the automated recommendations. Still, the system was scrapped after only one year in operation. The incident demonstrates how ADMS trained on incomplete or unrepresentative data can both perpetuate existing societal biases and create new ones (Leslie, 2019).

ADMS can also enable unethical behaviours. Consider the scandal surrounding Cambridge Analytica, a political consultancy hired to assist Donald Trump's 2016 election campaign. The firm used Facebook data to profile approximately 87 million US voters and target them with political ads (Lapowsky, 2019). The decision that was automated in this case was the personalisation of content. In 2018, the US Federal Trade Commission (FTC) stated that Facebook had violated its privacy consent agreement and failed to protect users' data. Facebook subsequently reached a $5 billion settlement with the FTC, and Cambridge Analytica's then-CEO was banned from running businesses in the UK for seven years due to 'unethical services' (Shabong & Aripaka, 2020).

The above examples come from different sectors and jurisdictions. What unites them is that the use of ADMS broke not only applicable regulations but also citizens' trust. Yet these are not isolated incidents. Examples of ADMS causing allocational, reputational, or financial harm abound (Holweg et al., 2022), and research has repeatedly shown that ADMS can:

- *Misjudge*, i.e., make incorrect predictions or classifications (Matthews et al., 2019),
- *Discriminate* against specific individuals or groups (DeVries et al., 2019),
- *Exploit* sensitive information without consent (Vincent et al., 2019),
- *Distort* information (Robertson et al., 2018),
- *Remove* human responsibility (Martin, 2019), and
- *Enable* human wrongdoing (Tsamados et al., 2021).

These ethical and social challenges have given rise to popular anger and scholarly criticism over the incautious use of ADMS. These sentiments are reflected in the titles of bestselling books like *Weapons of Math Destruction* (O'Neil, 2016), *Algorithms of Oppression* (Noble, 2018), and *Automating Inequality* (Eubanks, 2019). Nevertheless, some researchers have cautioned that we must not throw out the baby with the bathwater. Human decision-makers and ADMS both have strengths and weaknesses (Baum, 2017). Human judgement, for instance, tends to be influenced by prejudices, hunger, and fatigue (Kahneman, 2011). Using ADMS can, therefore, sometimes lead to fairer decisions (Lepri et al., 2018). ADMS also make bureaucratic procedures more efficient and allow human decision-makers to access real-time

data (Zerilli et al., 2018). Moreover, fear about ADMS could hamper the adoption of well-designed technologies, thus creating significant social opportunity costs (Floridi et al., 2018).[6]

To summarise, the fact that ADMS are both relatively autonomous and readily adaptable is a double-edged sword, enhancing their abilities to deliver benefits and cause harm. Against this backdrop, the overarching purpose of this thesis is to better equip societies to reap the benefits of ADMS while managing the associated risks. In Sections 1.6–1.11, I will explain how my thesis helps advance this aim by exploring how EBA can help technology providers design and deploy ADMS in ways that align with their organisational values. However, first, more should be said about what ADMS are and why they are so hard to govern.

1.3 Terminology: Conceptualising automated decision-making systems

With *automated decision-making systems* (ADMS), I refer to autonomous and self-learning systems that gather and process data to make or inform decisions that impact individuals, groups, or the natural environment with little or no human intervention. To perform tasks like classification, planning, communication, and predictive analytics (Corea, 2019), ADMS rely on multiple statistical techniques – from decision trees to deep neural networks (Lepri et al., 2018; Lomborg et al., 2023). However, my focus in this thesis will not be on the underlying technologies but instead on the features of ADMS – such as autonomy and adaptability – that underpin both their socially beneficial and ethically problematic uses.

This conceptualisation has three advantages for the purpose of my thesis. First, it encompasses both logic-based (symbolic) expert systems and connectionist (sub-symbolic) ML systems.[7] Second, it highlights how ADMS operate with varying levels of complexity and autonomy to automate both task execution and policy optimisation. Third, it bypasses any discussions about machine consciousness by focusing solely on the observable characteristics and operations of ADMS. As noted by Esposito (2022), if machines appear intelligent, this is not because they have learned how to think like us but because we have learned how to communicate with them in ways that advance our purposes.

In fact, the word *system* in ADMS indicates that I am talking about a class of systems that differ from other systems with respect to their autonomy and adaptability as opposed to a

[6] 'Opportunity cost' is an economic term that refers to the value of the best alternative foregone where a choice needs to be made between mutually exclusive alternatives (Haberler, 1936). In policy-oriented research, it is common to talk about 'trade-offs' when referring to the same dynamic (Nilsson & Weitz, 2019).

[7] See Section 6.2 for a discussion about the differences between symbolic and sub-symbolic systems.

kind of intelligence that differs from human intelligence.[8] In previous literature, many different names have been given to what I call ADMS, including *algorithmic systems* (Ananny & Crawford, 2018), *AI systems* (OECD, 2022), *AI-based systems* (Saleiro et al., 2018), and *autonomous/intelligent systems* (IEEE SA, 2020). Despite this variety, my use of the term ADMS is standard in the literature on fairness, accountability, and transparency in ML (see e.g., AlgorithmWatch, 2019; Cobbe et al., 2021; Zarsky, 2016)

For the overarching framing of my thesis, I will use the term ADMS consistently.[9] The reason is twofold: first, the term best captures the technical features of the systems under investigation, and second, talking about ADMS avoids the distracting questions about consciousness that tend to come when discussing *artificial intelligence* (AI). In two chapters, however, I will use the more popular term *AI systems* (which is often used interchangeably with ADMS in the literature) for both scholarly and pragmatic reasons. AstraZeneca uses the term AI systems internally. Adopting the same terminology thus allowed me to stay closer to my observational data during my descriptive case study in Chapter 4. Similarly, when analysing the role of auditing in the AIA in the Chapter 5, I adopt the terminology used by the European Commission, for the simple reason that it makes communication easier.

This section has sought to conceptualise ADMS and describe how I use the term in this thesis. However, it has not sought to provide a hard and fast definition. Different researchers and industry practitioners that develop and implement EBA procedures define ADMS in different ways. I, therefore, view the definitions of ADMS that these researchers and practitioners use not as a criterion for inclusion or exclusion in my research but as part of the design of the EBA procedures that I study.[10] Having made those terminological remarks, I now turn to the governance challenges posed by ADMS and how they can be overcome.

1.4 Research area: Governing automated decision-making systems

The ethical risks posed by ADMS are real and pressing. But why are ADMS so hard to govern? And how have different stakeholders responded to that difficulty? By discussing these

[8] The levels of autonomy and adaptability displayed by ADMS are a matter of degree (Tasioulas, 2018). In some cases, they are fully autonomous; in others, they 'only' provide recommendations to human (Cummings, 2004).
[9] To be clear, I will also discuss and refer to related works in the field in terms of ADMS, even when the cited works use any of the above-listed terminological variations.
[10] This choice is reflected in the formulation of my research questions (Section 1.6), and discussed in Chapter 6, which argues that how an audit's material scope is defined constitutes an integral part of the auditing procedure.

questions, this section provides the context necessary to introduce EBA in the next. But first, I must define some key concepts.

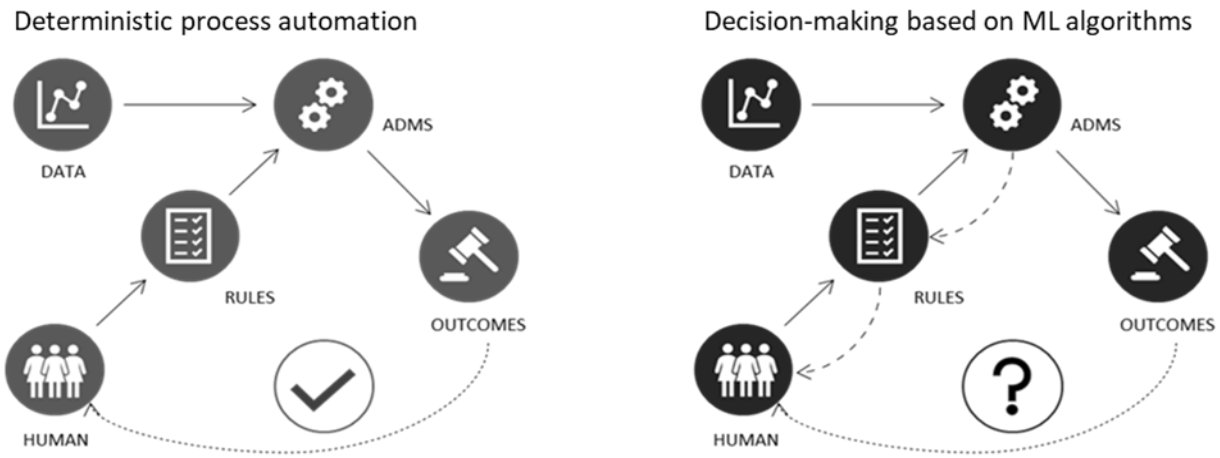
Following Baldwin and Cave (1999), I define *governance* as the process whereby various parties in society wield power, authority, and influence. The term *governance mechanisms* refers to the set of activities and controls different parties use to achieve their normative ends. Because governance is an interactive activity among stakeholders with competing interests, it is inevitably complex – even without the presence of new digital technologies (Chopra & Singh, 2018). However, as we will see, the appearance of ADMS that can perform morally relevant actions constitutes a major transformation in the fields of ethics and governance (Floridi & Sanders, 2004).

1.4.1 The governance gap

Since the 1960s, advances in computer science have been accompanied by concerns about the governance challenges ADMS pose (Samuel, 1960; Wiener, 1954). Hence, much has been written about what makes ADMS hard to govern by researchers from different disciplines. Recent contributions have been made by computer scientists (Russell, 2019), legal scholars (Pasquale, 2016), philosophers (Bostrom, 2014), communication scholars (Crawford, 2021), sociologists (Benjamin, 2019), and journalists (Christian, 2020). Here, I focus on two sets of governance challenges: those stemming from the features that characterise ADMS and those stemming from how ADMS transform decision-making processes. This discussion does not claim to be exhaustive but only to provide examples of how many governance mechanisms – originally designed to oversee human decision-makers – fall short when applied to ADMS.

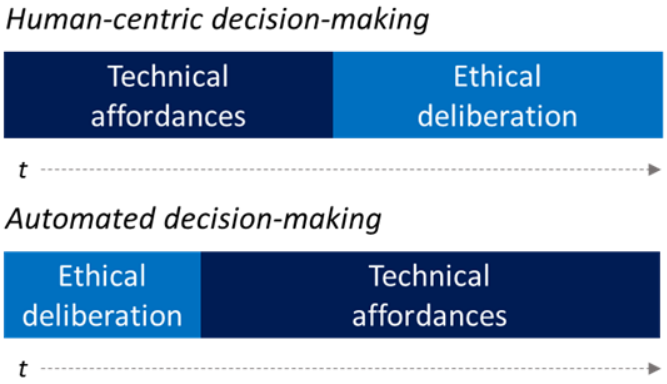
The first set of governance challenges posed by ADMS is rooted in their *autonomy*, *complexity*, and *adaptability*. For example, ADMS operating with a high degree of autonomy can display emergent behaviours as they interact with dynamic environments (Luckcuck et al., 2019). This makes their formal specification and verification uniquely challenging (EPRS, 2019). Further, the complex, often opaque, ways in which ADMS process information complicates the linking of system inputs and outputs (Citron & Pasquale, 2014). Finally, the adaptability of ADMS makes it hard to predict system failures and to assign accountability when harm occurs (Coeckelbergh, 2020). Traditionally, the actions of technical systems have been linked to its user, owner, or manufacturer. However, the ability of ADMS to learn, i.e., to adjust their decision-making logic over time, undermines existing accountability chains (Dignum, 2017). This shift from deterministic process automation to decision-making based on autonomous and self-learning ML algorithms is illustrated by Figure 1.

Figure 1. The difference between traditional process automation and autonomous, complex, and adaptable ADMS based on ML algorithms. Source: Verne and Mir (2019).



A second set of challenges stems from the ways in which ADMS transforming existing decision processes. While many different recruiters screen applicants' CVs at various companies, a single ADMS may screen all applicants' CVs at many companies (Bommasani et al., 2022). Using ADMS thus tends to reduce noise in decision processes (Kahneman et al., 2021). The flipside is that the effect of bias in an ADMS is likely to be amplified. Further, delegating tasks to ADMS erodes the room for discretion in decision-making processes (Dunleavy & Margetts, 2015) as norms that humans used to interpret are now embodied in ADMS (D'Agostino & Durante, 2018). Humans are reflective, meaning that they (often unconsciously) monitor the outcomes of their decisions and compare them with their values (van de Poel, 2020).[11] In contrast, ADMS execute tasks mechanically, meaning that any ethical deliberation must take place *upstream* in the technology design process. Figure 2 illustrates the logic behind this shift:

Figure 2. The use of ADMS both constrains the space for ethical deliberation in decision-making processes and shifts the burden from street-level bureaucrats to system developers.



[11] Humans possess emotions like empathy and a sense of duty that influence their behaviour (Lerner et al., 2015).

This governance challenge, i.e., the extent to which humans can ensure that the behaviour of the ADMS they design and deploy meet pre-specified objectives, has been termed the *alignment problem* (Kim et al., 2021). However, the main takeaway from this discussion is that our capacity to build ADMS has outpaced the development and adoption of governance mechanisms to ensure they are legal, ethical, and safe (Kroll et al., 2016).

1.4.2 Bridging the gap – from what to how

In response to these governance challenges, numerous institutions have published ethics principles that provide guidance to organisations that design and deploy ADMS (Fjeld, 2020; Jobin et al., 2019). Well-known examples include the *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019) and the *Asilomar AI Principles* (FLI, 2017). While differing in terminology, these tend to converge on five principles: beneficence, non-maleficence, autonomy, justice, and explicability (Floridi & Cowls, 2019). Additionally, many private companies – including Google (2018) and Microsoft (2019) – have published their own ethics principles.

This is a step in the right direction. However, to have real-world impact, ethics principles must be translated into verifiable criteria that can be locally enforced – and such translational attempts face many difficulties (Mittelstadt, 2019). For example, trade-offs are often required between different ethical principles, like privacy and fairness (Whittlestone et al., 2019a). Moreover, industries lack both incentives and tools to operationalise their ethical commitments (Raji et al., 2020). In short, there still exists a gap between the 'what' of ethics principles and the 'how' of designing and deploying ADMS in practice (Morley et al., 2020).

Promisingly, a wide range of governance mechanisms designed to bridge that gap have already been proposed. Some of these focus on interventions in the early stages of the ADMS development process, such as raising software developers' awareness of ethical issues (Floridi et al., 2018), creating more diverse engineering teams (Sánchez-Monedero et al., 2020), embedding ethical values into ADMS through proactive design (Aizenberg & van den Hoven, 2020), and screening input data for potential biases (AIEIG, 2020).

Other proposed governance mechanisms take the outputs of ADMS into account. One example is algorithmic impact assessments, i.e., procedures to evaluate the risks associated with an ADMS based on its design and intended purpose (Calvo et al., 2020). Another example is human-in-the-loop protocols, which imply that human operators can intervene to prevent or be held responsible for harmful outputs (Jotterand & Bosco, 2020; Rahwan, 2018).

Further, policymakers have proposed legally binding obligations. For example, the *Artificial Intelligence Act* (AIA) proposed by the (European Commission, 2021a) stipulates

that high-risk ADMS must undergo conformity assessments prior to their deployment. Similarly, the US Congress is currently debating the *Algorithmic Accountability Act of 2022* (AAA) (Office of US Senator Ron Wyden, 2022) which would require providers to justify how a specific ADMS would improve the decision process it has been designed to inform or replace.

In this section, I have not sought to provide an exhaustive overview of all ADMS governance mechanisms or to evaluate their relative merits and constraints. Instead, I wanted to stress two points that have direct implications for how I frame my research in this thesis. First, the fact that governments, private companies, and academic researchers are all working on developing new mechanisms to govern ADMS indicates that calls for EBA respond to a real-world problem. Second, addressing the ethical risks associated with ADMS will require multifaceted approaches. My focus on EBA in this thesis is thus not meant to diminish the merits of other governance mechanisms but rather to expand and sharpen the toolkit available to technology providers and policymakers wishing to analyse and evaluate ADMS.

1.5 Research topic: Ethics-based auditing

Many researchers have pointed towards EBA as a promising yet underexplored mechanism to govern ADMS (see e.g., Brundage et al., 2020; Raji & Buolamwini, 2019; Sandvig et al., 2014). In Chapter 2, I review previous literature in the field. The aim of this section is to make three more limited contributions: to define EBA; to establish EBA as an existing phenomenon; and to demonstrate that there remains a discrepancy between the attention EBA procedures attract and the lack of empirical research concerning their feasibility and effectiveness.

1.5.1 Calls for audits of ADMS

An audit is an independent examination of a phenomenon with the aim of expressing an opinion on it (Gupta, 2004). Auditing has a long history of promoting trust and transparency in areas like financial accounting and safety engineering (LaBrie & Steinke, 2019). The idea underpinning calls for audits of ADMS is thus straightforward: just as financial transactions can be audited for correctness and legality, so the design and use of ADMS can be audited for technical performance and adherence to relevant regulations or norms.

Following Sandvig et al.'s (2014) seminal article *Auditing Algorithms,* a growing body of research has focused on how audits can identify and mitigate the risks posed by ADMS. Broadly, contributions to this literature fall into two categories. The first consists of audit studies aiming to expose representational or allocational harms caused by ADMS. For instance, a much-cited study by Buolamwini and Gebru (2018) found that commercial face recognition

tools are less accurate for darker-skinned females than for lighter-skinned males. These studies tend to highlight the need for systematic and independent audits:

> *'Auditing is an essential strategy for detecting unintended bias and prompting the re-examination and revisions [of ADMS] to reduce discriminatory effects.'* (Kim, 2017, p.202)

The second category consists of articles that formulate theoretical justifications for why ADMS audits are required. The promise of auditing can be summarised by three core assumptions: that procedural regularity and transparency enable good governance (Floridi, 2017b); that proactivity in the design of ADMS helps prevent harm before it occurs (Kazim & Koshiyama, 2020); and that operational independence between auditors and auditees promotes objective and rigorous evaluations (Raji et al., 2022).

Policymakers too have shown a growing interest in auditing. For example, the UK Information Commissioner's Office (ICO, 2020) has issued guidance on how to audit ADMS. However, the most mature government regulation is currently found in the US In 2021, New York City enacted the *AI Audit Law* (NYC Local Law 144), requiring that ADMS used to inform employment-related decisions are made subject to independent audits:

> *'New York City's law will restrict employers from using ADMS in hiring and promotion decisions unless it has been the subject of a bias audit by an independent auditor no more than one year prior to use.'* (Gibson Dunn, 2023, p.1)

As these examples illustrate, audits can be used by different actors for different purposes: (i) by regulators seeking to assess whether an ADMS is legally compliant; (ii) by technology providers looking to mitigate technology-related risks; and (iii) by other stakeholders wishing to make informed decisions about how they engage with specific companies. By EBA, I refer to a subset of auditing procedures assess ADMS for adherence to voluntary adopted ethics principles, as opposed to legal compliance or technical robustness. To these, I turn next.

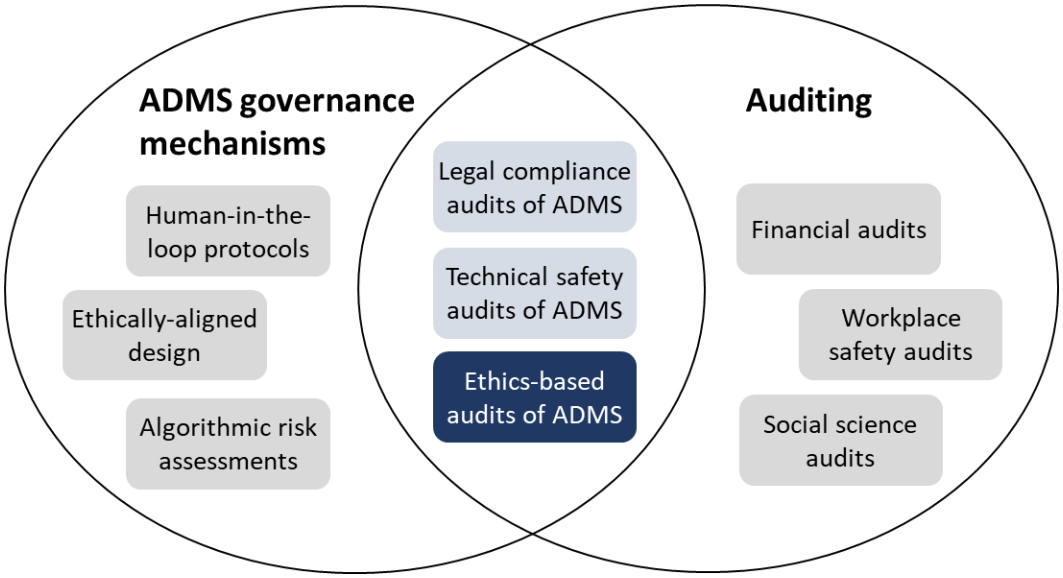
1.5.2 Defining ethics-based auditing

EBA can be defined both functionally (i.e., with respect to its intended function or purpose) and descriptively (i.e., with respect to its methodological characteristics). Functionally, EBA is a governance mechanism that organisations can use to operationalise their ethical commitments. Methodologically, EBA is characterised by a structured process whereby an entity's past or present behaviour is assessed for consistency with relevant principles or norms.

The subject of the audit can be a process, an organisation, or an ADMS. Yet, in each case, EBA is a structured and purpose-oriented process of demonstrating adherence to ethics principles.

Figure 3 illustrates how EBA is a subset both of auditing procedures (methodologically) and of mechanisms to govern ADMS (functionally). The literature review in Chapter 2 will explore what EBA can learn from audits in other fields, including finance and social science.[12]

Figure 3. A schematic overview of how EBA relates both to previous research in the field of ADMS governance and auditing in other fields of research and practice.



As Figure 3 shows, EBA is conceptually distinct from audits that assess ADMS exclusively based on legal compliance or technical robustness. In practice, however, there are overlaps since one auditing procedure can serve multiple purposes. However, by definition EBA concerns what ought and ought not to be done over and above existing regulations.

Different researchers use different terms to describe the same phenomenon. For example, LaBrie and Steinke (2019) use the term *ethical audits*. However, I prefer to use the expression 'ethics-based' instead of 'ethical' to avoid confusion: I do not refer to a kind of auditing conducted ethically but to auditing procedures that assess ADMS based on their adherence to relevant ethics principles.

[12] As a social science method, an audit study is a specific type of field experiment used to test for discriminatory behaviour in cases where survey or interview questions risk inducing undesirable biases (Gaddis, 2018).

1.5.3 A new industry emerges

From a societal perspective, the case for EBA is clear: the design and deployment of ADMS can be problematic and deserving of scrutiny even when not illegal. However, there are also many reasons why organisations subject themselves to EBA on a voluntary basis. Private companies have incentives to implement EBA procedures that improve business metrics like regulatory preparedness, data security, and reputational management (Holweg et al., 2022).

As a result of these pressures, a new industry is emerging. Professional services firms like Deloitte (2020), EY (2018), KPMG (2020), and PwC (2020) are all offering EBA services to clients on a commercial basis. At the same time, a growing number of startups – like Babl AI (2023), Credo AI (2023), Holistic AI (2023), and ORCAA (2020) – are attempting to fill the same market niche. On their official website, ORCAA describe their services as follows:

> *'ORCAA is a consulting company that helps companies and organizations manage and audit ADMS. ORCAA audits ADMS for accuracy, bias, consistency, transparency, fairness, and timeliness, following a 4-step approach.'*[13]

In parallel, academic researchers and NGOs are also developing and promoting different EBA procedures. Consider the following extract from an article by Raji et al. (2020) as an example:

> *'We introduce a framework for algorithmic auditing that supports the [software] development process end-to-end, to be applied throughout the internal organization development life-cycle. Each stage of the audit yields a set of documents that together form an overall audit report, drawing on an organization's values or principles to assess the fit of decisions made throughout the process.'* (Raji et al., 2020, p.1)

While Raji et al.'s procedure is sector-agnostic, other researchers have developed procedures to audit ADMS designed and used for specific purposes like medical diagnostics (Liu et al., 2022) or recruitment (Kazim et al., 2021). Another example is ForHumanity (2023), a non-profit organisation that relies on voluntary contributions from subject matter experts to develop procedures for independent audits of ADMS as well as to train and certify auditors. As of February 2023, ForHumanity has licenced their EBA procedure to 11 service providers and certified over 50 auditors.[14] In short, researchers and NGOs have not only stressed the need to audit ADMS but also provided blueprints for how to do so in practice.

[13] www.biasinai.com/directory/oneil-risk-consulting-algorithmic-auditing-orcaa/#about (Retrv. 19 Feb, 2023).
[14] https://forhumanity.center/certifications/certified-people/?v=920f83e594a1 (Retrv. 10 Feb, 2023).

1.5.4 The research gap

The above examples demonstrate that EBA of ADMS is an existing phenomenon, i.e., something that is already taking place and can be observed. They also suggest the existence of a strong case for why EBA procedures are needed. However, central claims about the feasibility and effectiveness of EBA as a governance mechanism have yet to be substantiated by empirical research. In what follows, I highlight five gaps in the existing literature.

First, there is widespread concern about conceptual confusion in the field (Landers & Behrend, 2022; Ng, 2021; Sloane, 2021). As noted by Vecchione et al. (2021):

> *'As [algorithmic] audits have proliferated, the meaning of the term has become ambiguous, making it hard to pin down what audits actually entail and what they aim to deliver.'* (Vecchione et al., 2021, p.1)

Without a shared understanding of what EBA is, let alone widely used standards for how it should be conducted, claims that an ADMS has been audited are difficult to verify and may potentially exacerbate rather than mitigate bias and harms (Costanza-Chock et al., 2022).

Second, there is a lack of empirically grounded research concerning the limitations of different EBA procedures. Raji and Buolamwini (2019) suggest that EBA *can* help check that ADMS engineering processes meet specific standards. Similarly, Brundage et al. (2020) argue that EBA *can* help verify claims about ADMS. However, the challenges organisations face in the process of implementing EBA procedures are best probed in applied contexts. The lack of case studies of real-world EBA thus constitute a critical gap in the existing literature, which has considered EBA as a theoretical proposition and discussed its merits in abstract terms.

Third, how the emerging EBA industry will be impacted by forthcoming regulations regarding the design and use of ADMS remains an open question. Previous research suggests that soft and hard governance mechanisms often complement and reinforce each other (Hodges, 2015). However, researchers like Munn (2022) have argued that soft governance mechanisms like EBA not only lack teeth but also are adopted by the private sector precisely to avoid or delay necessary regulation. Further research is, therefore, needed to understand the role EBA plays in relation to forthcoming regulations like the AIA.

Fourth, limited attention has been given to how specific policy design choices – like *who* conducts the audit, what its *material scope* is, and according to which *metrics* ADMS are evaluated – impact the feasibility and effectiveness of different EBA procedures. For example, Bandy (2021) stresses that researchers and practitioners have yet to develop robust ways of operationalising normative concepts during EBA:

> *'Audits should present clear metrics that quantify problematic behaviour of ADMS. Unfortunately, auditing presents the same potential for "p-hacking" that has plagued other scientific disciplines. Auditors can almost always find some metric that suggests inequity, discrimination, or other problematic behaviour.'* (Bandy, 2021, p.24)

Similarly, the lack of a clear material scope makes it difficult to enforce ADMS governance in practice (Kritikos, 2019; Scherer, 2016). This difficulty is especially acute for EBA procedures, which presuppose procedural transparency and operational regularity (Loi et al., 2020).

Finally, previous research has focused on developing procedures to audit ADMS that are employed for specific purposes. However, the capabilities of ADMS tend to become ever more general. In a recent article, Bommasani et al. (2021) coined the term *foundation models* to describe ADMS that can be adapted to a wide range of downstream tasks. Such models pose significant challenges from an auditing perspective (Bharadhwaj et al., 2021). For example, it is difficult to assess the risks that ADMS pose independent of the context in which they are deployed (Weidinger et al., 2021). This implies that further research is needed to develop auditing procedures for ADMS with highly general capabilities.

1.6 Research questions and limitations in scope

My research questions are intended to address some of the gaps in the literature identified in the previous section. However, it is important not to lose sight of the ethical challenges ADMS pose. After all, main purpose of this thesis is to better equip societies to reap the benefits of ADMS while managing the associated risks by exploring whether and how EBA can help organisations design and deploy ADMS in ways that align with their organisational values.

1.6.1 Research question(s)

Two overarching research questions (RQs) guid my research throughout this thesis:

RQ1 What are the limitations of EBA as a governance mechanism for identifying and mitigating the ethical risks posed by ADMS?

RQ2 How can EBA procedures be designed to effectively identify and mitigate the ethical risks posed by ADMS while being feasible to implement? [15]

[15] With effectiveness, I refer to the degree to which EBA produces its desired result. With feasibility, I refer to how easily EBA procedures can be implemented. Both are interpreted as qualitative, non-quantifiable, concepts.

These questions are intentionally broad, to reflect the larger purpose of my thesis. However, while such broad RQs are useful for directing a research project at a high level of abstraction (LoA), they need to be broken down into more specific questions that guide the data collection and analysis to address specific gaps in the academic literature (Agee, 2009).

To do so, I approach my RQs on three levels: *conceptual*, *descriptive*, and *applied*, each of which can be divided into several subsidiary research questions (SQs). The *conceptual level* concerns the proper meaning of EBA, i.e., what it is and how it works. As I demonstrated in Section 1.5, the affordances and constraints of EBA as an ADMS governance mechanism have yet to be systematically explored. The following SQ was thus formulated at this level:

SQ1 What are the affordances and constraints of EBA as a governance mechanism to address the ethical risks posed by ADMS?

The *descriptive level* concerns empirical questions that can be answered by observational or experiential data. In this case, the descriptive level highlights the reality faced by 'early adopters', i.e., organisations that are already designing and implementing EBA procedures. Again, as noted in Section 1.5, the existing literature on EBA contains few, if any, empirical case studies. To help bridge that gap, the following SQ was formulated at the descriptive level:

SQ2 How do organisations integrate EBA procedures with existing governance structures, and what challenges do they face in the process?

Finally, the *applied level* concerns the evaluation of different design decisions or policies in relation to a desired outcome. Taking an explicitly pragmatic stance (more on this in Section 1.7), my research focuses on real-world situations and goes beyond evaluating existing options to propose new solutions.[16] The following three SQs were formulated at this level:

SQ3 How can EBA complement legislative approaches to managing the risks ADMS pose?

SQ4 How can the material scope for EBA be demarcated?

SQ5 What could blueprints for feasible and effective EBA procedures look like for ADMS with highly general capabilities?

[16] In Chapter 2, I will demonstrate how each of these questions address not only gap in the current literature but also pressing challenges faces by industry practitioners, auditors, and policymakers.

In this section, the SQs have been listed in a logical order. For example, a grounded knowledge of the challenges organisations face when implementing EBA (SQ2) is a precondition for understanding what feasible and effective EBA procedures could look like (SQ6). However, not all SQs are mutually dependent. Hence, the numbering of the SQs does not perfectly map onto the chronological order in which studies have been undertaken.

Further, the distinction between conceptual, descriptive, and applied research must not be overemphasised: even theoretical notions can be interpreted by tracing their practical consequences (James, 1975). Still, I found the distinction useful since questions on different levels require different methods to be answered, as I will expand on in Section 1.7.

1.6.2 Limitations in scope

Three limitations help narrow the scope of my research. First, any comparison between – or normative evaluation of – different sets of ethics principles falls outside the scope of this thesis. Previous work has shown that the apparent consensus around high-level ethics principles hides deep political disagreements, e.g., in terms of how concepts like fairness should be interpreted (Schiff et al., 2021). Moreover, different principles that are all desirable in isolation may sometimes create tensions for which there are no fixed solutions (Whittlestone et al., 2019). In practice, this means that organisations must strike justifiable trade-offs within the limits of legal permissibility and operational viability. However, for the purpose of this thesis, I assume normative clarity, i.e., that an organisation seeking to implement EBA has already committed itself to a coherent set of ethics principles.

Second, the thesis does not address any legal aspects of auditing. I will neither study audits directed exclusively towards ensuring legal compliance nor discuss the liability auditors may have in their contractual relationship with clients. My research focuses on how organisations ensure and demonstrate adherence to voluntary ethics principles. This does not mean that regulation is unimportant: ADMS should be lawful, ethical, and robust (AI HLEG, 2019). Moreover, legal considerations influence both the incentives organisations face and the design of EBA procedures. Throughout the thesis, I will, therefore, frequently refer to regulatory proposals like the AIA and the AAA for illustrative purposes.

Third, I do not engage with questions concerning *good intent*. Ideally, organisations that publish ethics principles also seek to embody those values. In practice, however, ethical commitments can be undermined by malicious intentions or skewed incentives (Floridi, 2019b). Although important, these considerations lie outside the scope of my thesis. Instead, I

take as a starting point the premise that even well-designed EBA procedures may facilitate but never guarantee morally good outcomes.

1.7 Methodology

This thesis consists of five substantive chapters (Chapters 3–7), each addressing one of the SQs listed above. Because different SQs have different methodological considerations, I leave it to each chapter to describe its methods. My aim in this section is to discuss my research stance and outline my overall methodological approach.

1.7.1 Research stance

There is no one way to conduct research. Research differs both explicitly, in terms of the methods and theories applied, and implicitly, in terms of what motivates a researcher's choice of topic (Alford, 1998; Swedberg, 2014). In this thesis, I take an explicitly pragmatist stance.

With origins in nineteenth-century American thought – particularly the works of C. S. Peirce (1903), William James (1907), and John Dewey (1920) – pragmatism has a rich history in both philosophy and the applied sciences. Simplified, it holds that conceptual advances are only valuable insomuch as they are useful for some specific purpose.

I choose to adopt a pragmatist stance for my doctoral research for three reasons. First, the *pragmatist maxim* states that theories should be judged by their practical implications (Legg & Hookway, 2020). This resonates well with my research aim of helping organisations that design or deploy ADMS operationalise their ethical commitments.

Second, pragmatists maintain that research should not only be grounded in real-world problems but also solution-oriented (Salkind, 2010). According to Prasad (2021), problem-solving occurs when a community examines the empirical world with the aim of changing it. As previously discussed, the governance challenges posed by ADMS require urgent problem-solving. My SQs on the applied level were designed to address these real-world problems.

Third, pragmatists do not separate knowing the world from acting within it (James, 1907). Pragmatists thereby maintain the freedom to use any research methods, whether inductive or deductive, quantitative or qualitative, provided they produce actionable results (Hacking, 1983). My research uses a multimethod approach that affords triangulation of data from different sources (Webb, 1966). Whilst triangulating findings may be advisable in any research, it is particularly important in rapidly emerging fields like EBA.

A further clarification. My choice of topic neither endorses the claim that ADMS *should* be audited nor suggests that EBA is *more promising* than other ADMS governance

mechanisms. Instead, I seek to explore whether and how EBA can help organisations design and deploy ADMS in ways that align with their organisational values. I argue that this stance, combining critical and constructive elements, does more to further the emergence of rigorous mechanisms to govern ADMS than either normative endorsements or rebuttals.

1.7.2 Overarching methodological approach

As mentioned above, I approach my RQs on three levels: conceptual, descriptive, and applied. The studies on each level use different research methods and address different audiences.

Conceptual studies refer to research methods that analyse or synthesise already available information. By bridging existing theories, or linking work across disciplines, conceptual studies can provide new insights (Gilson & Goldberg, 2015). Conceptual studies are particularly useful in fields of research that draw upon multiple theoretical strands (Jaakkola, 2020). As a field of research, EBA is both inherently multidisciplinary and rapidly evolving. Hence, before conducting any descriptive or applied studies, I deemed it appropriate to systematise and synthesise existing knowledge about EBA.

The conceptual study of this thesis aims to address SQ1, i.e., what are the affordances and constraints of EBA as a governance mechanism to address the ethical risks posed by ADMS? To answer this question, I combined a *systematised literature review* (Grant & Booth, 2009) and *theory synthesis* (Jaakkola, 2020). The research process consisted of two steps. First, I systematically review previous work on EBA to define core concepts and structure the findings of previous work. Second, I synthesised the findings and theories found in previous research to achieve an improved understanding of what EBA *is*, or at least *ought to be*, in the context of ADMS. I will expand on the methodology used to address SQ1 in Chapter 3.

Descriptive studies aim to provide detailed descriptions of empirical phenomena without regard to a specific hypothesis. Consequently, they are often the first foray into a new area of inquiry (Grimes & Schulz, 2002). In Section 1.5, I demonstrated that many EBA tools and methods have already been developed. However, at the time I started my DPhil no study of how organisations implement EBA had been conducted. I thus set out to explore SQ2, i.e., how do organisations integrate EBA procedures with existing governance structures, and what challenges do they face in the process? To address this question, I conducted a longitudinal *industry case study* (Bass et al., 2018), studying the implementation of EBA in AstraZeneca.

The case study leveraged qualitative methods, including *participant observation* (Woodside, 2016) and *semi-structured interviews* (Edwards & Holland, 2013). Participant observation is a method well-suited to making sense of organisational practices (Vinten, 1994)

which works best when a researcher is embedded within an organisation long enough to observe how it operates. Therefore, I observed and analysed the activities of AstraZeneca's R&D Data Office team over a period of 12 months, as they prepared for and underwent an EBA. I also conducted semi-structured interviews with managers and software developers within AstraZeneca, enabling me to collect data on the motivations behind specific decisions on how to and the perspectives of different internal stakeholders.

In Chapter 4, I will describe the methodology used to conduct the case study in detail. Still, something should be said here about why I choose this case study, and how I secured access to my observational data. I gained access to this unique data through an institutional agreement between AstraZeneca and the University of Oxford, which also involved that AstraZeneca funded my DPhil research. It is often difficult for researchers to gain permission for conducting participant observation in industry settings and I was fortunate to get access to AstraZeneca's processes, documentation, and staff via this institutional arrangement.[17]

However, my choice to study the implementation of EBA in AstraZeneca was not just one of convenience. AstraZeneca's EBA constituted what Merton (1987) calls strategic research material for SQ2 for three reasons. First, AstraZeneca uses ADMS for a variety of tasks (e.g., to detect treatment response patterns), allowing me to study ADMS governance in an applied setting. Second, as a biopharmaceutical company, AstraZeneca has a long history of adhering to ethics principles. This meant that the practical challenges they faced with respect to governing ADMS overlapped with the theoretical problems I sought to address. Third, the timing was advantageous. As I embarked on my DPhil research in fall 2020, AstraZeneca had just decided to conduct an EBA in Q4 2021. This gave me the unique opportunity to observe the entire process as the organisation prepared for and conducted their first-ever EBA.

Finally, applied studies aim to find and evaluate alternative solutions for immediate problems facing individuals, organisations, or societies (Bickman & Rog, 2008). My pragmatist research stance outlined above is thus well-suited to conduct applied studies. To recap, I formulated three research questions on the applied level: SQ3, i.e., how can EBA complement legislative approaches to managing the risks ADMS pose?; SQ4, i.e., How can the material scope for EBA be demarcated?; and SQ5, i.e., what could blueprints for feasible and effective EBA procedures look like for ADMS with highly general capabilities?

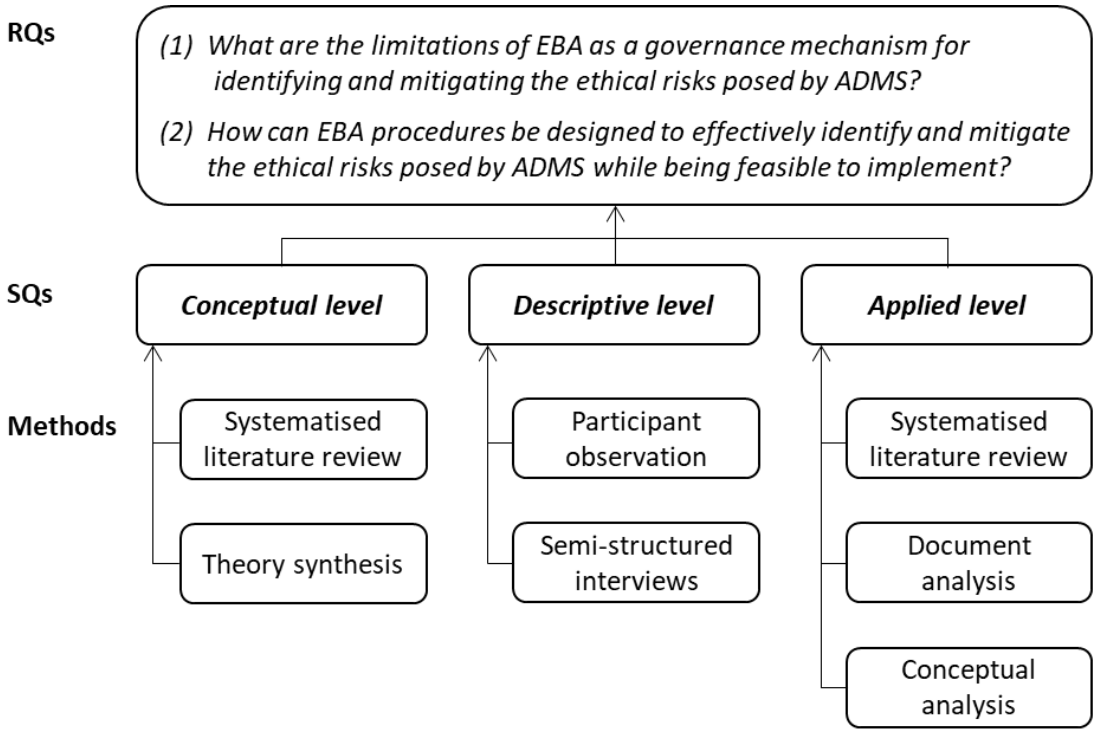
[17] In Section 1.8 I consider the ethical aspects of this arrangement.

To answer the SQs at the applied level, I used a combination of *systematised literature review* (Grant & Booth, 2009), *document analysis* (Karppinen & Moe, 2012) and *conceptual analysis* (Maggetti et al., 2015). I started by conducting a systematised review of previous academic literature to synthesise what is already known with respect to each SQ. Thereafter, I analysed documents written by people addressing these SQs from software engineering or policy perspectives.[18] Finally, I engaged in conceptual analysis, i.e., breaking down concepts into simpler elements to promote clarification or find new solutions (Furner, 2006).

Since EBA is a relatively new field of research, central concepts are still being shaped. Whilst this is a problem for practitioners, it is can be an opportunity for researchers. According to Merton (1948), research can play an active role in reformulating and clarifying practically applicable concepts. Hence, although each of the Chapters 5–7 address different SQs, they all contribute to sharpening and expanding the conceptual toolkit available to organisations who wish to audit the design and use of ADMS for alignment with specific ethics principles.

Figure 4 provides an overview of the different methods that I used to answer SQs at the conceptual, descriptive, and applied levels.

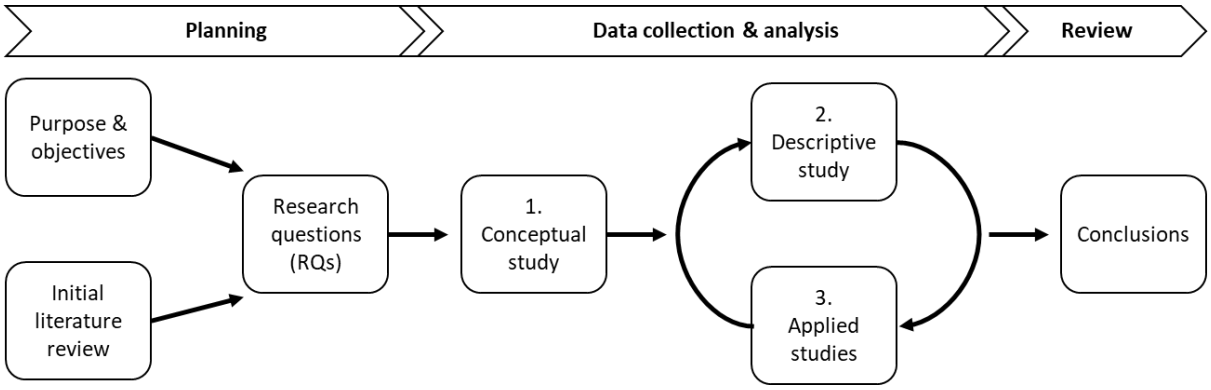
Figure 4. Overview of the research methods employed to address my SQs at the conceptual, descriptive, and applied levels, respectively.



[18] This includes documents such as EBA frameworks, ADMS demand specifications, and design logs produced by policymakers, think tanks, private consultancies, and industrial firms.

A final methodological remark. My research design leverages *instrument development*, which is one of the affordances of multimethod approaches. Instrument development refers to the process whereby different types of data collection and analysis inform each other (Pope & Mays, 1995). As Figure 5 sketches, I used a *parallel research design* (Creswell & Clark, 2011) in which research tied to the descriptive and the applied levels were conducted in parallel to form an iterative process. The purpose thereby was to ensure that my applied research was grounded in real-world problems experienced by organisations attempting to implement EBA.

Figure 5. High-level methodological flowchart: An iterative approach.



Section 1.9 summarises the results of my five substantive chapters and explains how they all fit together into an overarching thesis. Before doing so, however, the next section discusses ethical considerations relevant to my thesis research.

1.8 Ethical considerations

The ethical considerations relevant to this thesis are centred around the case study of AstraZeneca's 'AI audit.' To start with, the case study involved interviews with AstraZeneca employees. Following best practices for research with human participants, I accounted for two distinct ethical concerns: *informed consent* and *informational privacy*. When conducting the interviews, I secured informed consent from all participants and declared my role as an independent researcher. Data from the interviews were stored in a secure manner and will not be reused for secondary purposes without prior consent. Moreover, I did not record personal data, i.e., any information related to an identifiable natural person (ICO, 2018).

In addition to interviews, the industry case study relied on participant observation. Here, questions of positionality become important. Following Richards (2009) advice on conducting qualitative case studies, I did not set out to test a specific hypothesis. Instead, I took an

exploratory position, letting the data drive the analysis. Still, it is inevitable that my previous experiences influenced what I choose to focus on and shaped my conclusions (Given, 2008).

A particular concern relates to the fact that the case study was conducted in collaboration with a private company, AstraZeneca, and that my research is funded by an Oxford-AstraZeneca studentship. When such dependencies exist, researchers may feel pressure to produce results that are agreeable to their private sector partner (Maruyama & Ryan, 2014). To manage that risk, I took two proactive measures. First, I ensured that the studentship funded by AstraZeneca was administered by the University. Hence, there has been no direct transactions between AstraZeneca and me as a doctoral researcher. Second, I communicated clear boundaries regarding my role as an independent researcher and made sure that the people I worked with at AstraZeneca understood the critical nature of my work.

My research design has been considered by the OII's Departmental Research Ethics Committee (DREC) and was judged to meet appropriate ethical standards (research ethics approval reference number: SSH_OII_CIA_21_097).

A final set of ethical questions goes concerns the *transformative effects* on society that a research project may have (Denzin & Lincoln, 2018). The social implications of abstaining from my research on EBA would be an increased risk that ADMS continue to cause harm and amplify societal inequalities. Hence, my hope is that the ethical legacy of my research will be to mitigate some of the harm that ADMS may elsewise have caused.

1.9 Thesis structure and outline

In this section, I explain how the thesis is structured and provide an overview of the scope and contributions of each chapter. Taken together, my aim in this section is to help readers assess the contributions of this thesis in their proper context.

1.9.1 Thesis structure

I chose to write an integrated thesis for two reasons. First, because EBA is a rapidly developing field, it made sense to publish research findings as they appeared. Second, an integrated thesis allows for greater flexibility to adapt the framing based on the audiences of different studies. This is particularly important since my research is inherently multi-disciplinary in nature.

The thesis has eight chapters. Chapters 3–7 constitute the substantive part of my research, with the remaining chapters providing appropriate lead-in and lead-out material. The five journal articles on which the substantive chapters of this thesis are based have all been completed in collaboration with other researchers, including my supervisor Professor Luciano

Floridi. I am the sole first author of each article, meaning that I formulated the research questions, collected and analysed the data, drafted the manuscripts, and guided them through the peer review process. My co-authors contributed with domain expertise, reviewed the manuscripts, and suggested revisions that I incorporated. To further clarify my role in these co-productions, I have attached statements from all co-authors (see Appendix 1-9). At the beginning of each chapter, I credit my co-authors and cite the original publication on which it is based. For consistency, I use the first-person pronoun throughout the thesis.

1.9.2 Thesis outline

Chapter 2. Literature review

Chapter 2 reviews the literature relevant to this thesis. It consists of three parts. In the first I survey the evolution of auditing as a governance mechanism. I show that auditing has a long history of promoting transparency and accountability in areas like financial accounting and safety engineering – and argue that valuable lessons can be learned from these domains. In the second part, I draw on recent societal developments to demonstrate that the current drive towards developing ADMS auditing procedures results from a confluence of top-down and bottom-up pressures. I argue that both technology providers and policymakers have an interest in promoting auditing as a mechanism to govern ADMS, and – therefore – that it is left to academic researchers to study the feasibility and effectiveness of EBA procedures. In the final part, I review relevant academic literature. Because ADMS auditing is a multidisciplinary field of research, this part surveys developments in computer science, systems engineering, social science, political and moral philosophy, law, and organisational studies.

Chapter 3. Ethics-based auditing of automated decision-making systems: Nature, limitations and scope

Chapter 3 is the first of the five substantive chapters. Approaching my RQs at the conceptual level, it addresses SQ1, i.e., what are the affordances and constraints of EBA as a governance mechanism to address the ethical risks posed by ADMS? Based on a systematised literature review and theory synthesis, I make three contributions to the existing literature in this chapter. First, I provide a theoretical framework for how EBA contributes to good governance by promoting procedural regularity and transparency. Second, I propose seven criteria for successfully designing and implementing EBA procedures. Specifically, I highlight the need for EBA procedures to be continuous, i.e., to monitor and assess ADMS outputs over time.

Third, I demonstrate that existing EBA procedures are subject to both theoretical and practical constraints and provide a novel taxonomy to help researchers understand and account for these constraints. Taken together, the chapter provides the conceptual foundation for the thesis as a whole and contributes directly to addressing both of my RQs.

Chapter 3 is a lightly edited version of a peer-reviewed journal article published in *Science and Engineering Ethics*. For details, see:

> Mökander, J., Morley, J., Taddeo, M., Floridi. L. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Sci Eng Ethics* 27, 44 (2021). doi.org/10.1007/s11948-021-00319-4.

Chapter 4. Operationalising AI governance through ethics-based auditing: An industry case study[19]

Chapter 4 constitutes the main body of empirical research conducted for this thesis. It seeks to answer SQ2, i.e., how do organisations integrate EBA within existing governance structures, and what challenges do they face in the process? The chapter describes and analyses the internal activities of AstraZeneca as it underwent an EBA in collaboration with an external auditor. My findings suggest that the difficulties organisations face when conducting EBA include harmonising standards across decentralised organisations, demarcating the audit's scope, driving internal change management, and measuring outcomes. I argue that these findings from AstraZeneca can be generalised. Focusing on the descriptive level of my RQs, Chapter 4 anchors the research in an applied context and provides an empirical benchmark for the thesis. To my best knowledge, Chapter 4 of this thesis presents the first-ever case study of a real-world EBA published by any academic researcher.

Chapter 4 is a lightly edited version of a peer-reviewed journal article published in *AI and Ethics*. For details, see:

> Mökander, J, Floridi, L. Operationalising AI governance through ethics-based auditing: an industry case study. *AI Ethics* (2022). doi.org/10.1007/s43681-02200171-7.

Chapter 5. Conformity assessments and post-market monitoring: A guide to the role of auditing in the EU AIA[19]

Chapter 5 is the first of three consecutive chapters that focus on the applied level of my RQs. This chapter seeks to answer SQ3, i.e., how can EBA complement legislative approaches to

[19] In Chapters 4 and 5, I use the term 'AI system' instead of ADMS for the reasons described in Section 1.3.

managing the risks ADMS pose? It does this by analysing the EU AIA. My argument proceeds in three steps. First, I describe the two primary governance mechanisms proposed in the Act: 'conformity assessments' and 'post-market monitoring.' Second, I argue that the AIA de facto proposes to establish a Europe-wide ecosystem for conducting 'AI audits.' Third, I show that the AIA encourages providers of non-high-risk ADMS to adopt and adhere to voluntary codes of conduct. This indicates that EBA procedures are compatible with, and complementary to, hard regulations concerning the design and use of ADMS. The chapter concludes by highlighting areas where potential amendments to the AIA would help strengthen its overall effectiveness from an auditing perspective. In addition to answering SQ3, Chapter 5 thus contributes to one of my main objectives, which is to provide recommendations for how policymakers can support the emergence of EBA procedures that are feasible and effective in identifying and mitigating the ethical risks posed by ADMS.

Chapter 5 is a lightly edited version of a peer-reviewed journal article published in *Minds and Machines*. For details, see:

> Mökander, J., Axente, M., Casolari, F., Floridi, L. Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds & Machines* (2022). doi.org/10.1007/s11023-021-09577-4.

Chapter 6. The Switch, the Ladder, & the Matrix: Models for classifying automated decision-making systems

Chapter 6 responds to SQ4. i.e., how can the material scope for EBA be demarcated? The chapter's premise is that a major obstacle to implementing EBA is the lack of a clear material scope. I review previous attempts to classify ADMS for governance purposes. I find that such attempts use one of three models: *the Switch*, a binary approach according to which systems either are or are not considered ADMS depending on their characteristics; *the Ladder*, a risk-based approach that classifies systems according to the ethical risks they pose; and *the Matrix*, a multi-dimensional classification that takes various aspects into account, including context, data inputs, and decision-models. I conclude the chapter by discussing how different ways of demarcating the material scope of ADMS governance would impact the feasibility and effectiveness of EBA procedures. In doing so, I provide organisations that design or implement EBA procedures with the vocabulary they need to have an informed discussion about available policy options regarding how to demarcate their material scope.

Chapter 6 is a lightly edited version of a peer-reviewed journal article published in *Minds and Machines*. For details, see:

Mökander, J., Sheth, M., Watson, D.S., Floridi, L. The Switch, the Ladder, and the Matrix: Models for Classifying AI Systems. *Minds & Machines* 33, 221–248 (2023). doi.org/10.1007/s11023-022-09620-y.

Chapter 7. Ethics-based auditing of large language models: A three-layered approach

This chapter seeks to answer SQ5, i.e., what could blueprints for feasible and effective EBA procedures look like for ADMS with highly general capabilities? It approaches that question by investigating how EBA can help address the governance challenges posed by large language models (LLMs), a subset of foundation models that have recently attracted much attention from researchers and policymakers alike. The chapter's main contribution is a blueprint for auditing LLMs that draws on best practices from IT governance and system engineering. Specifically, I propose a three-layered approach, whereby *governance audits* (of technology providers that design and disseminate LLMs), *model audits* (of LLMs after pre-training but prior to their release), and *application audits* (of applications based on LLMs) complement and inform each other. I argue that this three-layered approach provides a blueprint not only for how to audit LLMs but also for how to design feasible and effective EBA procedures for ADMS that are highly adaptable to a wide range of downstream applications more generally.

Chapter 7 is a lightly edited version of a peer-reviewed journal article published in *AI and Ethics*. For details, see:

Mökander, J., Schuett, J., Kirk, H.R., Floridi, L. Auditing Large Language Models: A Three-Layered Approach. *AI Ethics* (2023). doi.org/ 10.1007/s43681-023-00289-2.

Chapter 8. Conclusions

In this chapter, I review my theoretical and empirical findings to answer my overarching RQs and address my larger research objectives. With respect to RQ1, I conclude that EBA procedures are subject to a wide range of conceptual, technical, economic, social, and institutional limitations. This implies that researchers, industry practitioners, auditors, and policymakers need to exercise caution and remain realistic about what EBA can reasonably be expected to achieve. With respect to RQ2, I conclude that – to be feasible and effective, should satisfy five conditions. They should (i) be structured and transparent, (ii) assess a clearly defined material scope according to an equally clearly defined normative baseline, (iii) incorporate elements of both technology-oriented assessments of ADMS and process-oriented assessments of organisations that design and deploy ADMS, (iv) include continuous monitoring of ADMS, and (v) be conducted by independent third-party auditors. I end the

chapter by providing recommendations for how policymakers can support and facilitate the emergence of feasible and effective EBA procedures.

1.10 Target audience and research objectives

In this section, I portray the different target audiences for this thesis. In doing so, I also articulate two larger research objectives that my research will help further. Given that each substantive chapter addresses its own SQ, I anticipate that different parts of this thesis will be of interest to different target audiences. However, at a high LoA, I envision four distinct target audiences for this thesis.

My first target audience consists of *academic researchers* who are interested in exploring the affordances and limitations of EBA of ADMS as a governance mechanism. This includes members of the FAccT community, i.e., scholars concerned with fairness, accountability, and transparency in ADMS, to which I have presented my work at several conferences over the last few years.[20] Like the FAccT community, my academic target audience is cross-disciplinary, brought together by shared research interests rather than disciplinary background. As my literature review in Chapter 2 will demonstrate, previous contributions to the EBA literature have been made by scholars from from computer science, systems engineering, law, media and communication studies, social science, philosophy, and organisational studies. All chapters of my thesis build on and add to this body of literature. However, I believe my academic target audience will find the conceptual contributions of Chapter 3 and the empirical case study presented in Chapter 4 most relevant to their work.

My second target audience consist of *auditors* that offer EBA services. Auditors in this context refers to a heterogeneous group that includes professional services firms (like PwC, Deloitte, EY, and KPMG), startups (like ORCAA, Holistic AI, and Babl AI) and non-profit organisations (like ForHumanity and the Algorithmic Justice League).[21] Despite having different incentives, all these organisations have developed EBA procedures. Hence, the criteria for how to design and implement feasible and effective EBA procedures listed and discussed in Chapter 8 are especially relevant to auditors, allowing them to improve and expand their service offering.

[20] FAccT doubles as the name for an annual conference on Fairness, Accountability, and Transparency in ML. I attended FAccT 2022 in Seoul, at which I presented the research underpinning Chapter 6 of this thesis.

[21] I have collaborated with several of the listed organisations throughout the course of my DPhil research. For example, one of my co-authors for Chapter 5, Maria Axente, is Responsible AI Lead at PwC.

Industry practitioners responsible for implementing EBA procedures in organisations that design or deploy ADMS constitute a third target audience for my thesis. Typically, this responsibility falls on managers with titles like Chief Information Officer, Responsible AI Lead, Head of IT Compliance, or Internal Audit Director. However, titles vary between companies, and sometimes the responsibility is shouldered by 'internal champions' who seek to implement EBA procedures not because of their official role description but rather their personal desire to do the right thing. Industry practitioners are likely to find my applied research (especially Chapters 5 and 6) most relevant to their work.

Finally, my research targets *policymakers* who mandate or design EBA procedures as part of larger efforts to govern ADMS. This includes the European Commission, the Federal Trade Commission (FTC) and the Government Accountability Office (GAO) in the US, as well as the Information Commissioner's Office (ICO) and the Center for Data Ethics and Innovation (CDEI) in the UK. Each of my substantive chapter provides recommendations to policymakers. That said, I believe that policymakers will find the novel blueprint for how to audit ADMS with highly general capabilities outlined in Chapter 7 particularly useful.

As I noted in the beginning of this chapter, the purpose of this thesis is to better equip societies to reap the benefits of ADMS while managing the associated risks by exploring whether and how EBA can help organisations design and deploy ADMS in ways that align with their organisational values. Having identified my target audiences, it is now possible to articulate two more specific research objectives that answering my RQs will help further.

My first research objective is to sharpen and extend the conceptual toolkit available to organisations who wish to audit the design and use of ADMS not only for legal compliance and technical robustness but also for alignment with specific ethics principles. My second research objective is to provide concrete recommendations for how researchers, auditors, industry practitioners, and policymakers can support the emergence of feasible and effective EBA procedures. Keeping these objectives in mind, I will return to discuss the implications of my research findings for readers from my different target audiences in Chapter 8.

1.11 Concluding remarks

The key message of this introductory chapter can be summarised as follows. Researchers and policymakers alike have pointed towards EBA as a promising governance mechanism for managing the ethical risks ADMS pose. However, despite a growing interest in EBA, many critical gaps remain in the existing literature. Most importantly, EBA has hitherto primarily

been referred as a theoretical proposition. This means that central claims regarding the effectiveness and feasibility of EBA as an ADMS governance mechanism have yet to be substantiated by empirical research.

In this thesis, I address that gap by carrying out problem-driven and empirically grounded research. My thesis thereby adds a perspective hitherto missing from studies of EBA. Specifically, this thesis sets out to explore two overarching research questions:

RQ1 What are the limitations of EBA as a governance mechanism for identifying and mitigating the ethical risks posed by ADMS?

RQ2 How can EBA procedures be designed to effectively identify and mitigate the ethical risks posed by ADMS while being feasible to implement?

In the process of addressing these RQs, the thesis will make original contributions of three kinds. At the conceptual level, it provides *conceptual clarity* about what EBA is, how it works, and what its limitations are. At the descriptive level, it provides new *qualitative knowledge* about how (and why) organisations implement EBA procedures – and what challenges they face in the process. Finally, at the applied level, it provides *actionable recommendations* on how to design EBA procedures that are feasible and effective in practice.

In Chapters 3–7, I will expand on how these contributions help sharpen and extend the conceptual toolkit available to organisations that wish to audit the design and use of ADMS for alignment with ethics principles. To further contextualise these contributions, however, I first review and summarise previous work in the field of ADMS auditing in Chapter 2.

CHAPTER 2

LITERATURE REVIEW

2.1 Synopsis

In this chapter, I review previous literature on auditing of *automated decision-making systems* (ADMS). That literature can, I argue, be divided into technical, legal, and ethics-based approaches. While my doctoral research focuses on *ethics-based auditing* (EBA), this chapter takes a wider remit and reviews the ADMS auditing literature more broadly. The reason for this is twofold. First, because this is an integrated thesis, each substantive chapter will engage with its own relevant literature. Taking a step back to review previous research in related fields thus helps avoid unnecessary overlaps. Second, EBA of ADMS is an inherently multidisciplinary field of research. Reviewing the literature on ADMS auditing more broadly is thus useful since it provides an overview of the different scholarly communities that are active in the discourse on EBA.

As this review will show, the literature on ADMS auditing is at once scarce and rich. It is scarce insofar as ADMS auditing is a relatively recent phenomenon that few researchers have explicitly addressed – much less studied empirically. In fact, much of the relevant literature has only been published during the three years I have worked on this thesis (e.g., Brown et al., 2021; Koshiyama et al., 2022; Raji et al., 2020). Still, the literature on auditing is rich in the sense that it intersects with almost every aspect of how to govern ADMS and relates to many different academic disciplines. To do justice to such a diverse body of research, the literature review presented in this chapter is divided into three parts.

First, in Section 2.2, I survey the evolution of auditing as a governance mechanism, discussing how it has been used to promote transparency and accountability in areas like financial accounting and safety engineering. In doing so, I argue that the track record of audits in these areas contains valuable lessons for how to design procedures to audit ADMS.

Second, in Section 2.3, I draw on recent societal developments to show that the need to audit ADMS results from a confluence of top-down and bottom-up pressures. Specifically, I demonstrate that both technology providers and policymakers have an interest in promoting

auditing as a mechanism for addressing the governance challenges ADMS pose. It is therefore left to academic researchers to study how feasible and effective these auditing procedures are.

Third, in Section 2.4, I review the academic literature in the field. Because auditing of ADMS is inherently multidisciplinary – both as a practice and a field of research – I survey relevant contributions from computer science, systems engineering, law, social science, media and communication studies, political and moral philosophy, and organisational studies. My aim is to showcase the full range of theoretical and methodological approaches different researchers have applied to develop ADMS auditing procedures.

As mentioned in Chapter 1, EBA is a governance mechanism that technology providers employ to demonstrate that the ADMS they design and deploy align with their organisational values. However, the effectiveness and feasibility of different EBA procedures have yet to be systematically explored. In this Chapter, I demonstrate through example that the existing literature in the field has two critical limitations: first, it contains few empirical case studies of the challenges organisations face when implementing EBA and, second, it provides limited guidance on how to design feasible and effective EBA procedures. I conclude this chapter by further contextualising the contributions of my thesis in light of those limitations.

A final methodological remark. Because the purpose of this chapter is to frame my thesis as a whole, I chose to write a *narrative literature review* (Stratton, 2019). Narrative literature reviews are useful since they pull many pieces of information together into a readable format (Grant & Booth, 2009). However, they do not involve a comprehensive search. The risk thereby is that the author's biases influence which studies are included and how these are assessed (Green et al., 2006). While this risk cannot be eliminated, I have taken two measures to minimise it. First, I have drawn on the systematised literature reviews conducted for each of my substantive chapters to ensure that my narration of previous work covers a wide range of perspectives. Second, I have sought input on earlier drafts of this chapter from both industry practitioners and academic researchers from different disciplines.[22] That said, any opinions expressed or omissions in this narrative literature review remain entirely my own.

[22] I am grateful to Luciano Floridi, Varun Rao, Margi Sheth, Josh Cowls, and Marta Ziosi for their comments on earlier drafts of this chapter. Their input has greatly improved it.

2.2 The evolution of auditing as a governance mechanism

The term audit stems etymologically from Latin *auditus*, meaning 'a hearing' (Merriam-Webster, 2023). During Roman times, the term was used with reference to juridical hearings, i.e., official examinations of oral accounts (Lee & Azham, 2008). With time, so-called auditors came to verify written records too. In the broadest sense, auditing thus refers to an independent examination of any entity conducted with a view to expressing an opinion thereon (Gupta, 2004). According to Flint (1988), auditing is a means of social control because it monitors conduct and performance to secure or enforce accountability. Auditing is thus a governance mechanism that various parties can employ to exert influence and achieve normative ends. Over time, the objectives and techniques of auditing have developed, reflecting society's changing needs and expectations (Brown, 1962).

In this section, I briefly review auditing in financial accounting, safety engineering, and social science research, discussing the lessons each has for auditing of ADMS. Before doing so, however, something should be said about why I focus on these three areas. To start with, auditing originated as a means to verify financial accounts (Knapp, 2021). Tracing the role auditing has played in financial accounting is thus useful to highlight its affordances as a governance mechanism. Safety audits and social science audit studies were included in the review since, as this section will show, methodologies and best practices developed in these areas have inspired and informed contemporary attempts to audit ADMS.

2.2.1 Financial audits

The close relationship between auditing and financial accounting is no coincidence. Throughout the Middle Ages, audits were used to verify the honesty of people with fiscal responsibilities (Brown, 1962). During the Renaissance, Italian merchants used auditors to verify the cargo imported from overseas to detect potential fraud. However, the rise of financial auditing – as we know it today – stems from shareholders' need to hold professional managers of large industrial cooperations accountable. In the words of Lee and Azham (2008):

> *'The emergence of a middle class during the industrial revolution period provided the funds for the establishment of large industrial and commercial undertakings. However, the share market during this period was unregulated and highly speculative […] and investors were in dire need of protection. Hence, the time was ripe for the profession of auditing to emerge.'* (Lee & Azham, 2008, p.3)

The modern history of auditing began in 1844, when the British Parliament passed the Joint Stock Companies Act, which required directors to issue audited financial statements to

investors (Smieliauskas & Bewley, 2010). Shortly thereafter, the first public accountant organisations – which certified independent auditors – were formed in the UK.

Another important transition took place in the 1980s with the rise of *risk-based* auditing (Turley & Cooper, 2005). Originally, audits were *compliance-based* in that they sought to verify previously occurring transactions against some pre-established baseline. In contrast, risk-based auditing assessed organisational processes to proactively mitigate risks. Hence, since the 1980s, auditors have not only been expected to enhance the credibility of financial transactions but also provide value-added services like identifying business risks and advising management on how to improve organisational processes (Cosserat, 2004).

In a book titled *The Audit Society*, Power (1997) described the key aspects of financial auditing procedures, three of which have direct implications for the contemporary discourse on how to audit ADMS. First, Power suggested that financial auditing is a 'ritual of verification.' Although auditors examine possible risks and potential fraud, their primary function is to produce comfort. As shown in Chapter 1, ADMS pose many different ethical and social challenges. While it may be impossible to identify and mitigate all risks associated with ADMS, systematised audits can promote trust between actors with competing interests through procedural transparency and regularity.

Second, Power explained that financial auditors do not verify every single transaction but rely on techniques like *sampling*, i.e., selecting a limited number of representative items, to draw conclusions about an entire population. However, the effectiveness of sampling when the audit subject is an autonomous, self-learning system remains unclear. For this reason, New York City's 'AI Audit Law' requires *all* ADMS used for hiring purposes to be audited pre-deployment (Gibson Dunn, 2023). However, Powers' account of financial audits illustrates the trade-off between rigour and feasibility when designing auditing procedures.

Third, Power argued that the auditor-auditee relationship has multiple layers. On the one hand, auditing presupposes operational independence between auditors and auditees. On the other hand, risk-based audits are most effective when the parties collaborate towards a common goal. That tension has created a model called *three lines of defence*; while management, internal auditors, and external auditors should all work to align organisational processes with the interests of different stakeholders, these three actors have complementary roles and responsibilities (IIA, 2009). Recent research suggests that this could also help reduce risks posed by ADMS (Schuett, 2022).

To summarise, financial auditing and accounting has grown into one of the world's largest industries, with an estimated market size of over $110bn (Grand View Research, 2017).

Consequently, the industry is highly professionalised. Many organisations with roots in that industry have utilised their know-how and strong market positions to expand horizontally by offering other auditing services. As a case in point, the Institute of Internal Auditors (IIA, 2018) has recently developed a framework for how to audit ADMS.

2.2.2 Safety audits

Although the modern history of auditing started with financial audits, safety audits represent an equally well-established area of theory and practice. While the former seeks to identify and mitigate financial risks, the latter aims to highlight health and safety hazards and assess the effectiveness of the mechanisms in place to address them (Allford & Carson, 2015). Examples include workplace safety audits (Gay & New, 1999), food safety audits (Dillon & Griffith, 2001), environmental impact and safety audits (De Moor & De Beelde, 2005), vehicle product audits within the automotive industry (Turley et al., 2007), and operation safety audits in the aviation industry (Klinect et al., 2003).

The history of safety audits stretches back to the Industrial Revolution in 19th-century Britain. At that time, the conditions for workers were poor: most people – including children – worked 12–16 hours per day, and the risk of injury or death following workplace accidents was high (Frey, 2019). With time, however, workers formed unions demanding better conditions. One of the mechanisms institutionalised to hold employers accountable was workplace safety audits. Allford and Carson (2015) defined the practice as follows:

> *'Safety audits check that what the business does in reality matches up to both what it says it does [according to its own policies] and what it [legally] should do to continuously ensure that major accident risks are reduced as much as possible.'* (Alford & Carson, 2015, p.1)

Ample evidence shows that systematic and independent audits have improved health and safety for industrial workers (Drudi, 2015). This suggests that the history of safety audits hold valuable lessons for how to design and implement feasible and effective auditing procedures. For example, auditors in this space rely on a plurality of tools (e.g., checklists) and methods (e.g., interviews) to assess the adequacy of organisational safety management systems (Kuusisto, 2001). The lesson that different auditing tools and methods must not be seen as mutually exclusive but rather complementary is something I will return to in Chapter 3.

The history of safety audits also suggests that no audit is stronger than the institutions backing it. Typically, safety audits are conducted by independent auditors, who either belong to or are certified by NGOs like the British Safety Council or government bodies like the US's

Occupational Safety and Health Administration (OSHA). Safety auditors rely on well-recognised standards published by institutions like the International Organisation for Standardisation (ISO) and the International Labour Organisation (ILO) to benchmark the adequacy of employers' health and safety management systems (Bennett, 2002). An equally rigorous institutional ecosystem has yet to emerge for ADMS audits (ICO, 2020).

Further, safety audits highlight the interdependence between technical and social systems. Most accidents involving engineered systems do not stem from the failure of technical components but from requirement flaws or handling errors (Leveson, 2011). Therefore, the main objective of safety audits is to assess and improve organisations' safety cultures. For example, although food safety auditors do sample products to identify toxic substances, they focus on ensuring producers follow best practices for sanitation, transportation, and employee training (Powell et al., 2013). This implies that audits must also consider the culture within organisations designing or deploying ADMS. This insight is something I will return to in both Chapter 4, when analysing my observational data from AstraZeneca's 'AI audit' and in Chapter 7, when proposing a novel blueprint for how to audit large language models.

Despite their merits, safety audits have limitations as a governance mechanism. For example, the history of food safety demonstrate that audits can reduce but never eliminate the risk of incidents occurring (Powell et al., 2013). Moreover, safety auditing may become a box-ticking exercise, which not only wastes resources but can also create a false sense of security that increases the risk of adverse events (Allford & Carson, 2015). Finally, because safety auditors rely on auditees' active cooperation, they often struggle to access the required evidence. As I will argue in Chapter 3, this especially applies to EBA of ADMS, since the access auditors have is limited by IP rights and privacy legislation.

While financial and safety audits differ in substance, they share both procedures and functions. In both cases, auditors seek to verify auditees' claims with the dual aim of reducing risks and providing a basis for holding management accountable. However, as discussed below, the term auditing has been used rather differently in other fields.

2.2.3 Audit studies in the social sciences

In the social sciences, the term 'audit study' refers to a research method, specifically a type of field experiment, which is used to examine individuals' behaviour or the dynamics of social processes (Gaddis, 2018). Field experiments attempt to mimic natural science experiments by implementing a randomised research design in a real-world (as opposed to a lab) setting (Baldassarri & Abascal, 2017). The advantage of field experiments – compared to surveys or

interviews – is that they allow researchers to study people and groups in their natural environment. Gaddis (2018) defined an audit study as follows:

> *'Audit studies [in the social sciences] generally refer to a specific type of field experiment in which a researcher randomizes one or more characteristics about individuals (real or hypothetical) and sends these individuals out into the field to test the effect of those characteristics on some outcome.'* (Gaddis, 2018, p.5)

Thus defined, audit studies have been employed by social scientists since the 1950s, typically to examine difficult-to-detect behaviours, such as racial and gender discrimination. For example, Bertrand and Mullainathan (2004) investigated racial discrimination in hiring across a wide range of sectors by designing an audit study in which they drafted and submitted fictitious résumés in response to job postings. They varied white-sounding and black-sounding names on similar résumés and measured the responses to those applications. Résumés with white-sounding names were 50% more likely to get call-backs from interviewers than those with black-sounding names (Bertrand & Mullainathan, 2004).

Many similar social science audit studies have been conducted. Although sharing a basic methodology, these studies vary in two dimensions. The first is the domain being studied. Beyond recruitment, audit studies have tested for discrimination in areas like access to healthcare (Kugelmass, 2016) and social housing (Ahmed & Hammarstedt, 2008). The second dimension is the choice of independent variable, i.e., the characteristic being manipulated by the researchers. In addition to race, the design of audit studies has included manipulation of gender (Neumark et al., 1996), age (Farber et al., 2017), religion (Pierné, 2013), and physical appearance (Patacchini et al., 2015), just to mention a few examples.

The social science audit study is a suitable methodology for gathering information about discrimination caused by ADMS too. In fact, this is already happening. Several examples of algorithmic audits discussed in Chapter 1 are of this kind, including Boulamwini and Gebru's (2018) study, which demonstrated that ADMS used to classify the images according to gender were significantly more accurate when applied to lighter-skin males than darker-skinned females. In Section 2.4, I will return to the literature on social science audits focusing specifically on ADMS. Here, however, I wish to make a further distinction that aids understanding of the different strands of auditing research.

There are many ways to conduct social science research. For example, there is a long-standing methodological tension between explanation-oriented research seeking to gather empirical evidence on social phenomena and activist research striving to advance a specific normative agenda or change the material conditions of the people and places they study (Hale,

2017). Both approaches have merits and – as the philosophy of social science has shown – are not mutually exclusive but overlap in practice (Cartwright & Montuschi, 2014). However, how researchers relate to their object of study matters, and the field of auditing is no different.

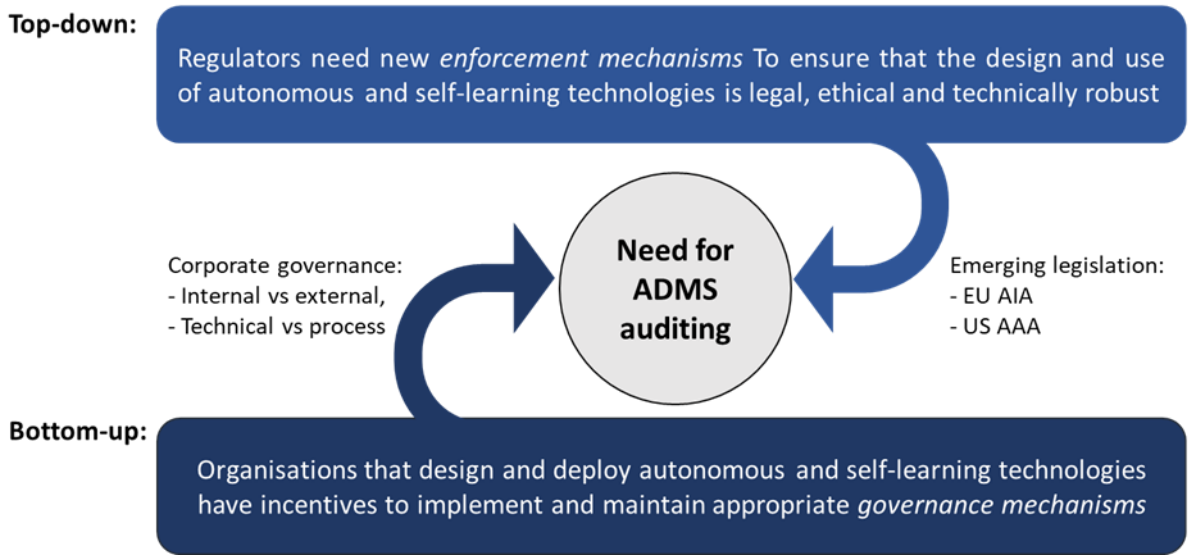
Historically, audit studies in the social sciences have been associated with so-called *activism research*. Cancian (1993) define activism research as research that aims to promote changes that equalise the distribution of resources by exposing inequalities. Audit studies conducted by activist researchers tend to be adversarial in nature, seeking to highlight injustices in ways that spark reactions. In contrast, audits conducted by professional service providers in industry settings aim to produce comfort (Power, 1997). There are thus deep tensions in the motivations different practitioners and researchers have for conducting audits. As the next section will show, these tensions also persist in the literature on ADMS auditing.

2.3 The need for auditing of ADMS: Top-down and bottom-up pressures

Auditing procedures are institutionalised in response to the perceived needs of individuals and groups who seek information or reassurance about the conduct or performance of others in which they have legitimate interests (Flint, 1988). In Section 2.2, that point was illustrated by how financial audits emerged in response to investors' needs and how safety audits were institutionalised in response to social and political pressures to improve working conditions.

In Chapter 1, I showed that the auditing of ADMS is not just a theoretical possibility but already a widespread practice. That sparks two questions: to which perceived needs do these auditing procedures respond? And which stakeholders are seeking information or reassurance through auditing of ADMS? In this section, I argue that the need for auditing of ADMS results from a confluence of top-down and bottom-up pressures. The former includes regulatory mandates and normative expectations placed on technology providers by external stakeholders like policymakers and social advocacy groups. The latter includes voluntary and proactive measures taken by technology providers to stay competitive in their industry, including continuous adaptations of quality management systems (QMS) and brand-building through corporate social responsibility (CSR). Figure 6 illustrates how this confluence of pressures results in a growing need to audit ADMS.

Figure 6. The need to audit ADMS is underpinned by both top-down and bottom-up pressures.



This distinction between top-down and bottom-up pressures is (of course) a simplification. Theoretically, even voluntary actions can be seen as reactions to outside pressures. For example, systems theory in sociology has a long tradition of interpreting all organisational behaviours as autopoietic responses to environmental pressures.[23] However, this line of reasoning generates philosophical questions about the nature of *agency* that are beyond the scope of this thesis.[24] Following the pragmatist stance outlined in Chapter 1, I do not claim that this top-down/bottom-up distinction is true in any ontological sense, but only that it is useful for my analysis. Specifically, it allows me to contrast EBA procedures with auditing procedures designed to test for legal compliance or technical robustness.

With that caveat clear, I now discuss the pressures that have contributed to the emergence of auditing of ADMS as a field of research and practice.

2.3.1 Auditing as a mechanism to implement legislation

ADMS have great potential to contribute to both economic growth and human well-being. By drawing inferences from the growing availability of (big) data, ADMS can improve the speed and accuracy of information processing and contribute to the development of new innovative solutions (Taddeo & Floridi, 2018). However, the ethical, social, and legal challenges ADMS

[23] Autopoiesis refers to the process whereby a system produces and maintains itself over time. See e.g., the works of Niklas Luhmann, in particular *Organization and Decision* (2018).
[24] Disagreement about the nature of agency was at the heart of Niklas Luhmann and Jürgen Habermas' debate in the 1980s (Harste, 2021). As noted by Magee (2016), we have no way of settling that debate empirically.

pose are equally evident. In Chapter 1, I provided several examples of the malfunctioning or misuse of ADMS – including the Dutch welfare benefit scandal and Amazon's discriminatory hiring algorithm. These examples illustrated that ADMS may not only cause harm related to bias, discrimination, and privacy violations but also enable human wrongdoing and undermine self-determination (Tsamados et al., 2021). Policymakers are thus faced with the challenge of balancing the prevention of harm against providing incentives for innovation.

Consider recent developments in the field of *large language models* (LLMs) as an example.[25] The release of ChatGPT has drawn public attention to the capacity of LLMs – such as OpenAI's GPT-3 (Brown et al., 2020) and Google's LaMDA (Thoppilan et al., 2022) – to generate human-like text based on the input provided to them. While such texts are not always semantically meaningful, they can still be used for tasks like text summarisation and translation (Floridi & Chiriatti, 2020). Yet there has been a strong backlash against how LLMs are designed and used. Some researchers have shown that LLMs can produce unethical language, including racist and sexist comments (Kirk et al., 2021). Others have proved that LLMs' answers often contain factual errors (Evans et al., 2021). The seriousness of these limitations is exacerbated by the fact that open-source business models allow LLMs to be used for tasks they were not originally designed to perform (Bommasani et al., 2021). For instance, in January 2023, a Columbian judge used ChatGPT to transcribe his interactions with witnesses, material which he later used to justify his verdict (Parikh et al., 2023). This and other similar examples have understandably sparked widespread public outcry (Kak & West, 2023).

It is important not to be carried away by the latest technological innovation or regulatory trends. Still, the case of LLMs illustrates a more general point, namely, that policymakers are facing increasing pressure to regulate the design and use of ADMS (Smuha, 2021). In many jurisdictions, this has meant drafting new legislation. The European AIA was the first comprehensive regulatory framework for ADMS proposed by any major global economy. In Chapter 6, I will discuss the role of auditing in the AIA in-depth. Here, it suffices to note that the conformity assessments for high-risk systems it mandates constitutes an example on top-down pressures for the institutionalisation of ADMS auditing procedures.

The AIA was published in April 2021. Subsequently, other countries and regions have followed suit. For instance, both Canada and Brazil published draft regulatory frameworks for

[25] LLMs is a subset of ADMS that uses deep learning algorithms trained on a large corpus of data to predict the most likely sequence of words given a specific input or prompt (Kojima et al., 2022). More on this in Chapter 7.

ADMS in 2022, and the US Congress is currently considering the *Algorithmic Accountability Act of 2022* (AAA) (Office of US Senator Ron Wyden, 2022). These draft regulations differ in scope and substance. However, they all stipulate rules and requirements that organisations designing or deploying ADMS must follow. In some cases, the focus is on substantive requirements. For example, ADMS used as components in medical devices must meet specific performance standards in both the EU (Niemiec, 2022) and the US (FDA, 2021). In most cases, however, the focus is on process-based rules (Veale & Borgesius, 2022).

To be successfully implemented, regulations must be linked to effective governance mechanisms (Baldwin & Cave, 1999). For example, the AIA threatens technology providers that fail to comply with its requirements with hefty fines (European Commission, 2021a). However, to determine compliance, one must first consider what mechanisms are available to establish what a provider is doing. This is where auditing comes in. As financial transactions can be audited for correctness, completeness, and legality, so the design and use of ADMS can be audited for technical robustness and legal compliance. Of course, the analogy must not be taken too far;[26] here, it is only intended to highlight how both types of audits respond to one actor's perceived need to gather information about another's conduct. As investors use audits to gather information about managers' conduct and hold them accountable for financial mismanagement or fraud, so policymakers can use audits to gather information about technology providers' conduct and hold them accountable for ADMS that cause harm.

As Chapter 5 will demonstrate, this development is already well underway. The EU AIA, for instance, mandates that high-risk ADMS undergo conformity assessments before deployment. By demanding that these assessments are conducted in a structured manner by independent third parties that have been accredited by national authorities, the European Commission is sketching an EU-wide auditing ecosystem in all but name.

However, there is a major difference between financial audits and legally mandated ADMS audits. Investors exert top-down pressure on managers motivated by the need to manage financial risk. In contrast, policymakers exert pressure on technology providers (in part) to maintain political legitimacy. As noted by Peter (2010), a government's legitimacy hinges partially on its success in solving social and economic problems. As ever more critical tasks become automated, policymakers' political legitimacy will increasingly depend on their

[26] Analogies can sometimes constrain our reasoning by uncritically transferring assumptions from one domain to another (Taddeo, 2016).

abilities to manage the ethical and social challenges ADMS pose. Consequently, the top-down pressure to institutionalise procedures to audit ADMS are likely to continue accumulating.

2.3.2 The role of auditing of ADMS in corporate governance

Private companies play a major role in designing and deploying ADMS (Cihon et al., 2021). Therefore, their design choices have direct and far-reaching implications for important issues, including social justice, economic growth, and public safety (Baum, 2017). However, the dominance of private sector actors holds true not only for the development of commercial applications but also for basic research on the computational techniques that underpin the capabilities of ADMS. For example, in 2018, private companies and labs published over 50% more research papers on ML than academics in the US (Perrault et al., 2019). Hence, the policies and governance mechanisms private companies employ to guide their design and use of ADMS are of profound societal importance.

In the previous section, I showed that policymakers have reasons for mandating audits of ADMS. However, previous research suggests that technology providers too have strong incentives to subject the ADMS they design and deploy to independent audits (Falco et al., 2021; Raji et al., 2020). To understand those incentives, it is useful to first consider the function of corporate ADMS governance, which Mäntymäki et al. (2022) defines as follows:

> *'[ADMS] governance is a system of rules, practices, processes, and technological tools that are employed to ensure that an organization's use of ADMS aligns with the organization's strategies, objectives, and values.'* (Mäntymäki et al., 2022, p.2)

As this definition suggests, corporate governance seeks to ensure that the conduct of an organisation aligns with its stated objectives (OECD, 2015). However, the environment in which corporate governance takes place is inherently dynamic (Arjoon, 2005). As Schumpeter (1942) argued, private companies face constant pressures to innovate and improve their products. Consequently, technology providers have developed mechanisms to ensure that the systems they design and deploy meet predefined quality standards and respond to consumers' needs. Since both the underlying technologies and consumer needs keep changing, the mechanisms employed to govern organisational processes must also be continuously revised.

This brief detour into the function of corporate governance has direct implications for why technology providers voluntarily subject themselves and their ADMS to audits. As noted by Russell et al. (2015), questions concerning corporate ADMS governance are of two kinds: (i) did we build the system right? and (ii) did we build the right system? The former is a

technical question; the latter is a normative one. Audits can provide answers to both kinds of question, as two real-world examples illustrate.

O'Neil (2016) told the story of a woman who, despite a competitive CV, could not get a job due to an error in the algorithmic vetting system used by many recruiters. It was eventually revealed that an alleged criminal offence in her file originated from a data-scraping program, which had conflated her and someone with the same name and postcode. This shows the dangers of negligent design, irresponsible data management, and questionable deployment of ADMS. It is important to note, however, that in this case the data controller, employer, and job seeker would all have benefited from a 'correct' classification. This type of poor-quality outcome constitutes a technical problem that developers, at least in theory, can address. To do so, however, developers need both be made aware of the limitations of the ADMS they design and incentivised to act on that information.

This is where auditing comes in. By assessing the capabilities and limitations of ADMS prior to deployment, auditing helps technology providers identify and mitigate risks before harm occurs (Wilson et al., 2021). Further, by providing a basis on which decision-makers within organisations that design or deploy ADMS can be held accountable, audits incentivise investments in adequate risk management (Shen et al., 2021). In fact, one of the main reasons organisations subject themselves to independent audits is to assess and improve their software development processes and QMS (Vlok, 2003). After all, it is often cheaper to address vulnerabilities early in software development processes. Dawson et al. (2010) estimated that it can cost up to 15 times more to fix a bug in an ADMS when it is found during the testing phase rather than the deployment phase.

In other cases, however, public outcry has been directed not against the technical failures of ADMS but against the purposes for and ways in which they were built in the first place (Keyes et al., 2019). In 2020, Clearview AI – a facial recognition company – faced backlash after investigations revealed that it had scraped billions of images from social media platforms without users' consent to assemble its training dataset (Hill, 2020). Clearview AI suffered significant reputational damage (Smith & Miller, 2022) and faced legal actions culminating in a settlement banning it from selling its technologies to private companies in the US (Robertson, 2022). While it remains unclear whether Clearview AI violated the law, it evidently violated customers' and citizens' normative expectations.

This brings us to the second point: audits focusing on not only technical but also ethical aspects of ADMS help technology providers manage financial and reputational risks (EPRS, 2019). Proactive communication of audit findings may help companies gain competitive

advantages: just as organisations seek to show consumers that their products are healthy through detailed nutritional labels (Holland et al., 2018), the documentation of steps taken to ensure that ADMS are ethical can play a positive role in both marketing and public relations.

To summarise, previous research suggests that structured and independent audits of ADMS can help organisations improve on several business metrics, including regulatory preparedness, data security, talent acquisition, reputational management, and process optimisation (EIU, 2020; Schonander, 2019). In the light of these bottom-up pressures, it is unsurprising that many technology providers have already voluntarily implemented procedures to audit their ADMS for alignment with different sets of ethics principles. AstraZeneca's 'AI audit' (which I describe and discuss in Chapter 4) is an example of an audit conducted voluntarily because of bottom-up pressures.

Yet researchers have also cautioned against this development. Sloane (2021) argued that audits commissioned by technology providers are insufficiently independent, and Bandy (2021) pointed out that, in the absence of agreed standards, technology providers' claims that their ADMS have been audited are hard to verify. These objections should be taken seriously, and in Chapter 3, I will discuss the limitations of EBA procedures in greater depth. However, this section has not sought to assess the merits of ADMS auditing as a governance mechanism but only to highlight that both policymakers and technology providers have an interest in developing and promoting procedures to audit ADMS. The study of how feasible and effective these auditing procedures are in practice is an exercise left to academic researchers.

2.4 Auditing of ADMS: Multidisciplinary foundations

In this section, I review what I refer to as the *ADMS auditing literature*. What unites all works in this body of literature is that they concern procedures to audit ADMS for consistency with relevant specifications, regulations, or ethics principles. To do so, however, I first revisit and expand the definition of auditing of ADMS introduced in Chapter 1.

2.4.1 The ADMS auditing literature

Auditing of ADMS is a governance mechanism that can be wielded by different actors in society in pursuit of different goals and objectives. It can be used by regulators to assess whether a specific ADMS meets legal standards, by technology providers to mitigate technology-related risks, or by other stakeholders to make informed decisions about how they engage with specific companies (Brown et al., 2021). Operationally, auditing of ADMS is

characterised by a structured process whereby an entity's past or present behaviour is assessed for consistency with predefined standards, regulations, or norms.

Three aspects of this definition demand further elaboration. First, the subject of the audit can be either a person, an organisation, a technical system, or any combination thereof. Second, whether conducted by an *external* third party or an *internal* audit function, auditing requires operational independence between the auditor and the auditee (Power, 1997). Third, auditing requires a predefined baseline to serve as a basis for evaluation (ICO, 2020). However, the nature of this baseline can vary between hard regulations, organisational values and policies, or technical standards and benchmarks.

Previous work on ADMS auditing constitutes a heterogeneous and multidisciplinary body of literature. It is heterogeneous in that it encompasses contributions from a diverse range of actors employing different methods and facing competing incentives. The ADMS auditing literature includes academic articles and books (Berghout et al., 2023; Metaxa et al., 2021; Mittelstadt, 2016), auditing tools and procedures developed by private companies (Babl AI, 2023; ORCAA, 2020; PwC, 2020), standards published by industry associations and professional standard-setting bodies (IEEE SA, 2020; ISO, 2022; NIST, 2022; VDE, 2022), and draft legislation and guidance documents issued by policymakers (EPRS, 2022; European Commission, 2021a; ICO, 2020), to mention just a few prominent examples.

The ADMS auditing literature is also multidisciplinary in that it harbours contributions from many academic disciplines, including computer science (Adler et al., 2018; Kearns et al., 2018), systems engineering (Dennis et al., 2016; Leveson, 2011), law (Laux et al., 2021; Selbst, 2021), media and communication studies (Bandy & Diakopoulos, 2019; Sandvig et al., 2014), social science (Metaxa et al., 2021; Vecchione et al., 2021), philosophy (Boddington, 2017; Dafoe, 2017), and organisational studies (Guszcza et al., 2018; Minkkinen et al., 2022).

Such a diverse body of literature can be sliced and diced in many ways. In what follows, I provide an overview of the ADMS literature in three steps. First, I distinguish between narrow and broad conceptions of auditing. Second, I distinguish between technical, legal, and ethical approaches to ADMS auditing. Finally, I distinguish between strands of research that (i) propose, (ii), develop, (iii) employ, or (iv) critique ADMS auditing procedures.[27]

[27] This taxonomy contains overlapping categories. A singe research contribution can, for example, use a *narrow conception* of auditing to *develop* a procedure for assessing whether an ADMS is *legally* compliant.

2.4.2 Narrow vs broad conceptions of auditing of ADMS

To start with, it is useful to distinguish between *narrow* and *broad* conceptions of ADMS auditing. The former is impact-oriented, focusing on probing and assessing the output of ADMS for different input data. The latter is process-oriented, focusing on assessing the adequacy of the software development processes and QMS technology providers employ.

In their book *Auditing algorithms: understanding algorithmic systems from the outside in*, Metaxa et al. (2021) provided an example of a narrow definition of auditing:

> *'[an algorithm audit is] a method of repeatedly and systematically querying an algorithm with inputs and observing the corresponding outputs in order to draw inferences to its opaque inner workings.'* (Metaxa et al., 2021, p.18)

Narrow conceptions of auditing are well suited to gathering evidence about unlawful discrimination and tend to be underpinned by experimental designs. For example, in an article titled *Algorithm auditing at large-scale: insights from search engine audits*, Ulloa et al. (2019) designed virtual agents to perform systematic experiments simulating human interactions with search engines. The authors demonstrated that such an audit design can be employed to monitor an ADMS's output over time and flag potential ethical concerns such as disparate treatment.

In contrast, broad conceptions of auditing focus not so much on the properties of ADMS as the governance structures of the organisations that design and deploy them. This practice has deep roots in conventional IT audits Zinda (2021) and technology risk management procedures (Senft & Gallegos, 2009). For example, when describing the role of such an auditor, (Jager and Westhoek (2023) wrote:

> *'It is not just about checking the algorithm itself and the management measures surrounding it, but also paying attention to the data used, the methods used in the development and the optimization of the algorithm. These aspects of management, process, and content should also be part of the assessment framework and thus the audit approach.'* (Jager & Westhoek, 2003, p.145)

Broad conceptions of auditing are useful since they allow researchers not only to detect illegal, erroneous, or unethical behaviours of ADMS but also to investigate the sources of such behaviours. For example, discriminatory behaviour of ADMS may be caused by incomplete or unrepresentative training datasets (Gehman et al., 2020) or inadequate ADMS testing and validation procedures (Myllyaho et al., 2021). For this reason, researchers like Koshiyama et al. (2022) have proposed procedures for auditing the entire process whereby ADMS are designed and deployed. Typically, this entails assessing the governance structures technology

providers have in place to train their staff, assemble training datasets, evaluate the limitations of ADMS prior to deployment, and monitor the behaviour of ADMS over their entire lifetime.

Both narrow and broad conceptions of auditing have generated flourishing strands of research. Some researchers have leveraged narrow conceptions of auditing to test for bias and discrimination in online ad delivery (Ali et al., 2019; Sweeney, 2013) and autocomplete algorithms (Robertson et al., 2018), for fairness in image classification systems (Morina et al., 2019), for accuracy in news curation systems (Bandy & Diakopoulos, 2019), for completeness in datasets (Coston et al., 2021; Sookhak et al., 2014), and for data privacy, e.g., how easy it is to reconstruct training data from ADMS (Kolhar et al., 2017; Narula et al., 2018).

Other researchers have leveraged broad conceptions of auditing to study how ADMS are being designed and the adequacy of technology providers' governance mechanisms. Ugwudike, (2021) studied how ADMS used for predictive policing are designed and deployed; Jager and Westhoek (2023) studied technology providers' mechanisms for testing image recognition algorithms; Mahajan et al. (2020) provided a framework for how auditors and vendors can collaborate to validate ADMS used in radiology; and Dash et al. (2019) demonstrated how audits of recommender systems can provide insights into how these systems affect users and societies over time.

For my purposes, this discussion has two key takeaways. First, narrow and broad conceptions of auditing have different affordances. The former allows researchers to audit the behaviour of ADMS without approval from, or the cooperation of, technology providers (Adler et al., 2018; Lee, 2021; Lurie & Mustafaraj, 2019). The latter enables researchers to study the real-world effects different auditing procedures have on how ADMS are designed and deployed (Ayling & Chapman, 2021; Fitzgerald et al., 2013; Stoel et al., 2012). Second, there is no contradiction between the two conceptions. In fact, they are both compatible and mutually reinforcing. As I will argue in Chapter 3, narrow testing of ADMS based on input-output relationships can (and should) be integrated into broader auditing procedures.

2.4.3 Technical, legal, and ethics-based approaches

In addition to having different methodological conceptions of what auditing *is*, researchers also differ in what they are auditing ADMS *for*. Per definition, auditing requires a predefined baseline against which the audit's subject can be evaluated (ICO, 2020). However, depending on the audit's purpose, this baseline can consist either of technical specifications, legal requirements, or voluntary ethics principles. Consequently, contributions to the ADMS auditing literature can be categorised into technical, legal, and ethical approaches.

Technical approaches refer to auditing procedures designed to quantify and assess the technical properties of ADMS, including accuracy, robustness, and safety. These build on tools and methods with proven track records in systems engineering and computer science, including model evaluation (Parker, 2020) and system verification (Luckcuck et al., 2019; Thudi et al., 2021). Within the realm of technical approaches, a distinction is often made between *ex-ante* and *ex-post* audits (Etzioni & Etzioni, 2016). The former evaluates an ADMS prior to its market deployment, the latter monitors its performance over time as it interacts with new input data in real-world environments (Minkkinen et al., 2022).

The idea of auditing software dates back several decades (see e.g., Hansen & Messier, 1986; Weiss, 1980). Still, the academic literature in this field has grown rapidly in recent years. Some research groups have developed open-source toolkits allowing technology providers to test and evaluate the performance of ADMS on different tasks and datasets (Cabrera et al., 2019; Saleiro et al., 2018). Others have developed auditing procedures for more targeted purposes, e.g., to test the accuracy of personality prediction in ADMS used for recruitment (Rhea et al., 2022), evaluating the capabilities of language models (Goel et al., 2021), providing explanations for black-box ADMS (Pedreschi et al., 2018), and conducting audits of clinical decision support systems (Panigutti et al., 2021). Again, what links all these procedures is that they audit ADMS against predefined technical, functionality, and reliability standards.

In contrast, *legal approaches* refer to auditing procedures that assess whether the design and use of ADMS comply with relevant regulations. Such procedures rely on different legal provisions, including those stipulated in: *data privacy regulations* like the GDPR (European Parliament, 2016); *discrimination laws* like the US's 1964 Civil Rights Act or Equal Credit Opportunity Act of 1974 (Barocas & Selbst, 2016); *sector-specific certification mandates*, as is the case for medical device software (FDA, 2021); or *general transparency obligations*, as found in the AIA (European Commission, 2021a).

Legal scholars have debated about when and how the above-listed regulations apply to ADMS (Durante & Floridi, 2021; Edwards & Veale, 2018; Pentland, 2020; Wachter et al., 2017). A review of legal scholarship on ADMS falls outside the scope of this thesis.[28] Nevertheless, I wish to highlight that a wide range of procedures to audit ADMS for legal compliance have already been proposed and, in some cases, implemented (Merrer et al., 2022).

[28] As mentioned in the Section 1.6, any legal analysis falls outside the scope of this thesis. Readers interested in the legal challenges posed by ADMS are referred to the overview provided by Barfield and Pagallo, 2018).

For instance, Mikians et al. (2012) developed a procedure to audit ADMS for unlawful price discrimination based on protected attributes. Similarly, Silva et al. (2020) audited Facebook's ad delivery algorithm, finding that it violated political advertising laws.

Finally, *ethics-based approaches* refer to auditing procedures for which voluntary ethics principles serve as the normative baseline. EBA can be either collaborative or adversarial. In the former case, audits are conducted in collaboration with technology providers to assess whether their ADMS adhere to predefined ethics principles (Berghout et al., 2023; Raji et al., 2020).[29] In the latter case, independent actors conduct the audits to assess an ADMS without access to its source code (Metaxa et al., 2021; Sandvig et al., 2014).[30] Collaborative audits aim to provide assurance, adversarial audits to expose harms. In both cases, however, EBA concerns what ought to be done over and above compliance with existing regulations.

During EBA procedures, ADMS are audited against either a technology provider's organisational values or ethics principles proposed by institutions like the IEEE (2019), OECD (2019), and the AI HLEG (2019). While these guidance documents vary in language (Jobin et al., 2019), they converge on a limited set of principles (Floridi & Cowls, 2019). Reflecting this convergence, previous research has developed procedures to audit ADMS for *transparency* and *explainability* (Cobbe et al., 2021; Mittelstadt, 2016), *bias* and *fairness* (Bartley et al., 2021; Morina et al., 2019), and *accountability* (Busuioc, 2021; Metcalf et al., 2021).

The boundaries between technical, legal, and ethics-based audits are often blurry in practice. Legal compliance audits typically rely on technical methods to gather evidence about the properties and impact ADMS have (Kim, 2017; Merrer et al., 2022). Similarly, technical robustness and legal compliance are often prerequisites for considering an ADMS ethical (Keyes et al., 2019). The three audit types are thus best viewed as a continuum of complementary approaches with different focal points.

That said, the distinction between technical, legal, and ethical approaches is useful here for two reasons. First, it mirrors the vocabulary adopted by policymakers. For example, AI HLEG, (2019) stipulated that ADMS should be lawful, ethical, and technically robust. Adopting this well-established vocabulary facilitates communication with my target audiences.

---

[29] Collaborative audits tend to be process-oriented (Kazim & Koshiyama, 2020), leveraging methods that require access to technology providers' internal processes, including ethical foresight analysis (Floridi & Strait, 2020) and algorithm design (Kearns & Roth, 2020).

[30] Adversarial audits tend to be impact-oriented (Jaiswal et al., 2022), leveraging both quantitative methods, like red-teaming (Perez et al., 2022), and qualitative methods, like ethical impact assessments (Reisman et al., 2018).

Second, the distinction helps demarcate the scope of my research – which focuses on EBA. That said, procedures to audit ADMS for technical robustness and legal compliance have a longer history and have been more widely implemented (Vecchione et al., 2021). Throughout this thesis, I will therefore continue referring to technical and legal approaches with the aim of identifying transferable lessons for how to design feasible and effective EBA procedures.

2.4.4 Who audits the auditors?

Contributions to the academic literature on ADMS auditing relate to the object of study in different ways. For example, distinctions can be made between contributions that (i) provide theoretical justifications for why audits are needed, (ii) develop procedures, tools, or methods to audit ADMS, (iii) employ available auditing procedures, tools, or methods, and (iv) study the effectiveness and feasibility of auditing ADMS as a governance mechanism. In what follows, I briefly review these different research strands.

To start with, there is a significant body of literature calling for ADMS to be audited (see e.g., Brown et al., 2021; Diakopoulos, 2015; Kim, 2017; Sandu et al., 2022; Sandvig et al., 2014). These contributions stress both the social, ethical, and legal risks ADMS pose and how audits can help identify and manage those risks. For example, research has suggested that auditing contributes to good governance through procedural regularity and transparency (Floridi, 2017b; Larsson, 2020; Loi et al., 2020) and prevents harm by ensuring proactivity in the design of ADMS (Kazim & Koshiyama, 2020). Such contributions are often commentary or viewpoint articles (see e.g., Falco et al., 2021; Guszcza et al., 2018; Kassir et al., 2022). The main argument advanced by this literature is that structured and independent audits constitute a pragmatic approach to managing the governance challenges of ADMS.

Responding to these calls, other researchers have developed tangible ADMS auditing procedures and tools. Such contributions can be divided into two broad categories. First, high-level procedures – often proposed by scholars from organisation studies or systems engineering – that outline the steps audits should include, what activities these entail, and the roles and responsibilities of different stakeholders (Felländer et al., 2022; Floridi et al., 2022; Zicari et al., 2021). Second, researchers have developed tools that can be employed by auditors for specific tasks, including detecting bias in ADMS (Saleiro et al., 2018; Sokol et al., 2022), documenting how ADMS are designed (Gebru et al., 2021; Mitchell et al., 2019), and simulating or monitoring their behaviour in real-world settings (Akpinar et al., 2022). These tools are typically developed by computer scientists or computational social scientists.

Yet other researchers employ existing auditing procedures and tools to conduct empirical studies (Aragona, 2022), including qualitative studies that assess how ADMS are designed (Christin, 2020; Marda & Narayan, 2021; Seaver, 2017) and quantitative audit studies that measure the properties of ADMS or their impact on users and societies (Abebe et al., 2019; Metaxa et al., 2021; Speicher et al., 2018). Contributions to this literature have been made by researchers from different fields. For example, labour economist Songül Tolan (2019) audited ADMS used by courts to predict criminal recidivism and found they discriminate against male defendants and people of specific nationalities, and a team of computer scientists led by Alicia DeVos et al. (2022) conducted user-centric audits to study ADMS, concluding that users were able to identify harmful behaviours that formal testing processes had not detected.

Finally, a small but growing community of researchers are interested in how feasible and effective auditing is as an ADMS governance mechanism (Costanza-Chock et al., 2022; Landers & Behrend, 2022). So far, such research has been dominated by theoretical critiques.[31] For example, Sloane (2021) argued that current auditing procedures are toothless and may even be counterproductive insofar as they legitimise the deployment of potentially harmful ADMS. To avoid that trap, Sloane (2021) suggested that standards for how to audit ADMS are urgently needed. Similarly, Engler (2021) argued that independent auditors struggle to hold technology providers accountable because – in the absence of sector-specific legislation – they can simply refuse access to their data and models. These important objections call for further inquiry. As of now, however, claims about the limitations of ADMS auditing as a governance mechanism have yet to be substantiated by empirical research (just as claims about its affordances).

My doctoral research relates to the four strands of research discussed above in several ways. As noted in Chapter 1, my research has conceptual, descriptive, and applied components. At the conceptual level, this thesis contributes *conceptual clarity* about what EBA is, how it works, and what its limitations are (Chapter 3). It thus builds on the literature that provides theoretical justifications for EBA as an ADMS governance mechanism and the discourse about what we can reasonably expect EBA to achieve. At the descriptive level, this thesis provides *qualitative knowledge* about the organisational contexts in which EBA must be integrated to be feasible and effective. In particular, my case study of AstraZeneca's EBA (Chapter 4) provides new empirical evidence about the challenges organisations face when implementing

[31] One exemption is Hasan et al. (2022) who, drawing on their own experience as auditors, have published generalisable lessons on how to audit ADMS in practice.

EBA, thus adding a perspective hitherto missing in the literature. Finally, at the applied level, this thesis provides *actionable recommendations* on how to design EBA procedures that are feasible and effective (Chapters 5, 6, and 7). Consequently, this thesis also adds to the literature that seeks to expand and sharpen the methodological toolbox available to ADMS auditors.

Other parts of the reviewed literature my thesis only touches upon tangentially. For example, I neither use auditing as a social science method to study the impact ADMS have on individual users, nor do I propose any new statistical techniques for evaluating ADMS during technical audits. While both are important research strands, they lie outside this thesis's scope. Instead, this thesis seeks to investigate the conditions under which EBA can be a feasible and effective governance mechanism for managing some of the ethical challenges ADMS pose.

2.5 Concluding remarks

This chapter aimed to showcase the multidisciplinary roots of ADMS auditing and provide an overview of the scholarly communities active in the discourse. So, what have we learned?

In Section 2.2, I showed that different stakeholders (like investors) have long used auditing as a mechanism for holding managers accountable for financial fraud or negligence (Lee & Azham, 2008). I also illustrated how auditing has evolved as a governance mechanism, from 'merely' verifying accounts to proactively identifying and managing risks (Turley & Cooper, 2005). Finally, I highlighted that – in addition to contributing to good governance – auditing is often employed to produce a sense of comfort (Power, 1997).

This historically grounded understanding of auditing enables us to better put contemporary calls for ADMS to be audited into perspective. The use of ADMS is associated with a wide range of ethical risks. Due to their autonomy, complexity, and adaptability, ADMS also pose significant governance challenges (Cath et al., 2018; Russell et al., 2015). This is a problem, both for technology providers seeking to ensure that their ADMS are ethical, legal, and safe (Dignum, 2017) and for governments facing increasing pressure to regulate the design and use of ADMS (Minkkinen et al., 2021; Smuha, 2021). Bringing all this together, I argued in Section 2.3 that both technology providers and policymakers have an interest in promoting auditing as a governance mechanism to manage the ethical challenges ADMS pose.

In Section 2.4, I reviewed previous academic literature on ADMS auditing. While showing that this literature harbours several distinct strands of research, I also demonstrated that it contains various critical gaps. For example, scholars have argued that ADMS should be audited for adherence to predefined ethics principles (Sandvig et al., 2014; Kim, 2017; Brown

et al., 2021; Falco et al., 2021). However, claims about the affordances and limitations of EBA as an ADMS governance mechanism have yet to be substantiated by empirical research. Further, both researchers and private companies have developed EBA procedures or offer EBA services to help technology providers design and use ADMS in ways that align with their organisational values (Raji et al., 2020; Zicari et al., 2021; Felländer et al., 2022). But again, the feasibility and effectiveness of those EBA procedures remain unclear.

Taken together, this literature review holds two key takeaways. The first is that my RQs address not only pressing social and practical problems but also distinct knowledge gap in the academic literature. Recall the two RQs that guide my research in this thesis:

RQ1 What are the limitations of EBA as a governance mechanism for identifying and mitigating the ethical risks posed by ADMS?

RQ2 How can EBA procedures be designed to effectively identify and mitigate the ethical risks posed by ADMS while being feasible to implement?

When formulating these RQs in Chapter 1, I stressed that the purpose of my research is to better equip societies to reap the benefits of ADMS while managing the associated risks by exploring whether and how EBA can help organisations design and deploy ADMS in ways that align with their organisational values. In this chapter, my review of previous work on EBA has demonstrated that my RQs are not only socially important but also academically relevant.

The second key takeaway from my review of previous work is that ADMS auditing is an inherently multidisciplinary field, incorporating insights from computer science, systems engineering, law, social science, media and communication studies, philosophy, and organisational studies. This thesis reflects that variety in several ways, from the choice of research topics, via the research methods employed, to the journals in which its findings have been published.[32] Given the multidisciplinary nature of my work – and that I approach my RQs on conceptual, descriptive, and applied levels – I anticipate that different elements of this thesis will be of interest to different target audiences.

[32] The article on which Chapter 3 is based was published in *Science and Engineering Ethics*, a journal aimed at computer scientists. The articles on which Chapters 4 and 7 were published in *AI and Ethics*, a multidisciplinary journal encouraging dialogue between academics, policymakers, and industry practitioners on how to govern ADMS. The articles on which Chapters 5 and 6 are based appeared in *Minds and Machines*, a philosophy journal.

In Chapter 1, I specified that my target audience includes: *academic researchers* who are interested in exploring the merits and limitations of EBA of ADMS as a governance mechanism; *auditors* who develop and offer EBA services to technology providers; *industry practitioners* who implement EBA procedures in organisations that design and deploy ADMS; and *policymakers* who draft legislation and guidance on how to govern ADMS.

Keeping those target audiences in mind, the following five chapters – i.e., the core of my research – explore my RQs at three different levels: conceptual, descriptive, and applied. Chapter 3, to which I turn next, starts at the conceptual level by exploring what EBA is, or at least should be, in the context of ADMS governance.

CHAPTER 3

ETHICS-BASED AUDITING OF AUTOMATED DECISION-MAKING SYSTEMS: NATURE, SCOPE, AND LIMITATIONS

Abstract

Previous research has pointed towards *ethics-based auditing* (EBA) as a promising governance mechanism for managing the ethical risks posed by *automated decision-making systems* (ADMS). However, while both researchers and policymakers have called for ADMS and technology providers to be audited, the affordances and constraints of EBA as a governance mechanism have yet to be systematically explored. In this chapter, I conduct a systematised literature review to address that gap. Building on previous work, I define EBA as a structured process whereby an entity's behaviour is assessed for consistency with relevant principles or norms. I then offer three contributions to the existing literature through theory synthesis. First, I provide a theoretical explanation of how EBA contributes to good governance by promoting procedural regularity and transparency. Second, I derive seven criteria for how to design and implement EBA procedures successfully. Third, I demonstrate that existing EBA procedures are subject to a wide range of conceptual, technical, social, economic, organisational, and institutional constraints, and I provide a novel taxonomy to help researchers understand and account for these. By articulating what EBA *is*, *how* it works, and *what* it can (and cannot) be reasonably expected to achieve, this chapter provides the conceptual foundation for the thesis as a whole.

Note

This chapter is based on a peer-reviewed journal article published in *Science and Engineering Ethics* (see Mökander et al., 2021).[33] While I have sought to minimise changes, this chapter differs from the original article in two ways. First, I have updated the bibliography to reflect recent technological and societal developments. Second, I have revised the text to ensure greater coherence with other chapters included in this thesis.

[33] The original article was co-authored with Jessica Morley, Mariarosaria Taddeo, and Luciano Floridi. Please see Appendix 1, 2 and 3 for authorship statements.

3.1 Introduction

3.1.1 Background

Automated decision-making systems (ADMS), i.e., autonomous self-learning systems that gather and process data to make qualitative judgements with little or no human intervention, increasingly permeate all aspects of society (AlgorithmWatch, 2019). This means that many decisions – which were previously made by human experts – are now made by ADMS (Krafft et al., 2020a). Already today, ADMS are used to inform decisions in areas like recruitment (Gupte & Mishra, 2023), medical diagnostics (Grote & Berens, 2020), and lending (Aggarwal 2019). As information societies mature, the range of decisions that can be automated in this fashion will increase, and ADMS will be used to make ever-more critical decisions.

From a technical perspective, the computational techniques used by ADMS vary from decision trees to deep neural networks (Lepri et al., 2018). However, my focus in this thesis is not on the underlying technologies but rather on the common features of ADMS from which ethical challenges arise. It is the combination of relative *autonomy*, *complexity*, and *adaptability* that underpins both beneficial and problematic uses of ADMS. Delegating tasks to ADMS can help increase consistency, improve efficiency, and enable new solutions to complex problems (Taddeo & Floridi, 2018). Yet these improvements are coupled with ethical challenges. As noted already by Norbert Wiener:

> *'The machine, which can learn and can make decisions on the basis of its learning, will in no way be obliged to make such decisions as we should have made, or will be acceptable to us.'* (Wiener, 1954, p.212)

Specifically, ADMS may leave decision subjects vulnerable to the harms associated with poor-quality outcomes, bias and discrimination, and invasion of privacy (Leslie, 2019). More generally, ADMS risk enabling human wrongdoing, reducing human control, devaluing human skills, and eroding human self-determination (Tsamados et al., 2020). If these ethical challenges are not sufficiently addressed, a lack of public trust in ADMS may hamper the adoption of such systems which, in turn, would lead to significant social opportunity costs through the underuse of available and well-designed technologies (Cookson, 2018). Addressing the ethical challenges posed by ADMS is therefore becoming a prerequisite for good governance in information societies (Cath et al., 2018).

However, traditional governance mechanisms designed to oversee human decision-making processes often fail when applied to ADMS (Kroll et al., 2016). One important reason for this is that the delegation of tasks to ADMS curtails the sphere of ethical deliberation in

decision-making processes (D'Agostino & Durante, 2018). In practice, this means that norms that used to be open for interpretation by human decision-makers are now embodied in ADMS. From an ethical perspective, this shifts the focus of ethical deliberation from specific decision-making situations to the ways in which ADMS are designed and deployed.

In response to the growing need to design and deploy ADMS in ways that are ethical, over 75 organisations – including governments, companies, academic institutions, and NGOs – have produced documents defining high-level guidelines (Jobin et al., 2019). Reputable contributions include *Ethically Aligned Design* (IEEE, 2019), the *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019), and the *Recommendation of the Council on Artificial Intelligence* (OECD, 2019). These guidelines converge around five principles: beneficence, non-maleficence, autonomy, justice, and explicability (Floridi & Cowls, 2019).

While a useful starting point, these principles tend to generate interpretations that are either too semantically strict, which are likely to make ADMS overly mechanical, or too flexible to provide practical guidance (Arvan, 2018). This indeterminacy hinders the translation of ethics principles into practices and leaves room for unethical behaviours like 'ethics shopping', i.e., mixing and matching ethical principles from different sources to justify some pre-existing behaviour; 'ethics bluewashing', i.e., making unsubstantiated claims about ADMS to appear more ethical than one is; and 'ethics lobbying', i.e., exploiting ethics to delay or avoid good and necessary legislation (Floridi, 2019b).

Moreover, the adoption of ethics guidelines remains voluntary, and the industry lacks both incentives and useful tools to translate principles into verifiable criteria (Raji et al., 2020). For example, interviews with software developers indicate that while they consider ethics important in principle, they also view it as an impractical construct that is distant from the issues they face in daily work (Vakkuri et al., 2019). Further, even organisations that are acutely aware of the risks posed by ADMS may struggle to manage these, either due to a lack of useful governance mechanisms or conflicting interests (PwC, 2019). Taken together, there still exists a gap between the 'what' (and 'why') of ethics principles and the 'how' of designing, deploying, and governing ADMS in practice (Morley et al., 2020a).

A vast range of governance mechanisms that aim to support the translation of high-level ethics principles into practical guidance has already been proposed. Some of these focus on interventions in the early stages of software development processes, e.g., by raising awareness of ethical issues among software developers (Floridi et al., 2018), creating more diverse teams of software developers (Sánchez-Monedero et al., 2020), embedding ethical values into technological artefacts through proactive design (Aizenberg & van den Hoven,

2020), screening potentially biased input data (AIEIG, 2020), or verifying the underlying decision-making models and code (Dennis et al., 2016). Other proposed governance mechanisms focus on the context in which ADMS operate. For example, so-called human-in-the-loop protocols imply that human operators can either intervene to prevent or be held responsible for harmful system outputs (Jotterand & Bosco, 2020; Rahwan, 2018).

While these examples do not provide a systematic account of the array of available governance mechanisms, they illustrate the point that researchers, technology providers, and policymakers are actively looking for and experimenting with new ways to address the governance challenges ADMS pose.

3.1.2 Research topic and research gap

Against this backdrop, several recent academic articles and policy proposals have pointed towards *ethics-based auditing* (EBA) as a promising governance mechanism to identify and mitigate the ethical risks posed by ADMS (Brown et al., 2021; Raji et al., 2020; Sandvig et al., 2014). As I explained in Chapter 1, EBA is characterised by a structured process whereby an entity's behaviour is assessed for consistency with predefined ethics principles.

Of course, the idea of auditing software is not new. Since the 1970s, computer scientists have audited software systems to check whether they adhere to relevant technical specifications (Weiss, 1980). However, in 2014, Sandvig et al. published a much-cited article titled 'Auditing Algorithms', in which they argued that ADMS should be audited not only for legal compliance and technical robustness but also for alignment with *ethics principles*.

EBA has since attracted much attention from policymakers, researchers, and industry practitioners alike. Policymakers like the UK Information Commissioner's Office (ICO) have drafted EBA procedures (ICO, 2020); academic researchers have developed tools and procedures to audit ADMS for their adherence to ethical principles like fairness, accountability, and transparency (Mittelstadt, 2016; Raji et al., 2020; Wilson et al., 2021). In parallel, a new industry is emerging, in which traditional accounting firms like PwC (2019) and Deloitte (2020), startups like ORCAA (2020), and NGOs like ForHumanity (2021) all offer EBA services to help technology providers verify claims about the ADMS they design and deploy.

These and other early works have made important contributions to the theory and practice of EBA. However, as I demonstrated in Chapter 2, there remains a discrepancy between the attention EBA has attracted and the fact that its merits and limitations as an ADMS governance mechanism have yet to be systematically explored (Landers & Behrend, 2022). The purpose of my thesis is thus to better equip societies to reap the benefits of ADMS while

managing the associated risks by exploring whether and how EBA can help organisations design and deploy ADMS in ways that align with their organisational values.

Pursuing that purpose, I seek to explore what the limitations of EBA as a governance mechanism are (RQ1) and, given those limitations, how feasible and effective EBA procedures can be designed (RQ2). Over the course of this thesis, I will approach these RQs at three different levels: conceptual, descriptive, and applied. In this chapter, however, I focus entirely on the conceptual level, which is concerned with what EBA is and how it works.

3.1.3 Research questions, methodology, and limitations

In this chapter, I address SQ1: What are the affordances and constraints of EBA as a governance mechanism to address the ethical risks posed by ADMS? To address this question, I proceeded in two steps. First, I conducted a *systematised literature review* (Grant & Booth, 2009) to define core concepts and structure the findings of previous work. To identify relevant literature for the review, I broke down SQ1 into three more specific questions:

- *What EBA tools and procedures have already been developed or proposed in the existing literature?*
- *How do researchers and policymakers advocating for EBA envision that it can contribute to identifying and mitigating the ethical risks posed by ADMS?*
- *What constraints of EBA as an ADMS governance mechanism have been discussed in the existing literature?*

The collection phase involved searching five databases (Google Scholar, Scopus, SSRN, Web of Science, and arXiv) for articles related to EBA of ADMS. Keywords for the search included (‘auditing, ‘evaluation’, OR ‘assessment’) AND (‘ethics’, ‘fairness’, transparency’, OR ‘robust) AND (‘automated decision-making’, ‘artificial intelligence’, OR ‘algorithms’). To limit the scope of the literature review, I focused on articles published after 2011, the year when IBM Watson marked the coming of the second wave of AI by beating the two best-ever humans to have competed in the TV quiz show Jeopardy (Susskind & Susskind, 2015). In total, 122 articles and reports were included in the systematised literature review.

In the second step, I synthesised the findings and theories found in previous research to achieve an improved understanding of what EBA *is*, or at least *ought to be*, in the context of ADMS. To do so, I followed an analytic technique called *theory synthesis* (Jaakkola, 2020). The aim of theory synthesis is to offer a new or enhanced view of a phenomenon by linking

previously unconnected arguments in a novel way. The methodology is particularly useful for identifying and underscoring commonalities and building coherence across fragmented bodies of literature and theory (Corpanzano, 2009).

According to Pund and Campbell (2015), theory synthesis includes three stages: (i) *synthesis preparation*, wherein relevant theories are extracted and summarised; (ii) *synthesis*, which involves comparing theories for points of convergence and divergence; and (iii) *synthesis refinement*, whereby the synthesis is critically interrogated to generate further theoretical insights. Following this methodology, I first used a predefined set of questions to extract information from all articles and reports found in my systematised literature review.[34] Subsequently, I identified all different claims made about EBA in the previous literature, clustered these based on shared assumptions, and assessed the underlying assumptions for soundness and consistency. Finally, I synthesised the theoretical affordances and constraints of EBA by articulating them in a mutually exclusive and collectively exhaustive way.

Building on the findings from my systematised literature review and theory synthesis, I offer three contributions in this review chapter. First, I provide a theoretical explanation of how EBA contribute to good governance by promoting procedural regularity and transparency. Second, I propose seven criteria for successfully designing and implementing EBA procedures. Third, I demonstrate that existing EBA procedures are subject to a wide range of constraints, and I provide a novel taxonomy to help researchers understand and account for these. Taken together, this chapter helps advance my larger research objective of sharpening and expanding the conceptual toolkit available to organisations that wish to audit the design and use of ADMS.

Two limitations help narrow down the scope of this chapter. First, I do not address any legal aspects of auditing. Rather, my focus in this chapter is on ethical alignment, i.e., on what ought and ought not to be done over and above compliance with existing regulations. Second, any review of normative ethics frameworks remains outside the scope of this chapter. When designing and operating ADMS, tensions may arise between different ethical principles for which there are no fixed solutions (Kleinberg et al., 2017). For example, different definitions of fairness (like individual fairness and demographic parity) are mutually exclusive (Friedler et al., 2016). It would be naïve to suppose that we must – or indeed even can – resolve disagreements in moral philosophy before we start to design and deploy ADMS (Binns, 2018).

[34] The questions that guided the systematized review are summarised in Appendix 10.

To overcome this challenge, I conceptualise EBA as a governance mechanism that can help organisations adhere to any predefined set of (coherent and justifiable) ethics principles. EBA can, for example, take place within one of the ethical frameworks already mentioned, such as the *Ethics Guidelines for Trustworthy AI* for countries belonging to the EU. But organisations that design and deploy ADMS may also formulate their own sets of ethics principles and use these as a baseline to audit.

The remainder of this chapter proceeds as follows. In Section 3.2, I define 'ADMS' and discuss the features of such systems that give rise to ethical challenges. In Section 3.3, I explain what EBA is (or should be) in the context of ADMS. I also clarify the roles and responsibilities of different stakeholders in relation to EBA. In Section 3.4, I provide an overview of currently available procedures and tools for EBA of ADMS and how these are being implemented. I then offer three novel contributions to the existing literature. First, in Section 3.5, I articulate how EBA can support good governance. Second, in Section 3.6, I identify seven criteria for how to implement EBA procedures successfully. Third, in Section 3.7, I highlight and discuss the constraints associated with EBA of ADMS. In Section 3.8, I conclude that EBA, as outlined in this chapter, can help organisations manage some of the ethical risks posed by ADMS while allowing societies to reap the economic and social benefits of automation.

3.2 Automated decision-making systems

For the purpose of this thesis, I define ADMS as autonomous and self-learning systems that gather and process data to make or inform decisions that impact individuals, groups, or the natural environment with little or no human intervention. In previous literature on EBA, terms like 'algorithms', 'AI' and 'ADMS' are typically used interchangeably. However, as I explained in Section 1.3, I prefer to use the term ADMS consistently because it captures more precisely the technical features of the systems under investigation.

From an ethical perspective, it is primarily the *autonomous*, *adaptable*, and *complex* nature of ADMS that introduces new governance challenges. The autonomous nature of ADMS makes it difficult to predict and assign accountability when harms occur (Coeckelbergh, 2020; Tutt, 2017). Traditionally, the actions of technical systems have been linked to the user, the owner, or the manufacturer of the system. However, the ability of ADMS to adjust their behaviour over time undermines existing chains of accountability (Dignum, 2017). Finally, the complex, often opaque, nature of ADMS may hinder the possibility of linking the outcome of an action to its causes (Oxborough et al., 2018). Specifically, the structures that enable ML,

including the use of hidden layers in neural networks, also contributes to the technical opacity that complicates the attribution of accountability for the action of ADMS (Citron & Pasquale, 2014). While it should be noted that opacity can also be a result of intentional corporate or state secrecy (Burrell, 2016), my main concern here relates to inherent technical complexity.

Although mutually reinforcing, the levels of autonomy, complexity, and adaptability displayed by ADMS are all matters of degree (Tasioulas, 2018). In some cases, ADMS act in full autonomy, whereas in other ADMS provide recommendations to a human operator who has the final say (Cummings, 2004). In terms of complexity, a similar distinction can be made between ADMS that automate routine tasks and those which learn from their environment – and adapt their internal decision-making logic – to achieve goals.

Further, the scalability of ADMS means that it will become more difficult to manage system externalities. The impact of easily scalable technologies is hard to predict and may spill over borders and generations (Dafoe, 2017). For example, a single ADMS can now be used to automate decisions previously made by many different human decision makers in many different organisations (Kleinberg & Raghavan, 2021). This makes it challenging to reconcile different legitimate values and interests. The problem posed by ADMS is thus not only that norms will become harder to uphold but also harder to agree upon in the first place.

A final clarification. From a governance perspective, it is useful to view ADMS as parts of larger sociotechnical systems. Because ADMS adapt their behaviour based on input data and interactions with their environments (van de Poel, 2020), important dynamics may be lost or misunderstood if technical subsystems are targeted separately (Di Maio, 2014). This risk is summarised by what Lauer (2020) calls the fallacy of the broken part: when there is a malfunction, the first instinct is to identify and fix the broken part. Yet most accidents associated with ADMS can be traced not to coding errors but requirement flaws (Leveson, 2011). This implies that no purely technical solution will be able to ensure that ADMS are ethical (Kim, 2017). It also implies that EBA procedures must consider the design of an ADMS, the purpose for which it is employed, as well as the impact it exerts on its environment.

3.3 Ethics-based auditing

EBA is a *governance mechanism* that can be used by organisations to control or influence the ways in which ADMS are designed and deployed, and thereby, indirectly, shape the resultant characteristics of these systems. As mentioned in the introduction, EBA is characterised by a *structured process* whereby an entity's behaviour is assessed for consistency with relevant

principles or norms. It is worth noting that the entity in question, i.e., the subject of the audit, can be a person, an organisational unit, or a technical system.

Further, I use the expression 'ethics-based' instead of 'ethical' to avoid any confusion: I do neither refer to a kind of auditing conducted ethically, nor to the ethical use of ADMS in auditing, but to an auditing process that assesses ADMS based on their adherence to predefined ethics principles. Thus, EBA shifts the focus of the discussion from the abstract to the operational, and from guiding principles to managerial intervention throughout the product lifecycle, thereby permeating the design, deployment, and use of ADMS.

While widely accepted standards for EBA of ADMS have yet to emerge, it is possible to distinguish between different approaches (Mantelero, 2018). For example, *functionality audits* focus on the rationale behind ADMS' decisions; *code audits* entail reviewing the model architecture source code of an ADMS; and *impact audits* investigate the types, severity, and prevalence of effects of an ADMSs outputs (Mittelstadt, 2016). Again, these approaches are complementary and can be combined to design and implement EBA procedures in ways that are feasible and effective (more on this in Section 3.5).

Importantly, EBA differs from merely publishing a code of conduct, since its central activity consists of demonstrating adherence to a predefined baseline (ICO, 2020). EBA also differs from certification in important aspects. For example, whereas certification typically aims at producing an official document that attests to a particular status or level of achievement (Scherer 2016). To this end, certifications are issued by a third party, whereas auditing can (in theory) be done by (parts of) an organisation over itself for purely internal purposes. In sum, understood as a process of informing, interlinking, and assessing existing governance structures, EBA can provide the basis for – but is not reducible to – certification.

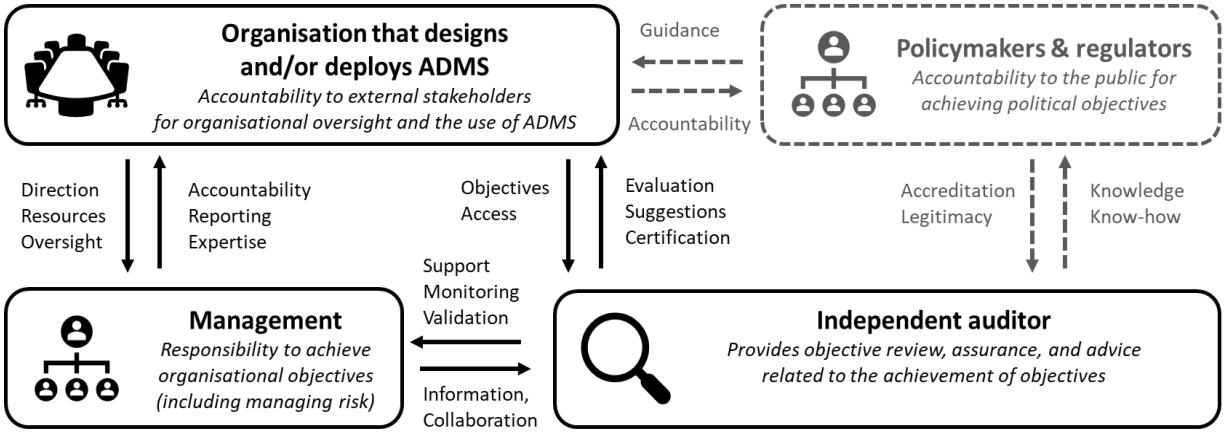
As a governance mechanism that aims to promote trust and transparency, auditing has a long history in areas like financial accounting and safety engineering (LaBrie & Steinke, 2019). As I showed in Chapter 2, valuable lessons can be learned from these domains. Most importantly, auditing is purpose-oriented process. In the case of EBA, this process is directed towards demonstrating that the design and use of ADMS align with specific ethics principles.

Throughout this purpose-oriented process, various *tools* (like software programs and standardised reporting formats) and *methods* (like stakeholder consultation or red teaming) are used to verify claims and create traceable documentation. Different EBA procedures employ different tools and contain different steps. In each case, however, this process of documentation enables the identification of the reasons why an ADMS was erroneous, which, in turn, help identify risks and mitigate future system failures (Felzmann et al., 2020).

Another lesson is that auditing presupposes operational independence between the auditor and the auditee. Whether the auditor is a government body, a third-party contractor, an industry association, or a specially designated function within larger organisations, the main point is to ensure that the audit is run independently from the regular chain of command within organisations (Power, 1997). The reason for this is not only to minimise the risk of collusion between auditors and auditees but also to clarify roles so as to be able to allocate responsibility for different types of harm or system failures (IIA, 2017).

Figure 7 below illustrates the relationships between organisations that design and deploy ADMS (who are accountable for their systems), the management of such organisations (who are responsible for achieving organisational goals, including adhering to ethical values), the independent auditor (who is tasked with assessing how well an organisation adheres to relevant principles), and regulators (who monitor organisations on behalf of the government and decision-making subjects). For EBA to be effective, auditors must be able to test ADMS for a variety of typical and atypical scenarios. Regulators can thus support the emergence and implementation of EBA procedures by providing the necessary infrastructure to share best practices and create standardised reporting formats (Keyes et al., 2019).

Figure 7. Roles and responsibilities during independent audits.



3.4 Status quo: Existing EBA procedures and tools

In the previous section, I defined EBA. In this section, I survey the landscape of currently available EBA procedures and tools. In doing so, I use examples from my systematised literature review to illustrate how EBA can provide new ways of detecting, understanding, and mitigating the unwanted consequences of ADMS.

When surveying the landscape, I distinguish between EBA *procedures*, i.e., protocols that define what is to be audited, when, by whom, and according to which standards, EBA *tools* i.e., conceptual models or software products that auditors can use to measure, evaluate, or visualise one or more properties of ADMS at various steps during the audit process. These are analytically useful since, as I will show, research on EBA procedures and EBA tools constitute two distinct bodies of literature that rarely are in dialogue with each other.

3.4.1 Ethics-based auditing procedures

Currently available EBA procedures originate from one of four processes. The first type originates from 'top-down' national and regional strategies and guidelines for how to design and use ADMS. For example, the Government of Canada (2019) has published a *Directive on Automated Decision-Making*, and the Government of Singapore (2020) has published guidelines on how to design and use ADMS responsibly. Similar strategies and guidelines have been published by the Governments of Australia (Dawson et al., 2019), Brazil (2021), and the UK (2021), amongst others.[35] To be clear, these strategies and guidelines are not EBA procedures in and of themselves. However, the ethics principles they stipulate form a baseline against which ADMS can be assessed during EBA.[36]

At a European level, this development was shaped by the AI4People project, which proposed that 'auditing mechanisms' should be developed to identify unwanted ethical consequences of ADMS (Floridi et al., 2018). Since then, the AI HLEG[37] has published not only the *Ethics-Guidelines for Trustworthy AI* (AI HLEG 2019), but also a corresponding *Assessment List for Trustworthy AI* (AI HLEG, 2020). This assessment list is intended for self-evaluation purposes and can thus be incorporated into EBA procedures.

Most recently, these efforts have culminated in *Artificial Intelligence Act* (AIA). The AIA takes a risk-based approach. While ADMS that pose 'unacceptable risk' are proposed to be completely banned, so-called 'high-risk' systems will be required to undergo legally mandated ex-ante and ex-post conformity assessments. However, even for ADMS that pose 'minimal' or 'limited' risk, the European Commission (2021a) encourages technology providers to adopt and adhere to voluntary codes of conduct (more on this in Chapter 5).

---

[35] For an overview, please see the OECD's (2023) live repository of '*National AI policies and strategies.'*

[36] For a review of existing sets of ethics principles, see e.g., Hagendorff (2020) or Floridi and Cowls (2019).

[37] The AI HLEG is an independent expert group set up by the European Commission in June 2018.

The second type of EBA frameworks emerges 'bottom-up', from the expansion of data regulation authorities to account for the effects ADMS have on informational privacy (CNIL, 2019; IAF, 2019). For example, in the UK the ICO (2020) has issued guidance on how to audit ADMS. Building on an extensive experience of translating principles into practice, governance protocols developed by data regulation agencies provide valuable blueprints for EBA procedures. For instance, the French CNIL's *Privacy Impact Assessment Framework* requires organisations to describe the context of their data processing when analysing how well it aligns with fundamental ethics principles (CNIL, 2019). This need for contextualisation applies not only to data management but also to the use of ADMS at large.

Another transferable lesson is that organisations should conduct an independent ethical evaluation of software they procure from – or outsource production to – third-party vendors (ICO, 2018). At the same time, EBA procedures with roots in data regulation tend to account only for specific ethical concerns, e.g., those related to privacy. This calls for caution. Since there is a plurality of ethical values which may serve as legitimate normative ends (think of freedom, equality, justice, proportionality, etc), an exclusive focus on one, or even a few, ethical challenges risks leading to sub-optimisation from a holistic perspective (Berlin, 1997).

The third type of EBA procedures stem from not from public but from the private sector. A new industry is emerging whereby professional service providers offer to help technology providers demonstrate that the ADMS they design or deploy are ethical (Landers & Behrend, 2022). As part of this movement, EBA procedures have been developed by accounting firms like Deloitte (2020), EY (2018), KPMG (2020), and PwC (2020), startups like Babl AI (2023), Holistic AI (2023), and ORCAA (2020), as well as by industry associations and professional standard-setting bodies (see e.g., IEEE SA, 2020; ISO, 2022; NIST, 2022; VDE, 2022). These EBA procedures tend to be extensions of auditing procedures historically used in the fields of systems engineering (Leveson, 2011) and IT governance (Senft & Gallegos, 2009), meaning that they focus less on the technical properties of ADMS and more on the adequacy of technology providers software development processes and QMS.

The final source of EBA procedures is academic researchers (Bandy, 2021). As I showed in Chapter 2, scholars from different disciplines have developed EBA procedures (see e.g., Felländer et al., 2022; Minkkinen et al., 2022; Zicari et al., 2021). Some of these procedures focus on *compliance assurance*, i.e., comparing organisational processes to existing standards or best practices. For example, Raji et al. (2020) suggested that process-oriented audits can check that the engineering processes involved in the design and deployment of ADMS meet specific standards. Other EBA procedures focus on *risk assurance*, i.e., asking

open-ended questions about how ADMS are designed and what risks they pose. Koshiyama et al. (2022), for instance, frame their EBA procedure as a governance mechanism for technology providers to manage legal, ethical, and technology-oriented risks.

To synthesise, currently available EBA procedures have emerged from different sources and for different purposes. As a result, they differ both in terminology and emphasis. However, despite these differences, my analysis suggests that currently available EBA procedures largely converge in terms of methodology. Whether developed by policymakers, professional service providers, or researchers, EBA procedures are based on process-oriented assessment that resemble traditional governance audits and technology impact assessments. Currently available EBA procedures can be thus summarised in 8 steps, whereby auditors assess whether technology providers have adequate processes in place to:

1) *Describe* the purpose of the ADMS they design or deploy.
2) *Define* the verifiable criteria based on which the ADMS should be assessed.
3) *Disclose* the process, including a full account of the data use and parties involved.
4) *Assess* the impact the ADMS has on individuals, communities, and its environment.
5) *Evaluate* whether the benefits and mitigated risks justify the use of ADMS.
6) *Determine* the extent to which the system is reliable, safe, and transparent.
7) *Document* the results and considerations, and
8) *Evaluate* periodically, i.e., create a feedback loop.

While imposing procedural transparency obligations on technology providers, existing EBA procedures leave many questions open. As a rule, they do not stipulate *how* ADMS should be assessed, or according to *which* criteria. This is where technically oriented EBA tools come in.

3.4.2 Ethics-based auditing tools

EBA *tools* are conceptual models or software products that help measure, evaluate, or visualise one or more properties of ADMS. With the aim to enable and facilitate EBA, a great variety of such tools have been developed by both academic and privately employed data scientists. While these typically apply mathematical definitions of principles like fairness, accountability, and transparency to measure and evaluate ADMS' properties (Keyes et al., 2019), different tools help ensure the ethical alignment of ADMS in different ways.

Through my literature review, I found that currently available EBA tools can be divided into five different categories depending on the purposes they serve. It would be possible to divide the literature in other ways, e.g., based on the statistical techniques different tools

employ. However, for my purposes it is more useful to provide an overview of the existing landscape based on instrumental categories, since these will feed into the methodological affordances of EBA discussed in Section 3.5. In what follows, I discuss the five different types of currently available EBA tools and illustrate these using real-world examples.

First, several tools have been developed that facilitate the audit process by visualising the outputs of ADMS. *FAIRVIS*, for example, is a visual analytics tool that integrates a subgroup discovery technique, thereby informing normative discussions about group fairness (Cabrera et al., 2019). Another example is *Fairlearn*, an open-source toolkit that treats any ADMS as a black box. Fairlearn's interactive visualisation dashboard helps users compare the performance of different models (Microsoft, 2020). These tools are based on the idea that visualisation helps spark ethical deliberation in the software development process.

A second category of EBA tools improve the interpretability of complex ADMS by generating more straightforward rules that explain their predictions. For example, Shapley Additive exPlanations, or *SHAP*, calculates the marginal contribution of relevant features underlying a model's prediction (Leslie, 2019). The explanations provided by such tools are useful, e.g., when determining whether protected features have unjustifiably contributed to a decision made by ADMS (Fabbri & LeFevre, 2011). However, such explanations also have important limitations. For example, tools that explain the contribution of features that have been intentionally used as decision inputs may not determine whether protected features have contributed unjustifiably to a decision through proxy variables.

The third category of EBA tools help convey the reasoning behind ADMS by applying one of three strategies: *data-based explanations* provide evidence of a model by using comparisons with other examples to justify decisions; *model-based explanations* focus on the algorithmic basis of the system itself; and *purpose-based explanations* focus on comparing the stated purpose of a system with the measured outcomes (Kroll, 2018). Different types of explanations are possible. However, EBA tend to focus on local interpretability, i.e., on explanations targeted at individual stakeholders – such as decision subjects or external auditors – and for specific purposes like reputation management or third-party verification. Here, a parallel can be made to what Loi et al. (2020) call *transparency as design publicity*, whereby organisations that design or deploy ADMS are expected to publicise the intentional explanation of the use of a specific system as well as the procedural justification of the decision it takes.

A fourth category consist of tools developed to help democratise the study of ADMS. Consider the *TuringBox*, which was developed as part of a time-limited research project at MIT. This platform allowed software developers to upload the source code of an ADMS to let

others examine them (Epstein et al., 2018). The TuringBox thus provided an opportunity for developers to benchmark their ADMS' performance with regards to different properties. Simultaneously, the platform also allowed independent researchers to evaluate the outputs from ADMS, thereby adding an extra layer of procedural transparency to the software development process. DeVos et al. (2022) go further still and propose a framework for '*user driven audits.'* The idea thereby is that users – through their interactions with ADMS – can detect and report vulnerabilities that formal auditing procedures have failed to detect.

The final category of EBA tools consists of templates or software programs that help organisations document the software development process or monitor ADMS throughout their lifecycle. *AI Fairness 360* developed by IBM, for example, includes metrics and algorithms to monitor, detect, and mitigate bias in datasets and models (Bellamy et al., 2019). Other tools have been developed to aid developers in making pro-ethical design choices by providing information about the properties and limitations of ADMS. Such tools include end-user license agreements, tools for detecting bias in datasets (Saleiro et al., 2018), and tools for improving transparency like datasheets (Gebru et al., 2018).

The literature review on which this overview of currently available EBA tools was conducted in March 2021. Since then, many more EBA tools have been put forward.[38] However, while tools with new names and increasingly sophisticated capabilities have appeared, they fit well into the categories of EBA tools outlined in this section.

A further observation. Most of the EBA tools discussed in this section were developed by academic researchers or private companies. However, governments have also contributed. For example, the Government of Singapore has developed *A.I.verify*, an open-source toolkit that enables industry to demonstrate adherence to voluntary ethics principles (PDPC, 2022). The toolkit includes metrics and methods to assess ADMS for fairness, explainability, and safety. A.I.verify is currently being piloted in together with private sector technology providers.

The key takeaway from this section is that EBA is not a theoretical proposition but an established practice that can be observed in applied settings. Policymakers, researchers, and professional service providers have proposed a plurality of EBA procedures, and computer scientists and industry practitioners have developed a wide range of EBA tools to help auditors measure, visualise, or evaluate the properties of ADMS. However, the motivations offered to

[38] See e.g., Ayling and Chapman (2022) or Liang et al., (2022) for more recent reviews.

produce these tools and procedures are often cursory. What is lacking in the existing literature is thus theoretical explanation of how EBA contributes to good governance.

3.5 A vision for ethics-based auditing of ADMS

The first half of this chapter has focused on describing existing literature. However, in this section, I go beyond description of the status quo to articulate a vision for what EBA should be in the context of EBA. To do so, I followed the methodology described in Section 3.1.2.

As I demonstrated in the previous section, many EBA procedures and tools have already been developed. The researchers and policymakers that develop or promote these procedures and tools have made different claims about how EBA contributes to good governance. However, while differing in terminology, my synthesis of previous work finds that these claims centre around a limited number of themes.

To start with, an important function of EBA is diagnostic (AIEIG, 2020). Before asking whether we would expect an ADMS to be ethical, we must consider which mechanisms we have to determine what it is doing at all. By gathering data on system states (both organisational and technical), EBA enables stakeholders to evaluate the reliability of ADMS. A systematic audit is thus a first step to make informed model selection decisions and to understand the causes of adverse effects (Saleiro et al., 2018). In short, EBA provide decision-making support to managers and software developers by defining and monitoring outcomes, e.g., by showing the normative values embedded in ADMS.

Further, EBA can increase public trust in technology by enhancing operational consistency and procedural transparency (Loi et al., 2020; Binns, 2018). Mechanisms such as documentation and actionable explanations are essential to help individuals understand why a decision was reached and how to contest it (Wachter et al., 2018). This also has economic implications. While there may be many justifiable reasons to abstain from using available technologies in certain contexts, fear and ignorance may lead societies to underuse available technologies even in cases where they would do more good than harm (Cookson, 2018; Floridi et al., 2018). In such cases, increased public trust in ADMS could help unlock economic growth. However, to drive trust in ADMS, explanations need to be actionable and selective (Barredo Arrieta et al., 2020). This is possible even when algorithms are technically opaque since ADMS can be understood intentionally and in terms of their inputs and outputs.

Another methodological affordance of EBA is that it allows for local alignment of ethics and legislation (Kazim et al., 2021; Lui et al., 2022; Mittelstadt, 2016). While some

normative metrics must be assumed when evaluating ADMS, EBA allows organisations to choose which set of ethics principles they seek to adhere to. This allows for contextualisation. Returning to the example with fairness above, the most important aspect from an EBA perspective is not which specific definition of fairness is applied in a specific case, but that this decision is communicated transparently and publicly justified. By focusing on identifying tensions and risks, as well as by communicating the same to relevant stakeholders like customers or independent industry associations, EBA can help organisations demonstrate adherence to both sector-specific and geographically dependent norms and legislation.

EBA procedures can also help relieve human suffering by anticipating potential negative consequences before they occur (Raji et al., 2020). There are three overarching strategies to mitigate harm: *pre-processing*, i.e., reweighing or modifying input data; *in-processing*, i.e., model selection or output constraints; and *post-processing*, i.e., calibrated odds or adjustment of classifications (Koshiyama, 2019; Zhang et al., 2023). These strategies are not mutually exclusive. By combining requirements on system performance with automated controls, EBA can help both developers test and improve the performance of ADMS (Mahajan et al., 2020) and enable organisations to establish safeguards against unwanted behaviours.

Moreover, EBA helps balance conflicts of interest. For example, data subjects' right to explanation must be reconciled with jurisprudence and counterbalanced with intellectual property (IP) rights (Wachter et al., 2018). By containing access to sensitive parts of the review process to authorised third-party auditors, EBA can provide a basis for accountability while preserving privacy and intellectual property rights (Imana et al., 2023). There are more ways in which EBA helps balance conflicts of interest. By subjecting themselves to independent audits, technology providers can display the tradeoffs involved in different design choices and communicate these to relevant stakeholders (Whittlestone et al., 2019). This way, EBA make visible implicit choices and tensions, and help technology providers strike justifiable tradeoffs within the bounds of legal permissibility and commercial viability.

Finally, EBA can help human decision-makers to allocate accountability by tapping into existing governance structures (Bartosch et al., 2018). Within organisations, EBA can forge links between non-technical executives and developers (Raji et al., 2020). Externally, EBA help organisations validate the functionality of ADMS. The main idea thereby is that the causal chain behind decisions made by ADMS can be revealed which, in turn, allow stakeholders to identify who should be held accountable for potential ethical harms (Vecchione et al., 2021). Taken together, EBA helps clarify the roles and responsibilities of different stakeholders involved in the process of designing and deploying ADMS.

To summarise, EBA contributes to good governance of ADMS by promoting procedural transparency and regularity. More specifically, EBA displays six, interrelated and mutually reinforcing, methodological affordances. EBA can help:

1) *Provide* decision-making support by visualising and monitoring outcomes,
2) *Inform* individuals why a decision was reached and how to contest it,
3) *Allow* for a sector-specific approach to ADMS governance,
4) *Relieve* human suffering by anticipating and mitigating harms,
5) *Balance* conflicts of interest, and
6) *Allocate* accountability by tapping into existing governance structures.

Considered in isolation, each of these methodological affordances have been formulated in one way or another different researcher in different contexts. However, it is one of the main contributions of this chapter to synthesise these into a coherent theoretical explanation of how EBA contributes to good governance.

The six methodological affordances articulated in this section provides a baseline against which the feasibility and effectiveness of specific EBA procedures can be evaluated. However, it is important to stress that the methodological advantages highlighted in this section are potential and far from being guaranteed. The extent to which these can be harnessed in practice depends not only on complex contextual factors but also on how EBA procedures are designed. To realise its full potential as a governance mechanism, EBA of ADMS needs to meet specific criteria. In the next section, I turn to specifying these criteria.

3.6 Criteria for successful implementation

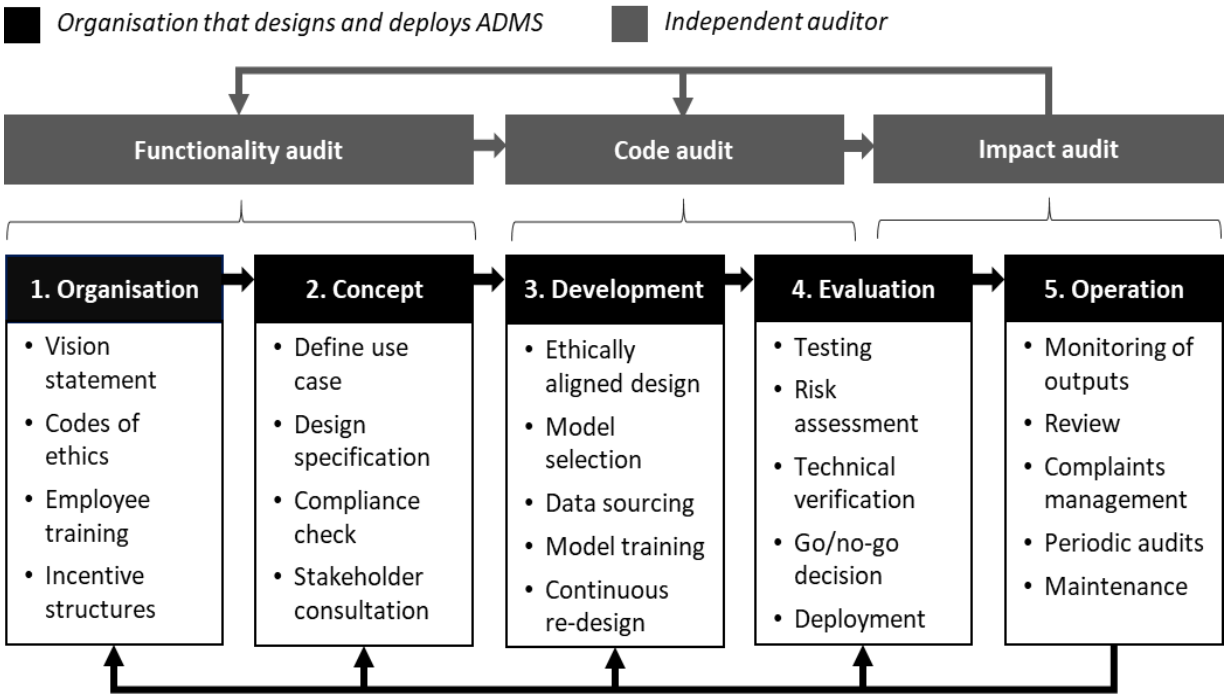
Best practices for EBA have yet to emerge. However, as discussed in Section 3.4, researchers have already developed and piloted a wide range of EBA tools and procedures. These early attempts hold valuable lessons for organisations that wish to implement feasible and effective EBA procedures. While some of these lessons concern how stakeholders view EBA, others concern the design of EBA procedures. In this section, I discuss the most important lessons from previous work and synthesise these into seven criteria for how to design EBA procedures.

As a starting point, it should be acknowledged that ADMS are not isolated technologies. Rather, ADMS are both shaped by and help shape larger sociotechnical systems (Dignum, 2017). As we have seen, a wide variety of EBA frameworks and tools have already been developed to help organisations and societies manage the ethical risks posed by ADMS.

However, my review of previous work suggest that these tools are often employed in isolation. To be feasible and effective, EBA procedures need to be holistic, i.e., combine existing tools and procedures into structured processes that monitor each stage of the software development lifecycle to identify and correct the points at which ethical failures (may) occur.

In practice, this means that EBA procedures should combine elements of (i) *functionality auditing*, which focuses on the rationale behind decisions (and why they are made in the first place); (ii) *code auditing*, which entails reviewing the source code of an algorithm; and (ii) *impact auditing*, whereby the severity and prevalence of the effects of an algorithm's outputs are investigated. This logic is illustrated in Figure 8 below.

Figure 8. EBA helps inform, formalise, and interlink existing governance structures.



A further lesson from previous research is that audits are only meaningful insofar as they allow technology providers to verify claims made about their ADMS. This implies that EBA procedures themselves must be *traceable*. By documenting the steps taken in the design and development of ADMS, audit trails can help organisations verify claims about their engineered systems (Brundage et al., 2020). However, a distinction should be made between traceability and transparency: while transparency is often invoked to improve trust (Springer & Whittaker, 2019), full transparency concerning the content of audits may not be desirable, e.g., with regards to privacy and IP rights. Instead, what counts is procedural transparency and regularity.

Further, to ensure that ADMS are ethically sound, organisational policies must be broken down into tasks for which individual agents can be held *accountable* (Ananny & Crawford, 2018). By formalising the software development process and revealing (parts of) the causal chain behind decisions made by ADMS, EBA helps clarify the roles and responsibilities of managers and data scientists. However, allocating responsibilities is not enough. Sustaining a culture of trust also requires that people who breach ethical norms are subject to proportional sanctions (Ellemers et al., 2019). By providing avenues for whistle-blowers and promoting a culture of ethical behaviour, EBA helps strengthen interpersonal accountability within organisations (Koene et al., 2019). At the same time, doing the right thing should be made easy. This can be achieved through strategic governance structures that align profit with purpose. The 'trustworthiness' of a specific ADMS is never just a question about technology but also about value alignment (Christian, 2020; Gabriel, 2020). In practice, this means that the checks and balances developed to ensure safe and benevolent ADMS must be incorporated into policies, organisational incentive structures, and individual paths.

Importantly, EBA does not provide an answer sheet but a playbook. Hence, EBA should be viewed as a *dialectic* process wherein the auditor ensures that the right questions are asked and answered adequately (Goodman, 2016). To manage the risk that independent auditors would be too easy on their clients, licences should be revoked from both auditors and system owners in cases where ADMS fail. However, it is difficult to ensure that an ADMS contains no bias or to guarantee its fairness (Microsoft, 2020). Instead, the goal from an EBA perspective should be to provide useful information about when an ADMS is causing harm or when it is behaving in a way that is different from what is expected. This pragmatic insight implies that audits need to monitor and evaluate system outputs *continuously*, i.e., through 'oversight programs' (Etzioni & Etzioni, 2016), and document performance characteristics in a comprehensible way (Mitchell et al., 2019). Hence, continuous EBA of ADMS implies considering system impacts as well as organisations, people, processes, and products.

Finally, the alignment between ADMS and specific ethics principles is a design question. Ideally, properties like interpretability and robustness should be built into systems from the start, e.g., through *Value-Aligned Design* (Bryson & Winfield, 2017). However, the context-dependent behaviour of ADMS makes it difficult to anticipate the impact ADMS will have on the complex environments in which they operate (Chopra & Singh, 2018). By incorporating an active feedback element into the software development process, EBA can help inform the continuous re-design of ADMS. Although this may seem radical, it is already

happening: most sciences, including engineering and jurisprudence, do not only study their systems: they simultaneously build and modify them (Floridi, 2017a).

It should be stressed that the primary responsibility for identifying and executing steps to ensure that ADMS are ethically sound rests with the management of the organisations that design and operate such systems. The independent auditor's responsibility is to (i) assess and verify claims made by the auditee about its processes and ADMS and (ii) ensure that there is sufficient documentation to respond to potential inquiries from public authorities or individual decision subjects. The process of EBA should also help spark and inform ethical deliberation throughout the software development process. The idea is that continuous monitoring and assessment ensures that a constant flow of feedback concerning the ethical behaviour of ADMS is worked into the next iteration of their design and application. Figure 8 above thus also illustrates how the process of EBA runs in parallel with the software development lifecycle.

Taken together, these generalisable lessons suggest that even imperfectly implemented EBA procedures can make a real difference to the ways in which ADMS are designed and deployed. However, my analysis of previous work also finds that, to be feasible and effective, EBA procedures must meet seven criteria. More specifically, to help organisations manage the ethical risks posed by ADMS, EBA procedures should be:

1) *Holistic*, i.e., combine both process-oriented and technology-oriented assessments,
2) *Traceable*, i.e., assign responsibilities and document decisions to enable follow-up,
3) *Accountable*, i.e., help link unethical behaviours to proportional sanctions,
4) *Strategic*, i.e., align ethical values with organisational strategies and incentives,
5) *Dialectic*, i.e., view EBA as a constructive and collaborative process,
6) *Continuous*, i.e., identify, monitor, and evaluate system impacts over time, and
7) *Driving re-design*, i.e., provide feedback to inform the continuous re-design of ADMS.

Let me immediately qualify the above contribution with two caveats. First, these seven criteria have been derived solely based on a systematised literature review and theory synthesis. They are theoretical propositions, not empirically established facts. As discussed in Chapter 1, I take a pragmatist research stance, and according to pragmatist epistemology theoretical propositions are only valuable insofar as they are useful. Hence, it remains a task for future applied research to test not so much the validity as the usefulness of the above listed criteria.

Second, the seven criteria for how to successfully design EBA procedures derived and discussed in this section are aspirational. In practice, they are unlikely to be satisfied all at once. Nevertheless, we must not let perfect be the enemy of good. The list would of course need to

be revised, should applied research conclude that one or more of the above-listed criteria do not help auditors and technology providers identify and mitigate the ethical risks posed by ADMS – or that other design criteria work even better. In the meanwhile, policymakers and organisations that design and deploy ADMS are advised to consider these seven criteria when developing and implementing EBA procedures.

3.7 Constraints associated with ethics-based auditing

Despite the methodological advantages identified in Section 3.5, it is important to remain realistic about what EBA can be expected to achieve. Without a shared understanding of what EBA is, let alone best practices for how to conduct it, claims that ADMS have been audited are hard to verify, and may even do more harm than good by giving a false sense of security.

This leads us to the third contribution of this chapter. As part of my review of existing EBA tools and procedures, I documented and synthesised the constraints of EBA as an ADMS governance mechanism discussed in the previous works. My findings suggest that EBA – even if implemented according to the seven criteria listed in Section 3.6 – is subject to a wide range of conceptual, technical, social and economic, as well as organisational and institutional constraints. In the remainder of this section, I discuss what these different types of constraints are, and highlight the most pressing constraints associated with EBA of ADMS.

3.7.1 Conceptual constraints

With conceptual constraints, I refer to constraints which cannot be easily overcome by means of technical innovation or political decision. Instead, they must be managed continuously by balancing the need for ethical alignment with tolerance and respect for pluralism.

To begin with, EBA is conceptually constrained by hidden political tensions. The apparent consensus surrounding high-level ethics principles often mask unresolved disputes about the definitions of normative concepts like fairness and justice (Mittelstadt, 2019). For example, the reviewed literature accommodates more than six definitions of fairness, including individual fairness, demographic parity, and equality of opportunity (Kusner et al., 2017). Some of these interpretations are mutually exclusive, and specific definitions of fairness can even increase discrimination according to others. While EBA can help ensure compliance with a given policy, how to prioritise between conflicting interpretations of ethical principles remains a normative question. This is because translating principles into practice often requires trade-offs between different legitimate, yet conflicting normative values. Using personal data, for example, may improve public services by tailoring them but compromise privacy.

Similarly, while increased automation could make lives more convenient, it also risks undermining human autonomy. How to negotiate justifiable trade-offs is a context-dependent, multi-variable problem. While audits cannot guarantee that a justifiable balance has been struck, the identification, evaluation, and communication of trade-offs can be included as assessment criteria. One function of EBA is thus to make visible implicit choices and tensions, give voice to different stakeholders, and arrive at resolutions that, even when imperfect, are at least publicly defensible (Whittlestone et al., 2019b).

EBA is also conceptually constrained by the difficulty to quantify externalities that occur due to indirect causal chains over time (Dafoe, 2017). This issue is exacerbated by the fact that the quantification of social phenomena strips away local knowledge and context (Mau, 2019). On the one hand, tools claiming to operationalise ethics mathematically risk falling into the trap of technological solutionism (Lipton & Steinhardt, 2019). On the other hand, tools that focus on only minimum requirements provide little incentives for organisations to go beyond legal compliance.

3.7.2 Technical constraints

With technical constraints, I refer to constraints tied to the autonomous, complex, and adaptable nature of ADMS. These constraints are time and context-dependent and thus likely to be relieved or transformed by future research. Three of them are worth highlighting.

First, consider how the opacity stemming from the technical complexity of ML models hinder their interpretation (Oxborough et al., 2018). This opacity of ADMS constrains the effectiveness of audits insofar as it makes it difficult to assign and trace responsibility when harm occurs. Technical complexity also makes it difficult to audit a system without perturbing it. Further, there is a risk that sensitive data may be exposed during the audit process itself (Kolhar et al., 2017). To manage this challenge, third party auditors can be given privileged and secured access to private information to assess whether claims about the safety, privacy, and accuracy made by the system developer are valid. As of today, however, most EBA schemes do not protect user data from third-party auditors.

A second technical constraint stems from the use of agile software development methods. The same agile qualities that help developers meet rapidly changing customer requirements also make it difficult for them to ensure compliance with pre-specified requirements. One approach to managing this tension is to incorporate agile methodologies (Strenge & Schack, 2020) that make use of ‘living traceability’ in the audit process. These methods provide snapshots of ADMS under development in real-time (Steghöfer et al., 2019).

Despite the availability of such pragmatic fixes, however, the effectiveness of EBA remains limited by the difficulty of providing assurance for ADMS that evolve over time.

Finally, EBA is technically constrained by the fact that laboratories differ from real-life environments (Auer & Felderer, 2018). Put differently, given the data- and context-dependent behaviour of ADMS, only limited reasoning about their later performance is possible based on testing in controlled settings. To manage this challenge, test environments for simulation can be complemented by continuous EBA of live applications which constantly execute the ADMS. One example is *live experimentation*, i.e., the controlled deployment of experimental features in live systems to collect runtime data and analyse the corresponding effect (Fagerholm et al., 2014). Still, meaningful quality assurance is not always possible within test environments.

3.7.3 Economic and social constraints

Economic and social constraints refer to those deriving from the incentives of different actors. Unless these incentives are aligned with the ethics principles guiding the design and use of ADMS, economic and social factors will constrain the feasibility and effectiveness of EBA.

EBA imposes costs, financial and otherwise. Even when the costs of audits are justifiable compared to the aggregated benefits, society will face questions about which stakeholders would reap which benefits and pay which costs. For example, the cost of EBA risks having a disproportionate impact on smaller companies (Goodman, 2016). Similarly, licensing systems for ADMS are likely to be selectively imposed on specific sectors, like healthcare or air traffic (Council of Europe, 2018). The point is that both the costs and benefits associated with EBA should be distributed to not unduly burden or benefit particular groups in, or sectors of, society. Similarly, demands for ethical alignment must be balanced with incentives for innovation and adoption. Pursuing rapid technological progress leaves little time to ensure that developments are robust and ethical (Whittlestone et al., 2019b). Thus, companies find themselves wedged between the benefits of disruptive innovation and social responsibility and may not act ethically in the absence of oversight (Turner Lee, 2018).

Moreover, there is always a risk of adversarial behaviour during audits. The ADMS being audited may attempt to trick the auditor (Rahwan, 2018). An example of such behaviour was the diesel emission scandal, during which Volkswagen intentionally bypassed regulations by installing software that manipulated exhaust gases during tests (Conrad, 2018). An associated risk is that emerging EBA end up reflecting and reinforcing existing power relations. Given an asymmetry in know-how and computational resources between data controllers and public authorities, auditors may struggle to review ADMS (Kroll, 2018). For example, industry

representatives may choose not to reveal insider knowledge but instead use it to obtain weaker standards (Koene et al., 2019). Sector-specific approaches may thus lead to a shift of power from juridical courts to private actors. Even if, in such a scenario, audits reveal flaws within ADMS, asymmetries of power may prevent corrective steps from being taken.

Another concern relates to the fact that ADMS mediate human interactions. From an EBA perspective, *nudging*, i.e., the process of influencing personal preferences through positive reinforcement or indirect suggestion (Thaler et al., 2008), may shift the normative baseline against which ethical alignment is benchmarked. This risk is aggravated *by automation bias*, i.e., the tendency of humans to trust information that originates from machines more than their own judgement (Cummings, 2004). Consequently, the potentially transformative effects associated with ADMS pose challenges for how to trigger and evaluate audits.

3.7.4 Organisational and institutional constraints

Organisational and institutional constraints concern the design and implementation of EBA procedures. Because these constraints depend on legal sanctioning, they are inevitably linked to questions about power. The central question here is who audits whom?

As of today, a clear institutional structure is lacking. To establish integrity and validity, EBA of ADMS must therefore adhere to a transparent and well-recognised process. However, both internal audits and those performed by professional service providers are subject to concerns about objectivity. A more plausible way to mandate EBA of ADMS would be the creation of a regulatory body to oversee system owners and auditors. Just as the Food and Drug Administration tests and approves medicines, a similar agency could be set up to approve specific types of ADMS (Tutt, 2017). Such an agency would be able to engage in *ex ante* regulation rather than relying on *ex post* judicial enforcement. However, the main takeaway is that EBA will only be as good as the institution backing it (Boddington et al., 2017).

In a similar vein, EBA is only effective if auditors have access to the information and resources required to carry out rigorous audits. Thus, EBA is infeasible without regulatory compulsion or cooperation from system owners. Data controllers have, for example, an interest not to disclose trade secrets. Moreover, the resources required to audit ADMS can easily exceed those available to auditors. If, for example, auditors have no information about special category membership, they cannot determine whether a disparate impact exists. Consequently, the effectiveness of EBA is constrained by a lack of access to both relevant information and resources in terms of manpower and computing power.

There are also fundamental tensions between national jurisdictions and the global nature of technologies (Erdelyi et al., 2018). Thus, rules need to be harmonised across domains and boarders. However, such efforts face a hard dilemma. On the one hand, the lack of shared ethical standards for ADMS may lead to protectionism and nationalism. On the other hand, policy discrepancies may cause a race to the bottom where organisations seek to establish themselves in territories that provide a minimal tax burden and maximum freedom for technological experimentation (Floridi, 2019a). As a result, the effectiveness of EBA of ADMS remains constrained by the lack of international coordination.

3.7.5 Summary of constraints

EBA is subject to a wide range of conceptual, technical, social and economic, as well as organisational and institutional constraints. To design feasible and effective EBA procedures, all these constraints must be understood and accounted for. However, different types of constraints require different responses. Technical constraints may be alleviated by further research, and institutional constrains may be addressed by policymakers' future efforts. In contrast, the conceptual constraints listed above will need to be continuously managed.

The constraints highlighted in this section do not seek to diminish the merits of EBA. In contrast, my hope is that the typology of constraints associated with EBA outlined in this section will serve three constructive purposes: first, to provide a roadmap for future research by drawing attention to the issues that currently constrain the feasibility and effectiveness of EBA; second, to guide policymakers' efforts to support the emergence of feasible and effective EBA procedures; and third, to caution against the overpromising of an emergent industry.

Professional service providers developing EBA procedures or offering EBA services to technology providers should acknowledge the constraints listed in this section, and stakeholders confronted with claims that ADMS have been audited should approach these with a healthy degree of scepticism. In Chapter 8, I will return to discuss these and other implications of my findings for readers from my different target audiences.

3.8 Concluding remarks

Researchers, policymakers, and countless op-eds have pointed towards EBA as a promising governance mechanism to identify and mitigate the ethical risks ADMS pose. However, what EBA is and how it contributes to good ADMS governance has remained poorly defined. To bridge that gap, the review and analysis presented in this chapter have sought to clarify what

EBA is – or at least should be – in the context of ADMS governance. This can be summarised in one paragraph:

The responsibility to ensure that ADMS are ethically sound lies with the organisations that develop and operate them. Functionally, EBA is thus to be understood as a governance mechanism that organisations can employ to demonstrate that their ADMS adhere to predefined ethics principles. Operationally, EBA is characterised by a structured and independent process. The subject of the audit can either be an organisation, a technical system, or a combination thereof. EBA procedures define what should be audited, when, by whom, and according to which standards. At various points in the assessment, auditors can leverage different EBA tools to identify, measure, document, or visualise the normative values embedded in ADMS or the impact these systems have on individuals and groups.

Equipped with this conceptualisation of what EBA is, we are now positioned to answer:

SQ1: What are the affordances and constraints of EBA as a governance mechanism to address the ethical risks posed by ADMS?

EBA has several theoretical affordances as an ADMS governance mechanism. Most importantly, it contributes to good governance by promoting procedural transparency and regularity. In doing so, EBA provides decision-making support to managers and software developers by visualising the normative tradeoffs involved in the design process; helps data subjects understand why a decision was reached and how to contest it; and allows external stakeholders to identify who should be held accountable for potential ethical harms by clarifying the roles and responsibilities within organisations that design or deploy ADMS. EBA also has several secondary methodological affordances. For example, by restricting access to sensitive information regarding the design of ADMS to authorised auditors, EBA helps balance the need for external scrutiny with the protection of data privacy and IP rights. Finally, by continuously assessing the adequacy of technology providers' software development processes and QMS, and by sparking ethical deliberation amongst software developers, EBA can help identify and mitigate risks before harm occurs.

Of course, this does not mean that traditional governance mechanisms are redundant. On the contrary, by contributing to procedural regularity and transparency, EBA of ADMS is meant to complement, enhance, and interlink other governance mechanisms like human oversight, certification, and regulation. For example, by demanding that ethics principles and codes of conduct are clearly stated and publicly communicated, EBA ensures that organisational practices are subject to additional scrutiny which, in turn, may counteract 'ethics

shopping.' Similarly, EBA helps reduce the risk of 'ethics bluewashing' by allowing organisations to validate the claims made about their ethical conduct and the ADMS they operate. Thereby, EBA constitutes an integral component of multifaceted approaches to managing the ethical risks posed by ADMS.

However, even in contexts where EBA is necessary to ensure ethical alignment of ADMS, it is by no means sufficient. As my review and analysis in this chapter have demonstrated, EBA is subject to a wide range of conceptual, technical, social, economic, organisational, and institutional constraints. For example, it remains unfeasible to anticipate all long-term and indirect consequences of a particular decision made by an ADMS. Further, while EBA can help ensure alignment with a given policy, how to prioritise between irreconcilable normative values remains a fundamentally normative question. Therefore, the design and implementation of EBA frameworks must be viewed as a part of – and not separated from – the debate about the type of society humanity wants to live in and what moral compromises individuals are willing to strike in its making.

In conclusion, structured and independent EBA procedures can help organisations validate claims about their ADMS. However, EBA will not and should not replace the need for continuous ethical reflection and deliberation among individual moral agents.

CHAPTER 4

OPERATIONALISING AI GOVERNANCE THROUGH ETHICS -BASED AUDITING: AN INDUSTRY CASE STUDY

Abstract

Ethics-based auditing (EBA) is a structured process whereby an entity's past or present behaviour is assessed for consistency with predefined ethics principles. Recently, EBA has attracted much attention as a governance mechanism with potential to help 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 chapter, I address that knowledge gap by providing insights from a longitudinal industry case study. Over 12 months, I 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 focused on proposing or analysing evaluation metrics or visualisation techniques, my 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 chapter constitutes the main body of empirical research conducted as part of this thesis and 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.

Note

This chapter is based on a journal article originally published in *AI and Ethics* (see Mökander & Floridi, 2022b).[39] While the chapter closely resembles the original article, I have expanded the methodology section. Please note that in this chapter I use 'AI systems' to refer to ADMS. AstraZeneca used the term AI systems internally, and adopting their terminology allowed me to maintain greater proximity to my observational data when reporting on this case study.

[39] The article on which this chapter is based was co-authored with my academic supervisor, Prof. Luciano Floridi. Please see Appendix 3 for an authorship statement.

4.1 Introduction

4.1.1 Background

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' (see e.g., Brown et al., 2021; Koshiyama et al., 2021; Raji et al., 2020). Chapter 3 in this thesis concerned what EBA is and how it works. In it, I defined EBA as a structured process whereby an entity's present or past behaviour is assessed for consistency with relevant principles or norms. I also argued that the promise of EBA is underpinned by three ideas. First, that procedural regularity and transparency contribute to good governance (Floridi, 2017b); second, that proactivity in the design of AI systems helps identify risks and prevent harm before it occurs (Kazim & Koshiyama, 2020); and third, that operational independence contributes to the objectivity and professionalism of the assessment (Raji et al., 2022).

Of course, the idea to audit software is not new. In fact, auditing ADMS for consistency with predefined requirements is a fundamental aspect of systems engineering (Leveson, 2011). More recently, however, Sandvig et al. (2014) and Diakopoulos (2015) helped popularise the idea that automated decision-making systems (ADMS) 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, and a range of EBA procedures have been developed (see e.g., Cobbe et al., 2021; ForHumanity, 2021; Zicari et al., 2021).

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 ADMS (ICO, 2020), and professional services firms like Deloitte (2020), EY (2018), KPMG (2020), and PwC (2019) have developed auditing (or 'assurance') procedures to help clients ensure that the ADMS they design and deploy are legal, ethical, and safe. In short, a new industry focusing 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 research. Raji and Buolamwini (2019) suggest that internal audits *can* help check that the engineering processes involved in designing ADMS meet specific standards. Similarly, Brundage et al. (2020) argue that external audits *can* help organisations verify claims about ADMS. 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 existing literature on EBA contains few case studies: Mahajan et al. (2020) conducted an audit of AI systems that replicate tasks in radiology workflows; and Buolamwini and Gebru (2018) assessed the efficacy of audits to address biases in facial recognition systems. Yet these audits were all designed and conducted by academic researchers. Hence, we still have only a limited understanding of how organisations in the private sector implement EBA.

4.1.2 Scope and contributions

To bridge that gap, this chapter seeks to address SQ2, i.e., how do organisations integrate EBA procedures with existing governance structures, and what challenges do they face in the process? To answer SQ2, I conducted a longitudinal industry case study. Over a period of 12 months, I observed and analysed the internal activities of AstraZeneca (a biopharmaceutical company) as it prepared for and underwent an ethics-based 'AI audit.'

This chapter describes and discusses the findings from that study to make two contributions to the existing literature. First, it provides a descriptive account of how (and why) a large, decentralised, and R&D-driven company like AstraZeneca implements EBA in practice. Second, by outlining the challenges and tensions involved in conducting a real-world EBA, it identifies transferable best practices for how to develop EBA procedures.

The findings from the case study suggest that the main difficulties organisations face when conducting EBA mirror well-known corporate governance challenges. 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 when designing and implementing EBA procedures. This chapter is thus aimed at researchers, auditors, and policymakers who design EBA procedures as well as industry practitioners tasked with the implementation of corporate AI ethics principles.

This chapter is structured as follows. Section 4.2 draws on previous research to establish the need for EBA. Section 4.3 introduces the case study by giving a descriptive account of AstraZeneca as an organisation and of the events leading up to the AI audit. Section 4.4 describes AstraZeneca's 2021 AI audit in greater detail, situating it relative to previous research on EBA. Section 4.5 describes the methodology used to conduct this study, which is based on participant observation and semi-structured interviews. Section 4.6 discusses the findings from the case study. Section 4.7 identifies the limitations of the approach taken in this chapter. Finally, Section 4.8 highlights current best practices and directions for future research.

4.2 The need to operationalise AI governance

AI holds great promise to support human development and prosperity (Dignum, 2020). Enabled by advances in machine learning (ML), access to increased computing power, the growing availability of data, and the ubiquity of digital devices, AI systems can improve efficiency, reduce costs, and help solve complex problems (Taddeo et al., 2018).

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 (Grote & Berens, 2020) and enable personalised treatments (Begoli et al., 2019). They also improve healthcare systems through better forecasting (Kaushik et al., 2020). In the pharmaceutical industry the combination of pattern recognition for molecular structures and laboratory automation promises faster drug discovery processes (Schneider, 2019). 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 (Jiang et al., 2017).

However, the use of AI systems in the healthcare sector is coupled with ethical challenges (Morley et al., 2020b; Topol, 2019). The use of AI systems may leave users vulnerable to discrimination and privacy violations (Leslie, 2019). It may also enable wrongdoing and erode human self-determination (Tsamados et al., 2020). Many of these risks apply to AI systems generally. But how AI systems process health data is particularly delicate (McLennan et al., 2022), since patients may be harmed by reputational damage and suboptimal care (Laurie et al., 2014). For example, recent studies have found racial biases in medical devices that provide pulse oximetry measurements (Sjoding et al., 2020).

While the adoption of AI systems has outpaced the development of governance mechanisms designed to address the associated ethical concerns (Taeihagh et al., 2021), abstaining from using AI systems in sensitive areas of application is not the way forward. As far as the use of AI in medicine is concerned, a 'precautionary approach' would cause significant social opportunity costs due to constraints that undermine the development of promising technologies, drugs, and treatments (Blasimme & Vayena, 2021). Moreover, AI systems are part of larger socio-technical systems that comprise other technical artefacts as well as human operators (Di Maio, 2014). No purely technical solution will thus be able to ensure that AI systems operate in ways that are ethically-sound (Schneider et al., 2020).

It is essential that 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 (Fjeld, 2020; Jobin et al., 2019).[40] These guidelines tend to converge on five principles: beneficence, non-maleficence, autonomy, justice, and explicability (Floridi & Cowls, 2019).[41] 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 (Mittelstadt, 2019). Moreover, translating principles into practice often requires trade-offs (Whittlestone et al., 2019a). Most critically, the industry lacks useful tools to translate abstract principles into verifiable criteria (Vakkuri et al., 2019).

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 (Ayling & Chapman, 2021; Morley et al., 2021b). These include *impact assessments lists* (AI HLEG, 2020; Koshiyama, 2019; Reisman et al., 2018), *model cards* (Mitchell et al., 2019), *datasheets* (Gebru et al., 2018; Holland et al., 2018), *human-in-the-loop* protocols (Jotterand & Bosco, 2020), *standards* and reporting guidelines for using AI systems (Cihon, 2019; Cruz Rivera et al., 2020; Liu et al., 2020), and the inclusion of broader *impact requirements* in software development processes (Prunkl et al., 2021).

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 (Cihon et al., 2021; Minkkinen et al., 2021). However, 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 (Floridi, 2019b). 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 (EIU, 2020a; Holweg et al., 2022).

How AI systems are designed and used is a concern not only for individual organisations but also for society at large (Floridi, 2021a). This insight has been reflected in recent regulatory developments. Both the EU Artificial Intelligence Act (AIA) (European Commission, 2021a) and the US Algorithmic Accountability Act of 2022 (AAA) (Office of

[40] Recent and influential contributions include AI HLEG (2019), IEEE (2019) and OECD (2019).

[41] Healthcare practitioners will note the overlap with the classical principles of bioethics (Dunn & Hope, 2018).

US Senator Ron Wyden, 2022) 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.'[42] 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 (Floridi, 2018). The 'ethics-based' approach studied in this chapter is thus compatible with – and complementary to – hard legislation.

4.3 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 (AstraZeneca, 2020a). As an R&D-driven organisation, AstraZeneca discovers and supplies innovative medicines. Its core business is using science and innovation to improve health outcomes through more effective treatment and prevention of complex diseases.[43] AstraZeneca has become a household name on account of the Oxford-AstraZeneca Covid-19 vaccine (Gilbert et al., 2021).

The biopharmaceutical industry has always been data-driven (Langkafel, 2015). To develop new treatments, researchers follow the scientific method by building and testing hypotheses about the safety and efficacy of various treatments.[44] 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 (Ashenden et al., 2021). 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. The company uses biological insight knowledge graphs (BIKG) to improve drug discovery and development processes (Crowe, 2020). Using BIKG helps synthesise and leverage prior knowledge to gain

[42] See Chapter 5 for an in-depth analysis of the role of auditing in the EU AIA.
[43] AstraZeneca is divided into three main therapy areas: Oncology; Cardiovascular, Renal and Metabolism; and Respiratory and Immunology diseases (AstraZeneca, 2021b).
[44] 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 (Ashenden, 2021).

new insights into disease characteristics and design smarter clinical trials (Vasetenkov, 2021). AI systems are also used for fast and accurate medical image analysis. Using AI systems based for image recognition cuts analysis time by over 30% and improves accuracy (AstraZeneca, 2021a). Moreover, AI systems help automate various tasks. For example, AstraZeneca use natural language processing to prioritise adverse event reports (Lea et al., 2021; Rizk et al., 2021). 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 the risks associated with AI systems. As discussed in Section 4.2, 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 below). 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 (AstraZeneca, 2020b).[45]

The primary aim of these ethics principles is to help employees and partners safely and effectively navigate the risks associated with AI systems.[46] 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 of the *ethics principles* strengthens AstraZeneca's brand.[47] Finally, the same internal processes that allow AstraZeneca to demonstrate adherence to its *ethics principles* also help it manage legal risks by 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 (Mittelstadt, 2019). Hence, AstraZeneca followed its commitment to its *ethics principles* by focusing on their implementation. However, doing so was not straightforward. AstraZeneca already had several related governance structures in place, e.g., regarding data management, QMS, CSR, sustainability, and product safety. Furthermore, AstraZeneca is a decentralised organisation in

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

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

[47] As noted by Slee (2021), 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.

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,*
3) The establishment of (i) an *AI resolution Board* and (ii) an internal *Responsible AI Consultancy Service,* and
4) The commissioning of an *AI audit* conducted by an independent party.

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

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

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 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 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

Second, a Responsible AI Playbook was developed to provide more detailed, end-to-end guidance on developing, testing, and deploying AI systems within AstraZeneca.[48] 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 the purpose of this thesis, AstraZeneca underwent an 'AI audit.' This audit constituted the research material for my case study, and framing it is the focus of the next section.

4.4 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 I 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. (Brown et al., 2021) 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

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

companies. The AI audit conducted within AstraZeneca corresponds to (ii) since it was directed towards demonstrating adherence to voluntary codes of conduct.[49]

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.[50] In the literature a distinction is often made between internal audits, based on self-assessment, and external audits conducted by expert organisations (Mantelero, 2018). The latter tend to be limited by reduced access to internal processes (Raji et al., 2020). However, involving external experts can address the confirmation bias that may prevent internal audits from recognising critical flaws (Bauer, 2016). By subjecting itself to external review, AstraZeneca thus got valuable feedback on how to improve its existing and emerging AI governance structures.[51]

As I argued in Chapter 3, 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.[52] However, transparency must always be understood in context, i.e., with regards to a specific audience and intended purpose (Larsson & Heintz, 2020). 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. It is worth noting that the subject of EBA can either be a process, an organisational unit, or a technical system.[53] AstraZeneca's AI audit focused on processes and people, i.e., on assessing the soundness and completeness of organisational processes and the extent to which different organisational entities adhered to these processes. During technical audits, in contrast, AI systems' source codes can be reviewed (Mittelstadt, 2016) or, alternatively, the behaviour (i.e., outputs) of such

[49] 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 (Kroll, 2021).

[50] 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.'

[51] Note that all the other three enterprise-wide activities conducted by AstraZeneca to operationalise AI governance (see Section 4.3) were internal in nature.

[52] As noted by Kroll (2021), 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.

[53] 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 (Powers & Ganascia 2021).

systems can be tested for a wide range of different input values (Kroll et al., 2016). However, no technical audits of individual AI systems were conducted during AstraZeneca's AI audit.

In Section 4.6, I discuss lessons learned from this audit. However, qualitative findings are best interpreted in the light of the context from which they emerged. Hence, before exploring the findings, the next section outlines how I collected and analysed the data.

4.5 Methodology: An industry case study

The aim of this chapter is to address SQ2, i.e., how do organisations integrate EBA procedures with existing governance structures, and what challenges do they face in the process? This is a descriptive question formulated to generate new qualitative knowledge about the organisational contexts into which feasible and effective EBA procedures must be integrated.

To address SQ2, I conducted an *industry case study* (Bass et al., 2018; Yin, 1994). Specifically, I observed and analysed AstraZeneca's internal activities as it prepared for and underwent an EBA in collaboration with an independent third-party auditor. The case study was *longitudinal* (Thomson et al., 2003) insofar as it lasted over 12 months (from November 2020 to December 2021). I chose to conduct a longitudinal case study because I wanted to capture not only the challenges AstraZeneca faced during the EBA but also the motivations different internal stakeholders had for structuring the audit one way over the other – as well as for conducting an EBA in the first place.[54]

To guide my data collection, I broke down SQ2 into three more specific questions:

- *How do industry firms integrate EBA within existing governance structures?*
- *What challenges do industry firms face when attempting to implement EBA?*
- *What are best practices for how to prepare for and conduct EBA?*

To answer these questions, I leveraged two qualitative research methods: participant observation and semi-structured interviews. However, before describing my methodology in greater detail, something should be said about what I mean by 'case study' in this context and why I chose to study AstraZeneca's EBA specifically.

[54] Similar longitudinal case studies have long been used to observe how different governance mechanisms impact organisational practices, see e.g., Jackall (2010).

4.5.1 Case study selection

In the social sciences, the term 'case study' is used in many different ways. For example, it is useful to distinguish between case studies used for pedagogical purposes and those conducted for research purposes, i.e., to generate new empirical data (Platt, 1992). My use of the term case study in this chapter corresponds to the latter. Further, a distinction is often made between *intrinsic* case studies, which aim to provide a better understanding of a particular organisation or group for its own sake, and *instrumental* case studies, which use particular cases to provide insights into a broader issue or to develop generalisations (Stake, 1995). In the former, a case is typically selected for its uniqueness, which is of genuine interest to the researcher (Crowe et al., 2011). However, my case study was instrumental insofar as it aimed to shed light on the broader question of how organisations implement EBA.

When selecting a case for an instrumental case study, several factors come into play. According to Crowe et al. (2011), the most important ones are *relevance, access,* and *ethical considerations*. Relevance means that the case should allow researchers to study the phenomenon they are interested in. This corresponds to what Merton (1987) refer to as strategic research material:

> *'[Strategic research materials are] research sites, objects or events that exhibit the phenomena to be explained or interpreted to such advantage and in such accessible form that they enable the fruitful investigation of previously stubborn problems and the discovery of new problems for further inquiry.'* (Merton, 1987, p.2)

With respect to my SQ, AstraZeneca's AI audit constituted 'strategic research material' for three reasons. First, AstraZeneca uses AI systems for a wide variety of tasks (e.g., to detect treatment response patterns and automate laboratory tasks), allowing me to study corporate AI governance in an applied setting. Second, as a biopharmaceutical company, AstraZeneca has a long history of following ethical standards. This meant that the practical challenges it faced with respect to governing AI systems overlapped with the theoretical problems I sought to address. Third, the timing was advantageous. As I started my DPhil research in the autumn of 2020, AstraZeneca had just decided that it would conduct an EBA the following year. That allowed me to observe the entire process – from the day AstraZeneca published its *ethics principles for data and AI* in November 2020 until the EBA was conducted during Q4 2021.

However, selected cases need not only be relevant but also hospitable to the inquiry (Stake, 1995). How to secure access to the research site is thus a central consideration in case study selection (Crowe et al., 2011). I gained access to AstraZeneca's processes, staff, and

documentation through an institutional agreement. My DPhil research at the Oxford Internet Institute (OII) had previously been awarded funding from AstraZeneca's doctoral scholarship program. Hence, AstraZeneca already had trust in the OII as an institution, and I could leverage the existing institutional agreement to pitch my research idea, which was well received.

As Darke et al. (1998) observed, organisations are more likely to open themself up to research that address questions they are interested in, or that they can otherwise benefit from. In my case, both criteria were satisfied. First, because AstraZeneca had just published its own *ethics principles for data and AI*, senior staff were interested in the question of whether and how EBA could help organisations design and deploy AI systems in ways that align with their organisational values. Second, from a communications perspective, AstraZeneca had an interest in being associated with research on how to design and use AI systems responsibly.

My supervisor and I considered alternatives. Amongst others, we had extensive dialogue with a German automotive company that was also due to conduct an EBA. However, they were planning to do an internal EBA, i.e., not one involving an independent third-party auditor. Also, we did not have the same stable institutional relationship to fall back on to secure access. In short, I picked AstraZeneca's EBA as a case study for my DPhil research both because it constituted strategic research material and because I was able to secure the required level of access to conduct my research. I discussed the ethical considerations of this choice in Section 1.8 and will return to reflect further on my positionality in Section 4.7. The remainder of this section focuses on how the research was conducted.

4.5.2 Data collection and analysis

As previously mentioned, my case study leveraged two qualitative research methods: participant observation and semi-structured interviews. Participant observation, in which research is carried out through the direct participation of the researcher in the situation under study, has a long history in organisational research (Vinten, 1994). It is a methodology particularly well-suited to making sense of organisational practices, framing problems, and evaluating outcomes (Woodside, 2016).

However, participant observation works best when researchers embed themself within an organisation for long enough to observe how it actually operates. For this reason, I observed AstraZeneca's R&D Data Office team – and their activities directed towards developing and implementing EBA procedures – over a period of 12 months. This meant partaking in weekly

meetings, reviewing working documents, and taking note of not only the actual development of EBA procedures but also the choices made along the way.[55]

I observed two types of meetings: *internal meetings*, in which AstraZeneca employees prepared, or evaluated the results from, the EBA, and *audit meetings*, in which external auditors asked questions to, and reviewed documentation provided by, AstraZeneca employees. I was invited to these meetings by AstraZeneca's internal audit team, which coordinated all EBA-related activities. Because AstraZeneca's employees are distributed internationally – and because of Covid-19 travel restrictions – all meetings took place online.

In addition, I conducted semi-structured interviews (Edwards & Holland, 2013) with different stakeholders involved in the audit. A list of the questions that guided my interviews in provided in Appendix 11. While semi-structured interviews take a broad set of questions assembled at the outset as a starting point, they allow for great flexibility in terms of follow-up questions. Another advantage of the semi-structured interview format is that the conversation is directed to the actual problem under investigation, as opposed to the preconceived interests of the researcher (Wang & Yan, 2012).

For these reasons, I did not interview a predefined list of people but instead used a snowballing technique (Given, 2008) to recruit new interviewees. Specifically, I reached out to both managers and software developers from different internal functions involved in the EBA, including the IT department and the R&D Data Office. This allowed me to follow up on themes emerging from regular audit meetings and explore different actors' motivations and perspectives. Further, I strove for a balance of different genders, ethnicities, and educational backgrounds amongst the interviewees. In total, 18 people were interviewed – some on several occasions. Each interview lasted 1–2 hours. To make the participants feel comfortable and avoid disturbing the flow of meetings, I did not record interviews, taking notes instead.

A separate NDA was signed with the external auditor, allowing me to join relevant meetings and study the entire process. In all meetings, participants were informed about my presence and the purpose of my research. No personal details were collected or stored during the research. Approval for this research was granted by the OII's departmental research ethics committee (research ethics approval reference number: SSH_OII_CIA_21_097).

The interviews were conducted and analysed in parallel with my ongoing participant observation. The aim of this parallel research design was to mitigate the risk of 'losing context'

[55] I also have online access to the (SharePoint) working folders of AstraZeneca's R&D Data Office team.

that is associated with qualitative research (Bryman, 2016). The data collected from both participant observation and the semi-structured interviews were imported, coded, and analysed in NVivo, in which I have previous experience. With 'code', I here refer to a word that assigns an essence-capturing attribute for a portion of language (Saldaña, 2009).

Because this is a descriptive case study, I used *thematic analysis* to capture the underlying meaning from the data collected interviews and participant observation. According to Braun and Clarke (2006), thematic analysis contains six steps: (1) get familiar with the data; (2) generate initial codes; (3) search for themes; (4) review themes; (5) define and name themes; and (6) produce a report by weaving the themes into a coherent narrative.

Following this procedure, I first familiarised myself with the data by rereading and transcribing my notes, before highlighting interesting and recurring topics or phrases in my data and assigning them initial codes. I then searched for common themes in my data and analysed how these themes relate to each other. In practice, this meant that I first collated my initial codes (like 'access' and 'data storage') into sub-themes (like 'data management') and, subsequently, clustered sub-themes into themes with the help of a provisional mind map. Having drafted this coding scheme, I then revised and refined the names and definitions of each theme to sharpen their descriptive precision (and reduce overlaps between themes).[56]

Through this coding process, nine themes emerged representing different challenges AstraZeneca employees faced when attempting to implement EBA procedures. The lessons learned from AstraZeneca's AI audit, which the next section reports on in narrative form, have been organised around these nine themes.

4.6 Lessons learned from AstraZeneca's 2021 AI audit

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

[56] See Appendix 12 for an overview of my code hierarchy, and Appendix 13 for a mind map of my initial codes, sub-themes, and themes generated in Nvivo.

4.6.1 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.[57] 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 (Nadler et al., 2021). 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.[58] However, even 'failed' R&D projects inform pharmaceutical innovation (Chiou et al., 2012). Hence, for AstraZeneca, risk per se is not undesirable. Rather, AstraZeneca's priority is to define and control the risk appetite in different projects. From an auditing perspective, this has two 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.

4.6.2 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 (Wang, 2019).[59] Nevertheless, every policy needs to define its

[57] Critically-oriented researchers often highlight AI systems' failures to stress the need for more regulation (Greene et al., 2019). In contrast, techno-optimists point towards the gains such systems bring and caution against red tape (Diamandis & Kotler, 2012).

[58] Pammolli et al. (2020) 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.

[59] Some researchers use the term AI to refer to a type of agents that display some levels of autonomy, adaptability, and problem-solving capacity (Legg & Hutter, 2007). Others take AI to demarcate the set of computational techniques designed to approximate cognitive tasks (US Defence Authorization Act, 2018). Yet others use the term to describe the science and engineering of making specific machines (McCarthy, 2007).

material scope (Schuett, 2021). 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 (Baum, 2017) and partly because ethical tensions can sometimes be intrinsic to the decision-making tasks at hand (Danks & London, 2017).[60]

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)

To solve this tension, AstraZeneca did not try to define what AI *is*.[61] 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.[62] 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

[60] As Bryson (2021) argues, problems associated with 'AI' have not so much been created as exposed by it.
[61] 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.
[62] A parallel can be made to the EU AIA, which also takes an explicitly risk-based approach to AI governance.

makes autonomous decisions without human judgement. The approach taken is pragmatic[63] 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, AI auditors depended on the business to identify and select relevant projects and systems for the in-depth audits.

The main takeaway 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.

4.6.3 Harmonising standards across decentralised organisations

A further challenge faced during AstraZeneca's AI audit was 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) only 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 decentralised, since collaborating with external partners has many advantages for them, including the possibility to leverage local market knowledge and health data.

These operational and structural differences between business areas are reflected in their capacities to manage AI-related risks. Different EBA procedures are thus needed to assess each business area's governance structure.[64] For example, AstraZeneca's AI audit showed that business areas understood risk differently. Within R&D, many employees work directly with

[63] Pragmatic problem-solving demands that things should be sorted so that their grouping will promote successful actions for some specific end (Dewey, 1920).

[64] 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.

patients and patient data. Hence, they typically see patient-centric risks. As one of the 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 (CDEI, 2021b). Using this distinction, current best practice would demand 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.

4.6.4 Internal communication as a key to operationalising AI governance

My observations of AstraZeneca's AI audit suggest that internal communication and training efforts are central to operationalising corporate AI governance. In AstraZeneca's case these communication efforts were continuous and happened on several different levels. For example, the ethics principles were agreed upon through a bottom-up process that included 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. 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 (Alshammari & Simpson, 2017; Floridi & Cowls, 2019) and (ii) the tools enabling managers and software developers to translate these principles into practice (Ayling & Chapman, 2021; Morley et al., 2020a). While these aspects remain important, my 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, my findings suggest three best practices. First, communication concerning AI governance is most effective when supported by senior executives.[65] 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.

4.6.5 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 (Crawford, 2021). 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 (BenevolentAI, 2019). Similarly, in 2021, AstraZeneca launched a collaboration with American healthcare company GRAIL to evaluate the effectiveness of early cancer detection technologies (GRAIL, 2021).

[65] This finding is supported by previous research. For example, Gasser and Schmitt (2021) have shown that the effectiveness of corporate governance mechanism depends on issues related to leadership, values, and culture.

External R&D collaborations offer numerous advantages.[66] However, such collaborations are 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)

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.

4.6.6 Ethics-based auditing as a catalyst for internal change

There are many reasons why organisations subject themselves to EBA. For example, such audits can help to control technology-related risks and inform AI design choices. However, my observations from AstraZeneca's AI audit suggest that organisations also have 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. However, as an organisation's technological resources evolve, old governance structures risk

[66] External R&D collaborations benefit innovation by increasing efficiency, reducing costs, and granting access to valuable resources not available internally (Grimpe & Kaiser 2010).

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 initial formulation of the ethics principles. Yet during roll-out, a more decentralised structure emerged, with each business area responsible for practically implementing the principles.

The point I 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.

4.6.7 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 (Kroll,

2018). Similarly, third-party auditors can make verifiable claims about AI systems without accessing the underlying data and computational models by analysing publicly available information (Dash et al., 2019).

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 (Pasquale, 2016), that facilitates the process of conducting AI audits. Moreover, EBA procedures focusing 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 (Chopra & Singh, 2018).[67]

Nevertheless, it is important to remain realistic about what EBA procedures focusing on organisational processes can be expected to achieve. Such procedures can verify claims about technology providers' QMS 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.

4.6.8 Measuring progress and demonstrating success

Social phenomena are increasingly measured, described, and influenced by numbers,[68] and the corporate governance field is no exception. Since 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 (Cugueró-Escofet & Rosanas, 2017; Islam & Greenwood, 2021).[69] Such metrics are relevant for EBA for two reasons. First, organisations investing in AI governance want to be able to point 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.

---

[67] According to Diakopoulos (2021), 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.

[68] See Mau (2019)for an excellent account of the growing tendency to quantify the social world and how that process changes our assignment of worth.

[69] Note that organisational performance metrics need not be based on financial measures alone. The perhaps most famous example of this is 'the balanced scorecard' (Kaplan & Norton, 1996).

Recently, much literature has focused on measuring and assessing the performance of different AI systems along normative dimensions such as fairness, transparency, and accountability (Hoffmann et al., 2018). For example, Wachter et al. (2021) 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 (Bellamy et al., 2019; Cabrera et al., 2019).

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 (Greenfield, 2017). Moreover, as Lee et al. (2021) 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 (Mittelstadt, 2019). Similarly, different definitions of fairness – like individual fairness and demographic parity – are mutually exclusive (Kusner et al., 2017; Verma & Rubin, 2018)

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 (Madaio et al., 2020). Yet simply having a checklist is insufficient to ensure that AI systems are designed and used ethically and safely (McNamara et al., 2018) and previous research has found that checklists risk reducing auditing to a box-ticking exercise (Raji et al., 2020).

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.[70] 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

[70] 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 (Floridi & Strait 2020).

evidence can we use to determine whether an AI system 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, my 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 (Kearns & Roth, 2020). Thus, a fruitful avenue for future research would be to develop a guide on when and how to use different metrics in the software development lifecycle and as part of holistic EBA procedures.

4.6.9 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 includes *initial costs* (e.g., time and resources invested in preparing for the audit as well as the procurement of audit services 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).[71]

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.

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

During the AI 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 documents and answer detailed questions during interviews. Taken together, around 2,000 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 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 (Renda et al., 2021). While those numbers have been supported by independent researchers (Haataja & Bryson, 2021), 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 (Mueller, 2021).

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 cost 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 (Dawson et al., 2010). This suggests that – despite the associated costs – businesses have clear incentives to design and implement effective EBA procedures.

4.7 Limitations and reflections

Conducting qualitative research is challenging and bound to result in methodological shortcomings (Miles & Huberman, 1994). Here, I discuss important limitations with regards to the validity, independence, and generalisability of the findings.

Consider validity first. Since this 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 (Smith, 2014). For example, it is difficult to assess the ethical risks posed by specific AI systems. Therefore, I exclusively focused 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 (Wolf, 2011). 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 (Levendusky, 2013).

Another limitation concerns the independence of the research. As mentioned, my doctoral 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 (Maruyama & Ryan, 2014). To manage this risk, I communicated clear boundaries regarding my role as an independent researcher. I also followed best practices in research ethics, e.g., informing all parties about the constructively critical nature of my work.

That said, it is useful to reflect a bit more about my positionality and how my presence influenced the process I studied. It was not easy – neither for me nor for AstraZeneca's staff – to separate my role as an independent researcher studying the EBA from my role as an academic with relevant expertise who happened to be in the room as real-world challenges were being addressed. Over the course of my research, I have accumulated some knowledge of how other organisations design their EBA procedures. Hence, I was often asked for advice by managers and software developers at AstraZeneca on how to approach specific issues, like how to demarcate the material scope of the EBA or what evaluation metrics to use during the audit. As a researcher, I did not want to engage directly with such questions, since my role could then have drifted into becoming a member of the team conducting the audit. However, as a human, I felt a desire to help. In the end, I came up with two strategies to manage this tension. First, while I did not give my opinion on how to address specific issues, I did provide direction to different sources where AstraZeneca could learn more about available policy design options. Second, I noted down questions that were of mutual interest to me and AstraZeneca, to be explored after the observational case study had been completed. For example, Chapter 6 in this thesis is the result of an industry-academia research collaboration between me and AstraZeneca that was informed by the findings of this case study and launched after its completion.[72]

[72] See also Mökander et al (2022b) *Challenges and Best Practices in Corporate AI Governance*.

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 (Morgan et al., 2016). Moreover, data controllers (like AstraZeneca) have an interest in not disclosing trade secrets (Flyvbjerg, 2001). I sought to reduce the risk that biased input distorts the analysis by triangulating the information provided by AstraZeneca employees with other sources. Still, the findings from the case study should not be treated as neutral, but rather as context-specific knowledge (Jackall, 2010).

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 BMW Group (2020), IBM (Cutler et al., 2018), Google (2018), and Microsoft (2019). Having published a series of articles on EBA, I am often approached by technology providers and auditors who are looking to conduct EBA. In some cases, they seek my advice on how to structure EBA and, in other cases, a more active collaboration. Because of time constraints, I have been very selective about which projects I engage with. That said, have always tried to make room for initial meetings with different actors to explore the motivations they have and the challenges they face. Over the last three years, I have therefore met with numerous industry firms seeking to implement EBA, professional service providers that offer EBA services, and policymakers seeking to enable the emergence of feasible and effective EBA procedures. My experience from these interactions is that most organisations seeking to develop or implement EBA procedures face similar challenges. The findings presented in Section 4.6 will thus be (at least in part) relevant to other large corporations attempting to integrate EBA procedures within existing governance structures.

4.8 Discussion

A new industry that focuses on auditing AI systems is emerging. The proposed EU AIA, which sketches the contours of a professionalised AI auditing ecosystem (more on this in Chapter 5), is likely to accelerate this trend. In such a fast-moving and high-stakes environment, it is essential that policymakers and industry practitioners understand the conditions under which EBA is a feasible and effective mechanism for operationalising AI governance. The findings from my industry case study helps further such an understanding.

To start with, my case study has illustrated that different EBA procedures serve different purposes. Process audits – such as that undertaken by AstraZeneca – are well suited

to verifying claims about the QMS 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, 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, my 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 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 – thereby failing 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.

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 multinational 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. As illustrated by AstraZeneca's difficulty to establish the material scope of their EBA, questions as to which systems and processes AI governance frameworks ought to apply to remain unanswered. 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. To 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 risky) development projects to

be outsourced. This is akin to what (Floridi, 2019b) 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 governance 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 legal theory as well as political and 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 chapter 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.[73]

4.9 Concluding remarks

As mentioned in Chapter 1, the purpose of this thesis is to better equip societies to reap the benefits of ADMS while managing the associated risks by exploring whether and how EBA can help organisations design and deploy AI systems in ways that align with their organisational values. This chapter has furthered that purpose by providing new qualitative knowledge about the organisational context in which EBA procedures must be integrated to be feasible and effective in practice. It has also filled an important gap in the existing literature. To the best of my knowledge, the case study of AstraZeneca's EBA reported on in this chapter constitutes the first study in which an independent researcher has been able to observe the internal activities of a technology provider before, during and after an EBA.

Focusing on the descriptive level of my research, this chapter set out to explore:

SQ2 How do organisations integrate EBA within existing governance structures, and what challenges do they face in the process?

[73] This conclusion reiterates findings from previous research. See e.g., Raisch & Krakowski (2021).

While a single case study cannot provide a conclusive answer to this question, the findings of my empirical research offer several clues. To begin with, my observational data indicated that technology providers have strong incentives to subject themselves to EBA. In AstraZeneca's case, the motivating factors included (i) the need to manage financial, legal, and reputational risks, (ii) the competitive pressures forcing technology providers to continuously improve their QMS, and (iii) the desire of individuals to design and use ADMS responsibly.

At the same time, my findings suggest that technology providers will face several challenges when seeking to implement EBA procedures and integrate them with existing governance structures. For example, AstraZeneca struggled to harmonise standards across a decentralised organisation, demarcate the material scope for ADMS governance, define key performance indicators for AI systems, and act on the results produced by EBA. In each case, the challenges faced mirrors well-known corporate governance challenges. This points towards a critical gap in the existing literature, which has focused on developing technical tools or step-by-step procedures for how to audit AI systems. The findings presented in this chapter, however, suggest that the main bottleneck to implementing EBA procedures is not a lack of tools but that these are not being employed in a rigorous and structured manner due to economic, social, and organisational factors.

The foregoing conclusion suggests that there are indeed serious difficulties involved in designing and implementing feasible and effective EBA procedures. However, the picture is not all bleak. When engaging in pragmatic problem solving, the first step is to correctly identify and adequately describe the problem at hand (Prasad, 2020). The implementation challenges identified and described in this chapter will thus form the basis for my applied research in the remaining chapters of this thesis. For example, Chapter 6 – which explores how the material scope of ADMS governance can be demarcated – was written in response to my empirical findings in this chapter.

Finally, not only the difficulties different actors face but also their motivations inform applied research. As I have shown in this chapter, one of the drivers behind AstraZeneca's decision to conduct an EBA was to anticipate forthcoming legislation. A potential objection to EBA would thus be that it will no longer be needed once hard legislation – like the AIA – comes into force. To assess the strength of that objection, the next chapter will consider SQ3, i.e., how can EBA complement legislative approaches to managing the risks ADMS pose?

CHAPTER 5

CONFORMITY ASSESSMENTS AND POST-MARKET MONITORING: THE ROLE OF AUDITING IN THE EU AIA

Abstract

The European Artificial Intelligence Act (AIA) is the first general legal framework for AI proposed by any major global economy. As such, it is likely to become a point of reference in the larger discourse on how AI systems can (and should) be regulated. In this chapter, I describe and discuss the two primary governance mechanisms proposed in the AIA: the conformity assessments that providers of high-risk AI systems are expected to conduct, and the post-market monitoring plans that providers must establish to document the performance of high-risk AI systems throughout their lifetimes. In doing so, I argue that the AIA can be interpreted as a proposal to establish a Europe-wide ecosystem for conducting AI auditing, albeit in other words. Subsequently, I offer three main contributions in this chapter. First, by framing the governance mechanisms included in the AIA in terminology from existing literature on AI auditing, I help providers of AI systems understand how they can demonstrate adherence to the requirements it sets out. Second, by examining the AIA from an auditing perspective, I provide transferable lessons from previous research and identify areas of the AIA in which further revisions or clarifications are needed. Third, by showing that the AIA also encourages technology providers to adopt voluntary codes of conduct even for non-high-risk AI systems, I demonstrate that voluntary, ethics-based, auditing procedures are compatible with, and complementary to, the governance mechanisms included in the AIA.

Note

This chapter is based on a peer-reviewed journal article published in *Minds and Machines* (see Mökander et al., 2022a).[74] In this chapter, I use the term 'AI systems' to refer to ADMS to reflect the vocabulary used by the European Commission in the AIA. When adapting the text into a thesis chapter, I took the opportunity to include references to amendments to EU AIA that have been submitted since my original article was first published.

[74] The article was co-authored with Maria Axente, Federico Casolari, and Luciano Floridi. Please see Appendix 3, 4, and 5 for an authorship statement.

5.1 Introduction

5.1.1 Background

On 21 April 2021, the European Commission published its proposal for a new Artificial Intelligence Act (AIA).[75] The AIA builds on several recent initiatives and publications that collectively have foreshadowed EU legislation on AI. For example, in the *Ethics Guidelines for Trustworthy AI*, AI HLEG (the European Commission's High-Level Expert Group on AI) stipulated that AI systems should be ethical, lawful, and technically robust. Building on these guidelines, the European Commission (2020b) subsequently published a *White Paper on AI*, in which the risk-based approach to AI governance that permeates the AIA was first outlined. Also significantly, the AIA is supposed to constitute a core part of the EU digital single market strategy; indeed, it aims at ensuring the proper functioning of the internal market by setting harmonised rules on the development and use of products and services that make use of AI technologies or are provided as stand-alone AI systems within the Union market. In short, the AIA is a natural continuation of what can be called an EU approach' to AI governance.[76]

The AIA has attracted much attention from policymakers, regulators, commentators, and businesses across the globe. It is the first attempt by any major economy to elaborate a general legal framework for AI. It is also expected to have a significant impact outside the EU's borders. This impact would be both direct, because the AIA applies to any AI system used in the EU irrespective of where providers are placed (AIA: Article 2); and indirect, because of the 'Brussels effect' (Bradford, 2012, 2020), whereby multinational organisations choose to harmonise all their international practices with EU laws.[77] Hence, in addition to constituting proposed legislation in its own right, the AIA is also likely to become a point of reference in the larger discourse on how AI systems can (and should) be regulated.

5.1.2 Scope and contribution

The initial reactions to the AIA have been many and disparate. I will return to highlight some of the points raised by different commentators. However, the purpose of this chapter is neither

[75] Its full name reads '*Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AIA) and amending certain Union legislative acts.'*

[76] Since the AIA's publication, this European approach to AI governance has been further extended by the *Digital Services Act* (European Parliament, 2022a) and the *AI Liability Directive* (European Parliament, 2022b) .

[77] Here, a parallel can be made to the (GDPR) (European Parliament, 2016) which, together with the *California Consumer Privacy Act* (2018), has become a de facto global standard for data regulation (Barrett, 2019).

to provide a general commentary on the AIA nor to review the initial reactions to the AIA from different stakeholders and interest groups. Instead, this chapter focuses on a more specific yet crucially important question: what is the role of auditing in the AIA?

The AIA only makes limited explicit references to auditing. However, in this chapter, I use the term 'auditing' in a broad sense to refer to structured processes whereby an entity's behaviour is assessed for consistency with relevant principles, standards, regulations, or norms (see Chapter 2). So understood, auditing encapsulates several governance mechanisms proposed in the AIA, including the 'conformity assessments' (AIA: Article 43) that providers[78] of high-risk AI systems are expected to conduct and the 'post-market monitoring plans' (AIA: Article 61) that providers must establish to document and analyse the performance of high-risk AI systems throughout their lifetimes. The AIA can thus be interpreted as a proposal to establish a Europe-wide ecosystem for conducting AI auditing, albeit in other words.

On a few occasions, the AIA does refer explicitly to auditing. However, these references are mostly to be found in the annexes. For example, paragraph 5.3 in ANNEX VII reads as follows: 'The notified body shall carry out periodic *audits* to make sure that the provider maintains and applies the QMS and shall provide the provider with an *audit* report.' Naturally, sentences like this cannot be understood except as part of the AIA as a whole since it requires an understanding of what exactly is meant by *notified body*, *provider*, and *QMS* in this context. In this chapter, I hope to contribute to such a clarification.

Understanding what role audits are expected to play in the proposed EU legislation is important for three reasons. First, organisations that design and deploy AI systems need clarity on how they can prove adherence to the rules laid out in the AIA. From a practical perspective, the questions thus centre around operational aspects like:

- *Material scope*: *what is being subject to evaluation?*
- *Normative baseline: according to which metrics are AI systems being evaluated?*
- *Procedural regularity: what are the roles and responsibilities of different stakeholders throughout the auditing process?*

The first goal of this chapter is to shed light on these operational questions and, thereby, help organisations interpret – and adapt to – the proposed EU legislation on AI.

[78] 'Provider' means a natural or legal person, public authority, or other body that develops AI systems or that has an AI system that it plans to place on the market, whether for payment or free of charge (AIA: Article 3).

The second goal is to inform the ongoing policy development process. By analysing the role of auditing in the AIA, I seek to anchor the proposed EU legislation in the vast and growing academic literature on AI auditing. To do so, I conduct a gap analysis, comparing the AIA's provisions with best practices for how to audit AI systems. Resulting from this analysis are seven recommendations on how to further refine the AIA (see Section 5.7).

The third goal of this chapter is to contribute to an improved understanding of what room the AIA leaves for self-regulation and co-regulation. As illustrated by AstraZeneca's AI audit reported on in Chapter 4, a new industry is already emerging whereby professional service providers offer ethics-based auditing (EBA) services to help organisations design and deploy AI systems in ways that align with voluntarily adopted codes of conduct or other ethics-principles. By analysing the role of auditing in the AIA, this chapter thus addresses SQ3, i.e., how can EBA complement legislative approaches to managing the risks AI system pose?

Before proceeding any further, it should be acknowledged that the AIA is a *proposal*. As such, it has been and will continue to be subject to negotiations and changes.[79] In fact, several amendments been submitted to the AIA since the first draft was published. For example, in November 2021 the Council of the EU shared a first compromise text on the AIA, which included changes to the definition of high-risk AI systems (Bertuzzi, 2021), and in May 2022, the French Presidency of the Council proposed that the AIA should be revised to better address the governance challenges posed by 'general purpose AI systems', like large language models (LLMs) (FLI, 2022). Most recently, in December 2022, the Council adopted its common position on the AIA. However, the proposal is currently under discussion in the European Parliament and the AIA will only become law once the Council and the Parliament agree on a common version of the text. My analysis in this chapter is based on the original draft.[80] However, where appropriate I have inserted footnotes to more recent amendments.

The remainder of this chapter proceeds as follows. Section 5.2 provides a high-level summary of the proposed EU legislation and the societal challenges that it attempts to address. Section 5.3 reviews previous research to define what I mean by 'governance mechanisms' and 'AI auditing' in this context. Section 5.4 describes and analyses the two governance

[79] Even after the European Parliament has given their approval, the AI act will have to pass through interinstitutional negotiations (trialogue). In the case of the GDPR, the process from first draft to becoming binding took over four years (EDPS, 2023).

[80] All references this chapter makes to page and article numbers of the AIA similarly refers to the original draft that was published by the European Commission on 21 April 2021, unless otherwise specified.

mechanisms currently included in the AIA that can also be understood in terms of AI auditing: conformity assessments and post-market monitoring. Section 5.5 describes the roles and responsibilities assigned to different actors at a corporate, national, and Union levels in the AIA. In doing so, it sketches the contours of an emerging European AI auditing ecosystem. Section 5.6 analyses the scope for EBA within the framework provided by the AIA. The focus is on the codes of conduct to which, according to the AIA, providers of non-high-risk AI systems are encouraged to adhere voluntarily. Section 5.7 moves beyond what is explicitly proposed in the AIA and provides a gap analysis that identifies areas omitted in the current proposal or where further clarification may help. Finally, Section 5.8 concludes that, while it constitutes a step in the right direction, the AIA could benefit from incorporating some lessons from previous research on auditing. These include, amongst others, translating vague concepts into verifiable criteria and strengthening the institutional safeguards concerning conformity assessments based on internal checks.

5.2 The Artificial Intelligence Act: A risk-based approach

The AIA represents the most ambitious attempt to regulate AI systems to date. It seeks to ensure that AI systems used by – or affecting – people in the EU are safe and respect existing laws and Union values. The scope of the AIA also includes the use of AI systems by EU institutions, bodies, and agencies (AIA: Recital 12). To prevent risks and harm to public interests and rights that are protected by Union law, the AIA proposes extensive documentation, training, and monitoring requirements on the AI systems that fall under its purview.

However, establishing safeguards against potential harms is not the only objective of the proposed EU legislation. The AIA also stresses that AI systems can support socially and environmentally beneficial outcomes and provide critical competitive advantages to European companies and economies. This claim is well-supported by previous research. For example, AI systems can improve efficiency and consistency in decision-making processes and enable new solutions to complex problems (Taddeo & Floridi, 2018). However, the same elements and techniques that power the socio-economic benefits of AI also bring about new risks for individuals and societies. Specifically, the combination of relative autonomy, complexity, and adaptability, underpins both beneficial and problematic uses of AI systems (Dignum, 2017; Floridi & Sanders, 2004; Russell & Norvig, 2015).

As a result, the capacity to manage the risks AI systems pose is becoming a prerequisite for good governance. The widespread repercussions AI governance has on other policy areas

are demonstrated by the fact that the AIA is closely linked to, and coherent with, other initiatives like the General Product Safety Directive (European Parliament/Council, 2001).

Well aware of this dynamic, the AIA takes as its starting point the twin objectives of promoting the uptake of AI systems and addressing the governance challenges they pose. This 'balanced approach' (AIA: p. 3) has been criticised both by those who contend that the AIA will ultimately stifle innovation (Dechert, 2021), and by those who argue that it leaves Big Tech virtually unscathed and that too little attention is paid to algorithmic fairness (MacCarthy & Propp, 2021). However, I disagree. As I shall argue in the following pages, the AIA is, on the whole, a good starting point to ensure that the development of AI in the EU is ethically sound, as well as environmentally and economically sustainable.

Any attempt to govern AI systems implies making a wide range of design choices. These choices are often difficult and require trade-offs. For example, every regulation needs to define its *material scope* (Schuett, 2021). While there is no commonly accepted definition of AI[81] (Buiten, 2019; Wang, 2019), the definition of AI systems originally proposed in the AIA is broad by any standard (CDEI, 2021a; Gallo et al., 2021).[82] Hence, the AIA is likely to capture decision-making systems that have been in place for decades, in ways that may be problematic.[83] On the one hand, a broad scope of application may prove to be more permanent, since it does not hinge on technical features which are likely to change in the near future. On the other hand, a broad definition risks being over-inclusive, applying to cases that do not need regulation with respect to the regulatory goal, adding unnecessary financial and administrative costs. Such burdens may, in turn, undermine the legitimacy of the regulation.

Trying to offset this risk, the AIA proposes different types of obligations for different types of AI systems. In fact, the most distinguishing characteristic of the proposed EU legislation is its proportionate, risk-based approach.[84] Simplified, the AIA clusters AI systems

[81] According to John McCarthy (2007), AI can be understood as the science and engineering of making intelligent machines. Machine learning (ML), i.e., the study of computer algorithms that can improve automatically through experience and by the use of data (Mitchell, 1997), could thus be viewed as a subset of AI.

[82] For the purpose of the proposed European legislation, the term 'AI system' refers not only to machine learning techniques but also to a wide range of statistical approaches (AIA: ANNEX I).

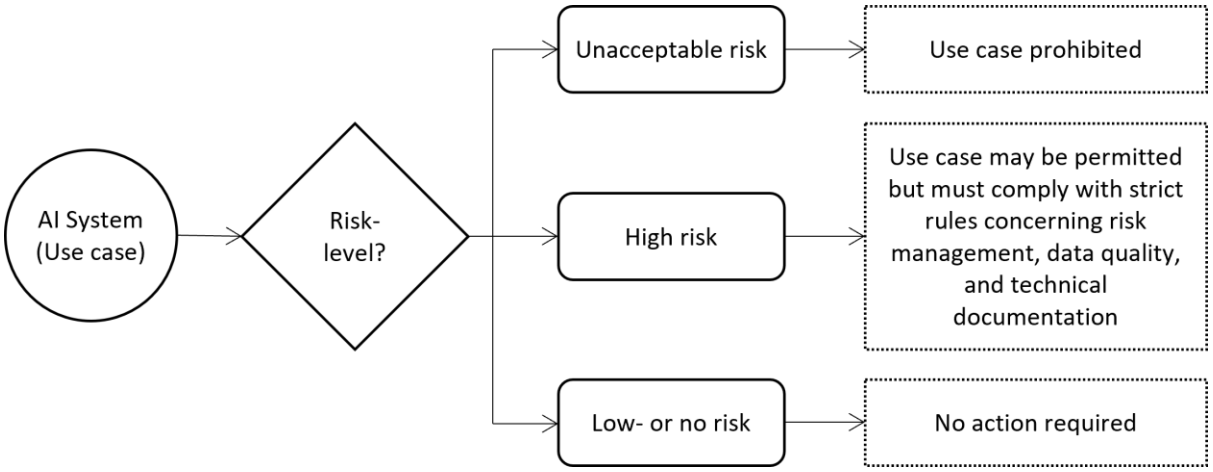
[83] The material scope has been subject to much debate (Bryson, 2022), and recent reports suggest that the European Parliament may settle for a narrower definition of AI systems, as proposed in the latest compromise text adopted by the Council (Bertuzzi, 2023).

[84] This risk-based approach can be traced back to publications like the White Paper on Artificial Intelligence (European Commission, 2020b) and the Recommendation of the Data Ethics Commission (DEK, 2018).

into three risk levels:[85] AI systems that pose 'unacceptable risk', 'high risk', and 'little or no risk' (AIA's Explanatory Memorandum: p. 12). The governance requirements differ between the three risk levels. AI systems that are deemed to pose an unacceptable risk, e.g., by posing a clear threat to people's safety, will be straight out banned. This includes the prohibition of AI systems used for general-purpose social scoring and real-time remote biometric identification of natural persons in public spaces for law enforcement.

In contrast, AI systems that pose little or no risk are not subject to any interventions stipulated in the AIA, exempt from some specific transparency obligations.[86] According to the AIA, the vast majority of AI systems are expected to fall into this category. However, in between these two extremes, there are a wide range of so-called 'high-risk' AI systems, that will be subject to strict obligations before they may be put on the market. To ensure a consistent level of protection from all high-risk AI systems, a common normative standard has been established. That standard is based on the EU Charter of fundamental rights and shall be non-discriminatory and in line with the EU's international trade commitments (AIA: Recital 13). Figure 9 below provides a simplified illustration of the risk-based approach.

Figure 9. The risk-based approach to AI governance proposed in the AIA.



The requirements for high-risk AI systems include the establishment of a risk management system, the identification and mitigation of known and foreseeable risks, an adequate testing

[85] When determining the risk-level, factors taken into account include the intended purpose of the system, the extent to which the system is likely to be used, and the potential for harm or adverse impacts (AIA: Article 7).
[86] For example, when using a chat bot, users should be made aware of the fact that they are interacting with a machine, rather than a human operator (AIA: Article 52).

and validation procedures (AIA: Chapter 2 of Title III). However, the AIA does not define rules for specific technologies. Instead, it seeks to establish processes for identifying those use cases requiring additional layers of governance to support specific policy goals. For example, the AIA demands that the technical documentation accompanying a high-risk AI system shall include 'a general description of its intended purpose' as well as 'a detailed description of the key design choices and assumptions made in the development process' (AIA: ANNEX IV).

While such measures contribute to procedural regularity and transparency, they also leave significant room for providers to develop and pilot new AI systems. A parallel can be made to what Loi et al. (2020) called *transparency as design publicity*, whereby organisations that design or deploy AI systems are expected to publicise the intentional explanation of the use of a specific system as well as the procedural justification of the decision it takes.

In Section 5.4, I will discuss how the AIA requirements on high-risk AI systems relate to AI auditing. However, to do so I must first clarify what is meant by AI auditing in this context. Thus, the following section provides a brief overview of previous research on AI governance in general and AI auditing in particular.

5.3 Previous research: AI governance mechanisms and AI auditing

To be successfully implemented, every regulation needs to be linked to effective governance mechanisms, i.e., activities, structures, and controls wielded by various parties to influence and achieve normative ends (Baldwin & Cave, 1999). Responding to the growing need for AI governance, a wide range of governance mechanisms have been developed that organisations can employ to ensure that the AI systems they design and deploy are legal, ethical, and technically robust. Some governance mechanisms focus on embedding ethical values into AI systems through proactive design (IEEE, 2019). Others are akin to what the (CDEI, 2021a) calls 'assurance techniques.' These include, amongst others, algorithmic impact assessment (ECP, 2018) and certification of AI systems (Scherer, 2016).

The proposed EU legislation on AI includes several governance mechanisms. Most notably, the providers of high-risk AI systems that fail to comply with the requirements stipulated in the AIA risk hefty fines. For example, non-compliance with the prohibition of specific uses of AI systems may subject providers to fines of up to 30,000 EUR, or 6% of their total annual turnover, whichever is higher (AIA: Article 71). However, before determining whether a specific AI system is legal, one must consider which mechanisms are available to establish its behaviour and performance (i.e., what it is doing at all).

This is where auditing comes in: auditing can be understood as a mechanism that helps organisations verify claims about the AI systems that they design and use. Building on previous work (Brundage et al., 2020), I define auditing as a structured process whereby an entity's present or past behaviour and performance is assessed for consistency with relevant principles, regulations and norms.[87] Note that while Brundage et al. focused on *organisational* audits, I stress that the entity in question, i.e., the subject of the audit, can be a person, an organisational unit, or a technical system.[88] Importantly, these different types of audits are not mutually exclusive but rather crucially complementary. To see that this is so, one need only consider the AIA, wherein some legal requirements concern the conduct of organisations that provide AI systems,[89] whereas others concern the technical properties of specific AI systems.[90]

Auditing differs from merely publishing a code of conduct because its primary goal is to show adherence to a predefined baseline (ICO, 2020). So understood, auditing has a long history of promoting trust and transparency in areas like financial accounting and safety engineering, as shown in Chapter 2. Concerning AI governance, auditing can be employed for several distinct yet related purposes. For example, Brown et al. (2021) noted that AI auditing could be used (i) by regulators to assess whether a specific AI system meets legal standards; (ii) by providers or end-users of AI systems to mitigate or control reputational risks; and (iii) by other stakeholders (including customers, investors, and civil rights groups) who want to make informed decisions about the way they engage with specific companies or products. The main takeaway is that all the above-listed applications of AI auditing align with – and have the potential to support – the stated objectives of the proposed EU legislation.[91]

5.4 Conformity assessments and post-market monitoring in the AIA

From an auditing perspective, two governance mechanisms in the AIA are especially relevant. First, the conformity assessments that providers need to conduct before putting high-risk AI

[87] A systems' behaviour and performance cover both 'what' it does and 'how' it does it.

[88] Different stakeholders are accountable for different steps in the process of developing AI systems. Hence, not only software developers and operators, but also managers and downstream users, could be subjected to EBA.

[89] For example, AI providers will be obliged to provide meaningful information about their systems and the conformity assessments carried out on those systems (AIA's Explanatory Memorandum: p. 12).

[90] AI systems should be resilient against risks connected to the limitations of the system and against malicious actions that may result in harmful or otherwise undesirable behaviour (AIA's Explanatory Memorandum: p. 30).

[91] In addition to ensuring that AI systems are safe and respect existing laws, the objectives of the AIA include facilitating investments, innovation, and – as already mentioned – the development of a single European market (AIA's Explanatory Memorandum: p. 3).

systems on the market (AIA: Article 43) and second, the post-market monitoring plans that providers shall establish to document the performance of high-risk AI systems throughout their lifetimes (AIA: Article 61). In this section, I will consider these in turn.

5.4.1 Conformity assessments

In line with the AIA's risk-based approach, high-risk AI systems are only permitted on the EU market if they have been subjected to (and withstood) an ex-ante conformity assessment.[92] Through such conformity assessments, providers can show that their high-risk AI systems comply with the requirements set out in the AIA. Once a high-risk AI system has demonstrated conformity with the AIA – and received a so-called CE marking – it can be deployed in, and move freely within, the internal EU market (AIA: Article 44).

There are three different ways in which these conformity assessments can be conducted. Which type of conformity assessment is appropriate in a specific case depends on the nature of the high-risk AI system. Consider first the many high-risk AI systems used as safety components of consumer products that are already subject to third-party ex-ante conformity assessments under current product safety law. These include, for example, AI systems that are parts of medical devices or toys. In these cases, the requirements set out in the AIA will be 'integrated into existing sectoral safety legislation' (AIA's Explanatory Memorandum: p. 4). The reason for this is to avoid duplicating administrative burdens and to maintain clear roles and responsibilities while ensuring a strong consistency among the different strands of EU legislation. However, it also implies that no 'AI specific' conformity assessments will take place. Instead, compliance with the AIA will be assessed through the third-party conformity assessment procedures already established in each sector.

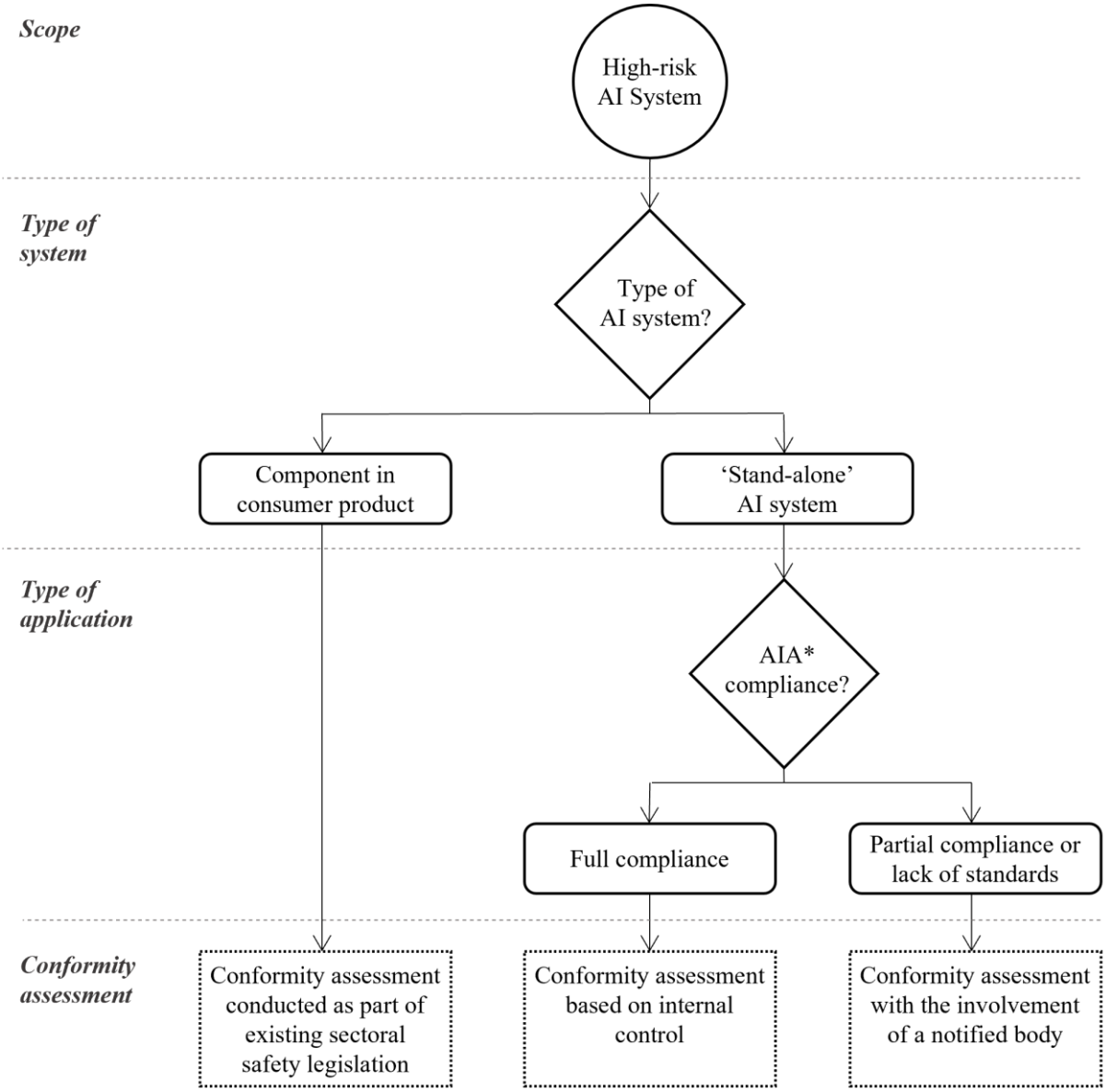
High-risk AI systems that do not fall into the first category are referred to as 'stand-alone' systems. The complete list of stand-alone, high-risk AI systems subject to conformity assessments is found in ANNEX III to the AIA. These include AI systems used in recruitment, determining access to educational institutions, and profiling persons for law enforcement, to mention a few notable examples. All stand-alone high-risk AI systems must comply with the requirements set out in the AIA. However, providers of stand-alone, high-risk AI systems have two options for how to conduct ex-ante conformity assessments. They can either (a) conduct ex-ante conformity assessments based on internal control, or (b) involve a third-party auditor

[92] Ex-ante or 'before the event' conformity assessments take place before a system is placed on the market. In contrast, post-market monitoring is a type of ex-post compliance check.

(i.e., a notified body, more on this in Section 5.5) to assess their QMS and technical documentation (AIA: Article 43).

Procedure (a) is only an option where the stand-alone, high-risk AI system is fully compliant with the requirements set out in Chapter 2 of Title III of the AIA. When, in contrast, the compliance is only partial (or harmonised standards do not yet exist) providers are obliged to follow procedure (b). This may seem opaque. However, Figure 10 illustrates through a simple flow-chart when different ways for conducting conformity assessments apply.

Figure 10. Ways to conduct conformity assessments for high-risk AI systems.



Ultimately, the legal requirements are the same for all high-risk AI systems. According to ANNEX IV in the AIA, these include, amongst others, obligations on the provider to:

(i) *Document* the intended purpose of the AI system in question,

(ii) *Provide* detailed user instructions,

(iii) *Disclose* the methods used to develop the system, and

(iv) *Justify* the critical design choices made by the provider.

However, in practice, not all high-risk AI systems will be subjected to third-party (i.e., external) ex-ante conformity assessments. The conformity assessments based on internal control that some providers of stand-alone, high-risk AI systems will have to conduct are more akin to what in the AI auditing literature is referred to as *internal auditing*. These internal checks would include properly documented ex-ante compliance with all requirements of the proposed EU legislation and establishing robust quality and risk management systems per Article 17 in the AIA. In addition, the internal conformity assessment should be accompanied by technical documentation concerning internal governance processes (AIA: Article 18).

Both external and internal audits come with their own sets of strengths and weaknesses. Because external audits help address concerns about the incentives for accuracy in self-reporting, they are typically required for formal verification and certification procedures (Brundage et al., 2020). However, external audits are fundamentally limited by a lack of access to internal processes at the audited organisation (Raji et al., 2020). Hence, they have a limited impact on how AI systems are designed. At the same time, organisations often employ internal audits to check the process in which AI systems are developed (Raji et al., 2020). While they run an increased risk of collusion between auditors and auditee, internal audits can thus constitute a first step towards making informed model design decisions (Saleiro et al., 2018).

Of course, the Commission is aware of the risks associated with internal audits. However, the AI sector is very innovative, and expertise for AI auditing is only now being developed. Hence, the choice of mechanism design is justified in the AIA by the fact that the providers of stand-alone, high-risk AI systems are best placed to intervene in the early stages of the system development process. Further, while internal conformity assessments rely on the active collaboration of providers of high-risk AI systems, the AIA includes several safeguards against negligent behaviour on their parts. After performing the conformity assessment, providers of high-risk AI systems must draw up an EU declaration of conformity[93] (AIA: Article 48). This declaration then becomes part of the required documentation accompanying

[93] A separate declaration of conformity shall be drawn up for each AI system and kept for 10 years after the AI system has been placed on the market or put into service.

the high-risk AI system, which, in turn, serves as a basis for the CE marking. Here, it should be noted that not only non-compliance but also the failure to communicate proactively and transparently can subject providers of high-risk AI systems to penalties. Specifically, Article 71 in the AIA stipulates that the supply of incorrect, incomplete, or misleading information in response to a request from relevant authorities shall be subject to administrative fines.[94]

The outline above provides only a brief sketch of the three different paths through which conformity assessments can be conducted. However, my aim here is only to extract and make visible the information available in the proposed AIA as currently drafted, not to 'fill in the gaps.' In Section 5.7, I will turn to discuss how the AIA could be amended. However, two areas where further clarification is needed should be highlighted already at this stage. First, the AIA only provides limited guidance on how sector-specific conformity assessments will be conducted in practice. Further, while stressing that that the types of risks posed by an AI system should be evaluated on a sector-by-sector basis, the AIA does not provide any sector specific guidance on what type of documentation is needed. Nevertheless, the European Commission stresses that the AIA will be complemented by other, ongoing, or planned, initiatives. This includes, for example, revisions of sectoral product legislation such as the *Machinery Directive* and the *General Product Safety Directive* (AIA's Explanatory Memorandum: p. 5).

Second, there is a lack of clarity about which AI systems, precisely, require conformity assessments conducted with the involvement of a third-party (procedure (b) in the typology above). The AIA, as currently drafted, displays a somewhat circular reasoning: procedure (a), i.e., conformity assessments based on internal control is sufficient for stand-alone AI systems that are in compliance with the AIA, but how can providers know if a specific AI system is compliant before the assessment is performed? In the *Commission Staff Working Document* accompanying the AIA,[95] it is suggested that whether the conformity assessment needs to follow procedure (b) hinges on the intended use of the AI system in question. For example, the AIA explicitly states that conformity assessments of AI systems intended for remote biometric identification in public spaces will require the involvement of a third-party (Haataja & Bryson, 2021). However, the list of high-risk areas is likely to change over time, and borderline cases are bound to emerge. Further clarification will thus be needed – both with regards to the criteria

[94] 10,000,000 EUR or, if the provider is a company, 2 % of its total worldwide annual turnover for the preceding financial year, whichever is higher (AIA: Article 71).

[95] See the European Commission (2021c) *Commission Staff Working Document: Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council*

used to determine the appropriate conformity assessment procedure for different types of stand-alone AI systems and the safeguards needed to ensure that technology providers don't opt for internal control in cases where the AIA mandates the involvement of third-party auditors.

5.4.2 Post-market monitoring

In addition to the ex-ante conformity assessments described above, providers of high-risk AI systems are also expected to establish and document post-market monitoring systems. The task of post-market monitoring is to document and analyse the performance of high-risk AI systems throughout their lifetime (AIA: Article 61). These ex-post assessments are complementary to ex-ante certifications since providers of high-risk AI systems are expected to report any serious incident or any malfunctioning that constitute a breach of Union law (AIA: Article 62). They are also obliged to take immediately any corrective actions needed to bring the AI system under conformity or withdraw it from the market (AIA: Article 21).

To detect, report on, and address system failures in effective and systematic ways, providers must first draft post-market monitoring plans that account for, and are proportionate to, the nature of their respective AI systems. The post-market monitoring plan is, in turn, part of the required documentation that constitutes the basis for the conformity declaration (AIA: ANNEX IV). It is important to note that such ongoing, post-market monitoring is intrinsically linked to quality management. According to the AIA (Article 17), the main objective of the QMS is to establish procedures for how high-risk AI systems are designed, tested, and verified. However, it should also include procedures for data management, record keeping, and procedures for how to conduct post-market monitoring of the high-risk AI system in question.

Legally mandated post-market monitoring adds a new element and new complexities to corporate QMS. Providers of high-risk AI systems are not necessarily the ones operating them. Hence, providers must give users instructions on how to use high-risk AI systems and cooperate with them to enable post-market monitoring. Consider the requirement that high-risk AI systems shall be designed with capabilities to automatically record (or 'log') their decisions (AIA: Article 12). These logs can either be controlled by the user, the provider, or a third party, as per contractual agreements. However, it remains the provider's responsibility to ensure *that*, and plan for *how*, high-risk AI systems automatically generate logs.

The post-market monitoring plan is complementary to the conformity assessment because it is based on a different logic. A distinction is often made between three complementary yet distinct approaches to AI auditing: *functionality audits* focus on the rationale behind using an AI system; *code audits* entail reviewing the source code of an AI

system; and *impact audits* investigate the types, severity, and prevalence of effects of an AI system's outputs. Whereas the conformity assessments mandated by the AIA entail elements of both functionality audits and code audits, the post-market monitoring plan adds the element of impact auditing. This element is specifically important for AI systems that continue to learn, i.e., update their internal decision-making logic, after being deployed at the market.

Combined, the ex-ante conformity assessments and the post-market monitoring mandated by the AIA constitute a coordinated and robust basis for enforcing the proposed EU regulation. However, an enforcement mechanism will only be as good as the institution backing it. Thus, in the next section, I examine the institutional structure proposed in the AIA, i.e., the roles and responsibilities of different stakeholders in ensuring that high-risk AI systems comply with the proposed EU regulations throughout their lifecycles.

5.5 The emergence of an EU AI auditing ecosystem

Ensuring that high-risk AI systems satisfy the various requirements set out in the AIA requires a well-develop auditing ecosystem that consists of two components. First, an institutional structure is needed that clarifies the roles and responsibilities of private companies, national and supranational authorities. This would also include ensuring accountability for different types of system failures. Second, the actors in the ecosystem need access to well-calibrated auditing tools and the necessary expertise to carry out the different steps in demonstrating that high-risk AI systems comply with the AIA. Unfortunately, as noted by the CDEI (2021c), such an ecosystem does not yet exist. Nevertheless, as well shall see in this section, the proposed EU legislation already sketches the contours of an emerging European AI auditing ecosystem.

According to the AIA, the ultimate responsibility to ensure compliance and identify and mitigate compliance breaches rests with the providers and users of high-risk AI systems. However, to ensure regulatory oversight, the Commission proposes to set up a governance structure that spans both Union and national levels (AIA 's Explanatory Memorandum: p. 15).[96] At a Union level, a 'European Artificial Intelligence Board' will be established to share best

[96] The auditing ecosystem described in the text is subject to some specific adjustments where the AIA interacts with other pieces of EU legislation. This is the case, for instance, of the Union legislation on financial services. According to the AIA, the authorities responsible for the supervision and enforcement of the financial services legislation, including the European Central Bank, should be designated as competent authorities (AIA: Article 63.4). Moreover, where Union institutions, agencies and bodies fall within the scope of the AIA, the European Data Protections Supervisor shall act as market surveillance authority (AIA: Article 63.6).

practices among member states and to issue recommendations on uniform administrative practices (AIA: Article 56). Quite significantly, the AIA does not adopt the 'agencification'[97] approach, which is inherent in the enforcement machinery established in other strands of EU legislation (including the GDPR). In fact, the European Artificial Intelligence Board is not conceived as an independent body having a legal personality. Rather, it is understood as a coordinating structure, chaired by the Commission, where Member States' and Commission's representatives are gathered to facilitate the effective implementation of the AIA.

In addition, the Commission will set up and manage a centralised database for registering stand-alone, high-risk AI systems (AIA: Article 60). The purpose of the database is to increase public transparency and enable ex-post supervision by competent authorities.

At a national level, member states will have to designate a competent national authority to supervise the application and implementation of the AIA. This national supervisory authority is not supposed to conduct any conformity assessments itself. Instead, it will act as a notifying authority (AIA: Article 59) that assesses, designates, and notifies third-party organisations that, in turn, conduct conformity assessments of providers of high-risk AI systems. In the proposed EU legislation, these third-party organisations are sometimes referred to as 'conformity assessment bodies', but, more often, they are simply called 'notified bodies' (AIA: Article 3.22). To become a notified body, an organisation must apply for notification to the notifying authority of the member state in which they are established.[98]

The main task of a notified body is to assess and approve the QMS that providers of high-risk AI systems use for the design, development, and testing (AIA: ANNEX VII). Further, the notified body shall examine the technical documentation for each high-risk AI system produced under the same QMS. Based on these assessments, the notified body shall then determine whether both the QMS and the technical documentation satisfy the requirements set out in the AIA. Where conformity has been established, the notified body shall issue an EU technical documentation assessment certificate.[99] Figure 11 below provides an overview of the relationship between different private organisations and institutional bodies in the process of assessing and certifying stand-alone, high-risk AI systems.

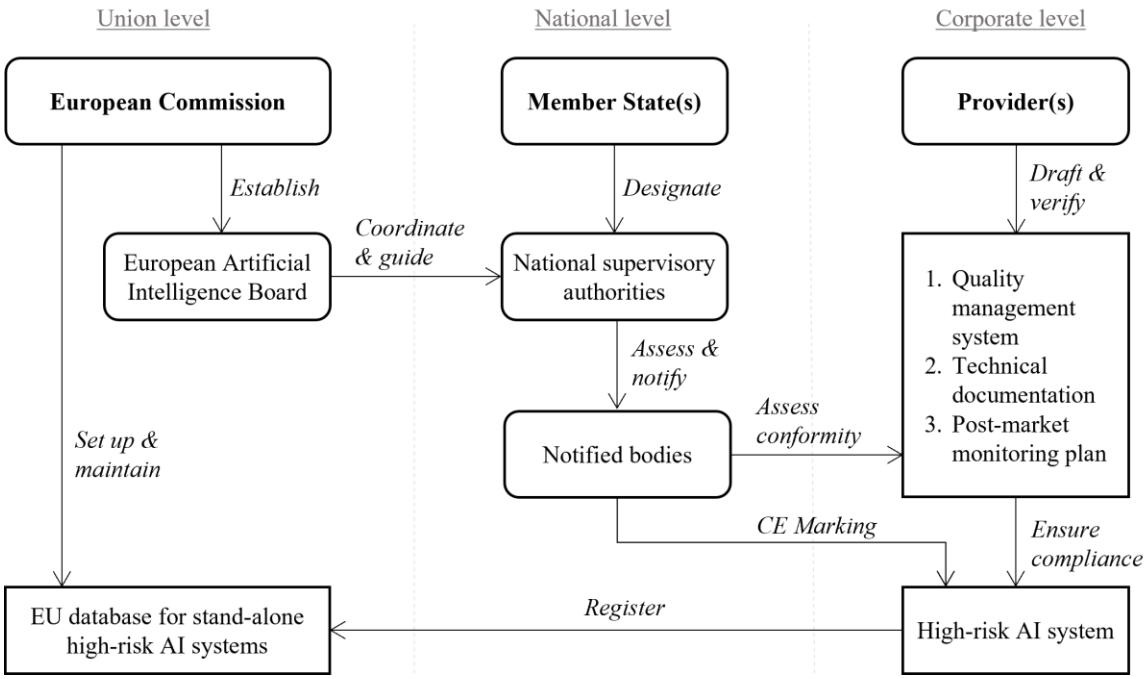
---

[97] The expression 'agencification' is normally used to refer to the proliferation of EU agencies within the EU legal order which has gained a terrific momentum from the 1990s (Chamon, 2016).

[98] Conformity assessment bodies established in third countries with which the Union has an agreement may be authorised to carry out the activities of notified bodies under this Regulation (AIA: Article 39).

[99] Note that this procedure is only applicable for the conformity assessments of stand-alone high-risk AI systems that require the involvement of third-party auditors (see Figure 10 on page 125 for guidance).

Figure 11. Roles and responsibilities during conformity assessments with third-party auditors.



It is important to note that Figure 11 gives a somewhat idealised picture of the roles and responsibilities outlined in the AIA. The relationships – here indicated by directional arrows – are in reality bidirectional. For example, although the national supervisory authority is responsible for assessing conformity assessment bodies, it does so based on the application and material submitted by organisations that wish to be notified.[100] After notification, each notified body is then assigned a unique identification number by the Commission (AIA: Article 35). Similarly, while the notified body is responsible for carrying out conformity assessments, providers of high-risk AI systems have an obligation to make the relationship work. That includes collaborating with the notified body, providing it with timely access to all resources and documents that are necessary for a comprehensive assessment to take place,[101] and reporting any severe incidents or malfunctioning of their high-risk AI systems directly to the national surveillance authority.[102] To deliver on these expectations, providers and users of AI systems may need to appoint new roles within their organisations.

[100] To become notified, conformity assessment bodies must demonstrate that they have the structure, competence, and resources required to fulfil their tasks.

[101] When public authorities and notified bodies need to be given access to confidential information or source code to examine compliance, they are placed under binding confidentiality obligations.

[102] Such notification shall be made as soon as possible and, in any event, no later than 15 days after the provider becomes aware of the serious incident or the malfunctioning (AIA: Article 62).

There is also a second sense in which Figure 11 is a simplification. It makes the process looks clear and solidified. In reality, the proposed EU legislation is quite vague and leaves significant room for interpretation: the language used in the AIA is highly technical, and, in several instances, multiple terms are used to refer to the same concept. For example, in the AIA, the terms 'notified body' and 'conformity assessment body' seems to be used interchangeably (AIA: Article 3.21 and 3.22). However, based on the tasks ascribed to the notified bodies, they could also have been called 'auditing bodies.' Similarly, what the AIA calls' notifying body' is equivalent to what is commonly known as 'accreditation body.' Most EU member states already have national accreditation bodies, and the AIA (Article 30) even highlights that these can be designated as notifying authorities. In 5.7, I shall discuss different points that demand clarification in greater detail. Before doing so, however, the next section will explore the potential scope for soft governance within the AIA.

5.6 The scope for soft governance within the AIA

In this section, I argue that the AIA should be seen as a complement to – and reinforcement of – the wide range of initiatives launched by both regulators and technology providers in recent years to ensure that AI systems are legal, ethical, and technically robust. Specifically, I stress that there will remain a demand for voluntary 'ethics-based' audits that allow organisations to validate claims about their AI systems and demonstrate adherence to ethics principles that go over and above compliance with the AIA.

To do so, it is useful to first take a step back and consider the distinction between *hard* and *soft* governance mechanisms. Hard governance refers to systems of rules elaborated and enforced through institutions to govern agents' behaviour (Floridi, 2018). Examples of hard governance mechanisms range from legal restrictions on system outputs to the prohibition of AI systems for specific applications (Koene et al., 2019). Both the conformity assessments and the mandatory post-market monitoring procedures discussed in the previous section fall into the category of hard governance mechanisms. In contrast, soft governance embodies mechanisms that exhibit some degree of contextual flexibility, like subsidies. Put differently, while hard governance refers to legally binding obligations, soft governance includes non-binding guidelines, incentives, or support infrastructure (Erdelyi & Goldsmith, 2018).

While analytically useful, the distinction between soft and hard governance is a simplification of what in reality is a more nuanced spectrum. Within the field of AI governance, there is a vast literature on both *self-governance*, which refers to collective, voluntary actions

of industry members (Rolski et al, 2021) and *co-governance*, which relies on cooperation between state and non-state actors to address the social, ethical, and environmental challenges AI systems pose (Corrigan, 2022). For my purposes, this literature holds two key takeaways. The first is that the boundary between soft and hard governance mechanisms is seldom hard and fast. As Corrigan (2022) notes, co-governance refers to a wide spectrum of mechanism that combine various elements of organisational policies, sector-wide standards, and state-led regulation. To capture this spectrum, Cave et al. (2008) proposed a 'Beaufort Scale' of self-regulation, that distinguishes between different degrees of government involvement, ranging from 0 (pure, unforced, self-regulation) to 11 (government-imposed mandates). With 'soft governance', I thus refer not only to self-regulation but also to the many intermediate degrees of co-governance in Cave et al.'s taxonomy that allow private actors some degree of flexibility.

The second takeaway is that hard and soft governance mechanisms often complement and reinforce each other (Hodges, 2015). As Cave et al. (2008) concludes, one of the main advantages of soft governance is that delegating responsibility to those 'closer to the action' enhances the effectiveness of regulatory activities. This is especially important in the case of AI governance, since laws may not always be up to speed in sectors that experience fast-paced innovation. Further, decisions made by AI systems may deserve scrutiny even when they are not illegal. Hence, there will remain room for EBA procedures, whereby organisations can demonstrate adherence to voluntary standards that go over and above existing regulations.

In and of itself, the AIA constitutes a proposal for hard governance. However, the AIA also leaves room for soft governance in general and EBA procedures in particular. Most notably, providers of non-high-risk AI systems are encouraged to draw up and apply voluntary codes of conduct (AIA: Article 69) related to their internal procedures and the technical characteristics of their AI systems. The critical difference between these voluntary codes of conduct and the other requirements in the AIA is that they focus on *process management* rather than *goal management*. This leaves individual organisations free to either draw up ethics principles of their own, adopt principles recommended by the European Artificial Intelligence Board, or declare adherence to any other set of standards relevant for their specific industry.

In the context of the AIA, the European Commission has at least two reasons for encouraging the voluntary use of codes of conduct. The first is to foster the voluntary application of the requirements set out in the AIA, even to use cases not subjected to mandatory conformity assessments. It should be noted that – as of now – the proposed EU legislation imposes no restrictions or obligations on AI systems that are deemed to constitute little or no risk, such as AI-enabled video games and spam filters (O'Donoghue et al., 2021). However,

depending on their technical specifications and intended purpose, such systems may also benefit from compliance with the requirements set out in Chapter 2 of the AIA concerning data quality, traceability, technical robustness, and accuracy.

The second objective is to promote post-compliance ethical behaviour. Even providers of high-risk AI systems may benefit from adopting voluntary codes of conduct that go over and above the requirements set out in the AIA.[103] Providers have good reasons to subject themselves to EBA: just as organisations seek to certify that their operations are sustainable from an environmental point of view (IEEE, 2019), or demonstrate to consumers that products are healthy through detailed nutritional labels (Holland et al., 2018), the documentation and communication of the steps taken to ensure that AI systems are ethically-sound can play a positive role in both marketing and public relations. By contributing to procedural regularity in, and transparent communication about, how AI systems are designed and deployed, EBA can help organisations manage financial and legal risks (Koene et al., 2019), improve public relations (EIU, 2020), and gain competitive advantages (European Commission, 2019).

As the analysis in this section has demonstrated, the European Commission encourages the voluntary adoption of codes of conduct and supports the emergence of complementary, soft governance mechanisms that sit on top of the AIA. This is promising. However, the AIA does not provide guidance on *whether* and *how* adherence to voluntary codes of conduct will be assessed. This is a missed opportunity, since there is in fact much that regulators can do to support the feasibility and effectiveness of EBA procedures that emerge bottom up from the assurance needs of different stakeholders in the AI ecosystem. In the next section, I discuss this omission alongside other areas where further guidance may be required.

5.7 The need for further guidance

The overall strategy to implement the AIA is clear. Nevertheless, further guidance is needed in several areas. This section highlights seven such areas with implications for the effectiveness and feasibility of AI auditing. Before doing so, however, it is worth reiterating that the recommendations provided in this section feed into an ongoing policy process. Hence, some of the gaps identified below may yet be ironed out as the AIA is still being revised and amended.

[103] Such codes of conduct may, for example, concern commitments to environmental sustainability, stakeholders' participation in the design of AI systems, or the diversity of development teams.

5.7.1 Level of abstraction

As many commentators have already noticed, some expectations in the AIA seem 'too idealistic' and will thus require 'a lot more guidance' (Gallo et al., 2021). Consider the data quality requirement that 'training, validation, and testing data sets shall be relevant, representative, free of errors, and complete' (AIA: Article 10.3). While this is a laudable vision statement at a high level of abstraction (LoA),[104] it may not be feasible to expect data sets to be completely 'free of errors' in practice. Setting the bar too high, or articulating requirements in too abstract terms, can backfire since rules that cannot be translated into operational terms are likely to be regarded only mechanically as a box-ticking exercise. Moreover, unrealistic expectations may undermine the legitimacy of the framework as a whole (Power, 1997).[105] Again, setting high-level expectations is useful, since expectations help shape the behaviour of different actors in multi-agent ecosystems (Minkkinen et al., 2021). Nevertheless, the AIA needs to provide further guidance on lower, and more detailed, LoAs. That is, high-level visions for data management and software development need to be broken down into applicable industry standards and evaluation metrics.[106] This is particularly important from an auditing perspective, since audits presuppose a realistic benchmark against which to audit.

5.7.2 Material scope

The material scope of the original draft of the AIA is opaque. In some cases, it looks very broad. For instance, the definition provided in ANNEX I to the AIA encapsulates several software-developing techniques, including machine learning approaches like deep neural networks, logic- and knowledge-based approaches like expert systems, and statistical approaches like Bayesian search and optimisation methods. The idea that a single regulatory approach could be designed in such a way as to tackle the issues associated with each one of these technologies is problematic. Ultimately, all that these technologies have in common is

[104] A level of abstraction (LoA) is a finite but non-empty set of observables, which are expected to be the building blocks in a theory characterised by their very choice (Floridi, 2008). Different LoAs can be nested, disjoined, or overlapping and need not be hierarchically related (Floridi, 2017a). However, this is not a relativist approach: a question is always asked for a purpose, and different LoAs can 'fit' the purpose more or less successfully.

[105] Audits can be viewed as rituals of verification that build trust through procedural regularity (Power, 1997). Hence, it is essential for the legitimacy of the process that the standard outcome is positive.

[106] While standards play an important role in any coordinated response to the risks posed by AI systems (Cihon, 2019), they are of particular importance for audits, since these presuppose a sound baseline to audit against.

that they process data.[107] A more narrowly defined scope may help providers of AI systems, third-party auditors, and national authorities direct their resources more effectively. The most recent compromise text approved by the Council of the EU (2022) addresses this point by proposing a slightly narrower definition of AI systems.[108] However, which definition will be included in the final version agreed to by the Council, the Parliament, as well as the member states remains to be seen.

Also problematic is the decision to include an exhaustive list of high-risk AI systems in the proposed legislation. As stressed by the European Data Protection Board (EDPB) and the European Data Protection Supervisor (EDPS) in a Joint Opinion on the AIA (EDPB/EDPS, 2021) adopted on 18 June 2021, this technique might create a 'black-and-white effect', undermining the risk-based approach of the Proposal. Further, the list misses some types of uses which are likely to involve significant risks, e.g., the use of AI for military applications, for determining insurance premiums, or for health research purposes.[109] Problems related to the material scope of the AIA may also come from the decision to exclude explicitly the international law enforcement cooperation (AIA: Article 2.4). Taken together, both the broad definition of 'AI and the attempt to exhaustively list applications that fall within the material scope of the AIA risk undermining the purpose of the regulation. Given that the AIA is justified with reference to the challenges associated with the complexity, unpredictability, and autonomy of specific AI systems (AIA's Explanatory Memorandum: p. 2), further clarification is required as to how the material scope (as currently defined), is linked to that regulatory goal.

5.7.3 Conceptual precision

At times, the language used in the AIA is vague and imprecise. For example, the vague terminology used in AIA Article 5 to identify the prohibited uses of AI runs the risk of making such limitations meaningless in practice. In fact, pursuant to that article, AI systems that 'deploy subliminal techniques beyond a person's consciousness to distort a person's behaviour in ways that may cause harm' are prohibited. However, digital mediation inevitably influences

[107] But so do human decision-makers. And, since they also make mistakes and produce discriminatory or inconsistent outcomes (Kahneman, 2011), the use of AI systems can sometimes lead to more objective and potentially fairer decisions (Lepri et al., 2018).

[108] Definition of AI systems proposed in the latest compromise text still includes both systems developed through ML approaches and systems developed through logic- or knowledge-based approaches.

[109] In fact, since AI technologies are quickly evolving, there is a real risk that any detailed list of so-called high-risk use-cases will be obsolete by the time the AIA comes into effect.

human users, e.g., by nudging an individual's preferences through positive reinforcement or indirect suggestion (Thaler & Sunstein, 2008b; Yeung, 2017). Hence, further guidance is needed regarding which kinds of distortions the AIA refers to as prohibited.[110]

A further point that should be better clarified is related to the control of conformity of AI systems already in use. According to Article 83.2 of the Proposal, those systems should be excluded from the scope of the Regulation, unless they are subject to 'significant changes in their design or intended purposes.' The provision does not offer further details thereon and the related threshold remains unclear. Some additional elements contributing to clarify the wording of Article 83.2 may be inferred from Recital 66 of the Proposal, which specifies that conformity re-assessment shall take place 'whenever a change occurs which may affect the compliance.' Even though that threshold is related to AI systems which were already subject to a conformity assessment, it could be also applied to pre-existing AI systems. A more accurate definition of the situations covered by Article 83 would be in any case necessary. Most importantly, the AIA would benefit from further guidance on how vague concepts like 'subliminal distortion techniques' or 'causal links' should be interpreted in practice.

5.7.4 Procedural guidance

While the logic behind the conformity assessments and the post-market monitoring activities mandated in the AIA is clear, many details concerning how these should be conducted have yet to be spelt out. For example, Article 20 in the AIA stipulates that the logs shall be kept for a period that is 'appropriate in the light of the intended purpose of the high-risk AI system.' However, the AIA does neither say how long is appropriate nor suggest who is responsible for determining this (e.g., the provider, a notified body, national authorities, or the Commission).

Further, in ANNEX VII to the AIA, it is stated that notified bodies shall carry out periodic audits to make sure that the provider maintains and applies the QMS following the technical documentation provided during the conformity assessment. However, the AIA does not specify how often periodic audits should be conducted or how such audits are triggered.[111]

---

[110] In a recent article in Nature, Köbis et al. (2021) distinguished between four main roles through which both humans and machines can influence ethical behaviour. These are role model, advisor, partner, and delegate. It is, in particular, AI agents acting as enablers of unethical behaviour (partners or delegates) that may let people reap unethical benefits while feeling good about themselves, a potentially perilous interaction.

[111] Article 44 in the AIA states that certificates shall only be valid for the period they indicate, which shall not exceed five years. However, it is unclear whether the periodic audits mentioned in ANNEX VII refer to the re-assessments that are required to extend the validity of a certificate for further periods, or to periodic audits during the continuous operation of an already certified AI system.

Finally, the AIA focuses exclusively on AI systems aimed for the market.[112] However, the distinction between basic and market-oriented research is not always clear – and even AI systems used for internal purposes may pose ethical risks. These examples highlight a need for further procedural guidance on how conformity assessments and post-market monitoring should be conducted in practice.

5.7.5 Institutional mandate

The European Commission and the national supervisory authorities, supported by The European Artificial Intelligence Board, have mandates to implement the hard governance mechanisms proposed in the AIA.[113] However, while the Commission's powers are clearly identified in the Proposal, the mandate of the European Artificial Intelligence Boards remains unclear.[114] The decision to exclude the independence of the Board, which is subject to a significant control by the Commission, is also debatable. Strong criticisms concerning that institutional solution may be found in the EDPB/EDPS Joint Opinion on the AIA, where the two entities stress the need to recognize more autonomy to the Board through a clearer identification of its nature and powers. Again, promisingly, the latest compromise text issued by the Council takes a step in this direction.

On a different note, the AIA does not prevent national authorities from keeping specific regulatory prerogatives in implementing the relevant obligations.[115] This risks reproducing the same fragmented approach emerging from the national implementation of GDPR (European Parliament, 2021). Moreover, the AIA does not include any institutional safeguards to maintain the integrity of the voluntary codes of conduct that it encourages providers of non-high-risk AI systems to adopt. This is problematic since the adoption of voluntary codes of conduct can be undermined by unethical behaviours like 'ethics bluewashing', i.e., an organisation making unsubstantiated claims about AI systems to appear more ethical than one is (Floridi, 2019b). Hence, any set of ethical principles will only be as good as the public institution backing it (Boddington, 2017). A potential solution would be to create or designate an independent entity that authorises organisations that conduct EBA to check whether providers of non-high-risk AI

[112] For example, while AI systems intended to distort human behaviour are prohibited, the European Commission explicitly states that research for legitimate purposes should not be stifled by the prohibition (AIA: p 18).
[113] Except in cases where the AIA interacts with specific sectoral policies of the Union.
[114] The European Artificial Intelligence Board should be 'responsible for a number of advisory tasks' (AIA: p. 35). However, the AIA does not specify *how*, and with which *mandate*, the Board will operate in practice.
[115] AIA: Recital 71, recognizing the Member States' right to elaborate artificial intelligence regulatory sandboxes.

systems adhere to their stated codes of conduct.[116] Given that the AIA already sketches the contours of a Europe-wide AI auditing ecosystem, one opportunity would be to leverage the same institutional structure to provide assurance also for post-compliance, EBA.

5.7.6 Resolving tensions

When designing and operating AI systems, tensions may arise between different ethical principles for which there are no fixed solutions (AI HLEG, 2019). For example, a particular ADMS may improve the overall accuracy of decisions but discriminate against specific subgroups in the population (Whittlestone et al., 2019a). Similarly, different definitions of fairness – like individual fairness and demographic parity – are mutually exclusive (Friedler et al., 2016; Kusner et al., 2017). Given these unresolved normative tensions, it is encouraging that the conformity assessments proposed in the AIA focus on making implicit design choices visible through the disclosure of technical documentation. Organisations are, and should be, free to strike justifiable ethical trade-offs within the limits of legal permissibility and operational viability. However, organisations that develop AI systems respond to various stakeholders who often have divergent interests. European regulators could help providers of AI systems understand and account for these diverse sets of interests, e.g., by complementing the requirements set out in the AIA with further guidance on how to resolve tensions between conflicting values, such as accuracy and privacy, as well as on how to prioritise between conflicting definitions of normative concepts, like fairness, in different situations.

5.7.7 Checks and balances

Although high-risk AI systems are subject to conformity assessments, the enforcement of the requirements set out in the AIA is less stringent than it appears (MacCarthy & Propp, 2021). This is because (for most high-risk AI systems) the conformity assessments will be based on internal checks conducted by the system provider itself. Further, while providers must draw up an EU declaration of conformity and give a copy of it to the relevant national authorities upon request (AIA: Article 48), *how* providers ensure compliance with the AIA is not disclosed to the public. This lack of checks and balances is problematic because pursuing rapid technological progress leaves little time to ensure that AI systems are robust and ethical

[116] Note that an obligation to demonstrate adherence to officially communicated codes of conduct is compatible with the voluntary nature of the code of conduct itself.

(Whittlestone et al., 2019b). Companies thus find themselves wedged between the benefits of innovation and social responsibility and may not act ethically in the absence of oversight.

Fortunately, there are several ways of strengthening the conformity assessment process outlined in the AIA. One way would be to impose even stricter transparency obligations so that the conformity assessment process – including the trade-offs made in designing a specific high-risk AI system – are disclosed to the wider public. Another option would be to subject the QMS put in place by individual providers of high-risk AI systems to ad-hoc audits by independent third parties. Some guidance could also be provided by the ECJ's case-law. However, if one considers the balancing exercise showed so far by the Luxembourg judges in the digital domain, it appears evident that their contribution will be far from decisive. Not only is the relevant case law fragmented, which prevents the emergence of a unitary approach to be replicated in the different strands of EU legislation (Fontanelli, 2016), but it also does not remove the need for private parties to engage in a delicate and unpredictable balancing act.[117]

5.8 Discussion

In this chapter I have argued that, on the whole, the proposed EU legislation is a good starting point for balancing the prospective benefits from promoting responsible innovation and providing proportionate safeguards against the risks posed by AI systems. In particular, the risk-based approach taken in the AIA is promising because it shifts the focus from technology to policy. This means that it will be less important to label a specific technical system 'AI' and more important to scrutinise the normative ends for which the system is employed.

Further, my analysis in this chapter suggest that the governance mechanisms proposed in the AIA (the conformity assessments and the post-market monitoring) bridge a critical gap. Hitherto, providers of AI systems have been encouraged to adopt and adhere to voluntary ethics principles (Hagendorff, 2020). However, central questions – like according to which metrics AI systems should be evaluated and who should be accountable for system failures – have remained unanswered (Floridi & Cowls, 2019). By proposing tangible governance mechanisms an institutional structure with the mandate to implement these, the AIA provides a framework for preventing, reporting on, and allocating accountability for different kinds of system failures.

Most importantly, my analysis of the role of auditing in the proposed AIA suggest that EBA procedures are not only compatible with but also complementary to regulatory

[117] See for instance Case C-507/17 *Google LLC* EU:C:2019:772 and (Susi, 2019).

approaches to managing the social and ethical risks posed by AI systems. This conclusion is supported by established theory as well as by formulations in the AIA. As noted by Cave et al. (2008) both pure industry self-regulation and exclusive reliance on government-imposed regulation are found infrequently in practice. More common are co-governance approaches, in which a plurality of soft and hard governance mechanisms complement and reinforce each other (Corrigan, 2022). This means that in moving from policy to implementation, no single governance mechanism will in isolation be able to address all the risks posed by AI systems. The AIA acknowledges as much. For example, while requiring providers of high-risk AI systems to undergo conformity assessments, the European Commission also encourage providers of non-high-risk AI systems to adopt and adhere to voluntary codes of conduct.

However, despite these merits, the proposed EU legislation still leaves some room for improvement. In this chapter, I have argued that the AIA de facto sketches an EU-wide ecosystem for auditing AI systems, albeit in other words. Conformity assessments based on internal checks, for example, are akin to what in the AI auditing literature is called *internal audits*; conformity assessments based on technical documentation with the involvement of a notified body resemble what is known as *external audits*; and the post-market monitoring that providers of high-risk AI systems will have to conduct follows the same methodological logic as *continuous auditing*. I believe the European Commission should make this explicit and plan it strategically. It may even be preferable to move further in the direction of 'conformity assessment', avoid any reluctance, and commit to fully supporting an EU-wide auditing ecosystem that is able to provide both compliance and post-compliance levels of assurance.

Of course, there may be good reasons for choosing different terminology and the language used in the AIA echoes the solutions adopted in other pieces of the EU legislation, starting from the legislation concerning the market surveillance and compliance of products.[118] However, by anchoring the AIA in the existing literature on AI auditing, valuable lessons can be learned from previous research. For example, auditing presupposes a predefined baseline to audit against. Hence, vague concepts like 'distorting behaviours' or 'causal links' must be translated into practically verifiable criteria for providers of AI systems to demonstrate adherence to the AIA. Similarly, the risks associated with internal audits are well known. Hence, the AIA would benefit from the inclusion of additional institutional safeguards concerning the enforcement of conformity assessments based on internal control.

[118] See Regulation (EU) 2019/1020 with which the AIA presents strong interactions.

By discussing the limitations and omissions of the original draft of the AIA, I do not seek to diminish its many merits. In contrast, I support the approach adopted. That is why I have highlighted areas where potential amendments to the AIA would help strengthen its overall effectiveness in contributing to good AI governance in the EU and beyond.

I want to end this discussion by reminding readers that the AIA analyzed in this chapter constitute a draft legislation that is still being discussed in the European parliament. As part of this process, several amendments to the AIA have been proposed since the article on which this chapter is based was first published. Based on the latest compromise text published by the Council of Europe (on 6th December 2022), three changes are worth highlighting. First, the Council is in favor of a more narrowly defined material scope. The direction of this change aligns with my recommendations in this chapter. Second, the Council proposes that the AI Board should be given greater autonomy than originally proposed. If provided with a strong enough mandate, the AI Board could help address some of the institutional shortcomings of the AIA that I discussed in Section 5.7. Finally, new provisions have been added to account for AI systems with highly general capabilities, i.e., systems that can be easily adapted to perform a wide range of different tasks. As of now, the two governance mechanisms – conformity assessments and post-market monitoring – only go so far in addressing the governance challenges posed by such systems. In Chapter 7 of this thesis, I will build on and expand the regulatory approach outlined in the AIA to propose a three-layered approach for how to audit AI systems with highly general capabilities.

5.9 Concluding remarks

In this chapter, I have shown that the AIA sketches the contours of an EU wide AI auditing ecosystem in all but name. The two governance mechanisms included in the AIA (conformity assessments and post-market monitoring plans) closely resemble ex-ante and ex-post audits as conceptualised within the AI auditing literature. Moreover, the institutional relationship between what the AIA refers to as ‘notified bodies’ and ‘notifying bodies’ mirrors that between auditors and national accreditation bodies. These observations are practically useful, since they help private sector actors interpret the AIA and understand how they can demonstrate compliance with it. However, the analysis provided in this chapter also has direct implications for the purpose of this thesis.

In the process of analysing the role of auditing in the AIA, I have also addressed:

SQ3 How can EBA complement legislative approaches to managing the risks posed by AI systems?

The AIA constitutes an appropriate backdrop against which to answer SQ3 for two reasons. First, it constitutes the most mature and ambitious AI regulation proposed by any major global economy to date. Second, given the 'Brussels effect' (Bradford, 2020), the AIA is likely to have implications for other jurisdictions as well. So, what have we learned with respect to SQ3?

EBA complement and reinforce legislative approaches in at least three ways. First, EBA provides assurance for AI systems that are not covered by the legislation. For example, the conformity assessments and post-market monitoring plans mandated in the AIA only apply to high-risk AI systems – which only constitutes a small subset of all AI systems. In the AIA, the European Commission explicitly stresses this point by encouraging providers of non-high-risk AI systems to adopt and adhere to voluntary codes of conduct.

Second, EBA allows providers to demonstrate adherence to ethics principles that go beyond legal compliance. This is an important function for both social and economic reasons. The use of AI systems may be problematic and deserving of scrutiny even when not illegal. Further, as my case study of AstraZeneca in Chapter 4 demonstrated, private companies have an interest in ensuring good AI governance to manage financial and reputational risk.

Third, EBA procedures inform policymaking by allowing researchers to study the feasibility and effectiveness of different governance mechanisms. This point is illustrated by the fact that both the conformity assessments and the post-market monitoring plans proposed in the AIA are adaptations of well-established AI auditing tools and methods.

At the same time, policymakers can do much to foster good AI governance without relying solely on legislation. AI governance is most effective when technology providers and policymakers work together to address issues of mutual concern (Corrigan, 2022). But as Cave et al. (2008) stress, the possibility of soft governance mechanisms being effective depends in part on the alignment of interest. It is, in other words, not independent of the guidance, support, and incentives policymakers provide. For instance, policymakers can support the emergence of feasible and effective EBA procedures by (i) creating standardised evaluation metrics and reporting formats, (ii) investing in infrastructure that auditors can use to share best practices, and (iii) establishing an institutional ecosystem to authorise auditors.[119]

[119] These three examples correspond to (i) standardised, (ii) co-funded, and (iii) recognized self-governance in Cave et al.'s (2008) taxonomy of degrees of government involvement in different co-governance mechanisms.

In the AIA, the European Commission goes some way to achieving that. For example, it envisages setting up coordinated 'regulatory sandboxes' which would allow providers to experiment with new AI systems in a safe setting. It also proposes to establish a database where providers can share information on serious incidents. Most importantly, it requires EU member states to designate competent authorities to accredit AI auditors. As of now, all these proposals are primarily intended to support conformity assessments and post-market monitoring of high-risk AI systems. However, as I will expand on in Chapter 8, the AI auditing ecosystem sketched by the European Commission can also be used to support EBA.

Finally, as my analysis in this chapter has shown, how to define the material scope of the AIA has been one of the most contentious issues European policymakers have faced in the process. surrounding the AIA has been its material scope. A good overview of this discussion is provided in Joanna Bryson's (2022) article with the memorable title *Europe is in danger of using the wrong definition of AI*. But the question remains: how can the material scope of AI governance be demarcated? In Chapter 6, I will address the question of how that can be done.

CHAPTER 6

THE SWITCH, THE LADDER & THE MATRIX: MODELS FOR CLASSIFYING AUTOMATED DECISION-MAKING SYSTEMS

Abstract

Ethics-based auditing (EBA) is a governance mechanism that organisations designing or deploying *automated decision-making systems* (ADMS) can use to operationalise their ethical commitments. However, there still exists a gap between principles and practice. A major obstacle organisations face when attempting to implement EBA is the lack of a well-defined material scope. This difficulty is rooted in the more general problem of how to demarcate the material scope of ADMS governance. Of course, there exists no universally accepted definition of ADMS. Yet pragmatic problem-solving demands things to be sorted so that their grouping promotes successful actions for some specific end. In this chapter, I review previous attempts to classify ADMS for the purpose of implementing ADMS governance. I find that such attempts use one of three approaches: *the Switch*, i.e., a binary approach according to which systems either are or are not considered ADMS depending on their intrinsic characteristics; *the Ladder*, i.e., a risk-based approach that classifies systems according to the ethical risks they pose; and *the Matrix*, i.e., a multi-dimensional approach that accounts for various aspects, like input data, decision-model, and decision task, when classifying ADMS. Each of these approaches comes with its own set of strengths and weaknesses. By conceptualising different ways of classifying ADMS in terms of simple mental models, I hope to provide organisations that design, deploy, or regulate ADMS with the vocabulary needed to demarcate the material scope of not only EBA procedures but also ADMS governance more generally.

Note

This chapter is based on a journal article published in *Minds and Machines* (see Mökander et al., 2023a).[120] I have edited the original manuscript to ensure greater consistency with the other chapters. I have also harmonised the vocabulary to fit the overarching framing of this thesis.

[120] The article was co-authored with Margi Sheth, David Watson, and Luciano Floridi. Please see Appendix 3, 6, and 7 for authorship statements.

6.1 Introduction

6.1.1 Background

The use of automated decision-making systems (ADMS) is increasingly reshaping societies and transforming economies (AlgorithmWatch, 2019). The drivers behind this development are clear: the delegation of tasks to ADMS can improve efficiency, reduce costs, and enable new solutions to complex problems (Taddeo & Floridi, 2018). Already today, ADMS are employed to help improve health outcomes (Schneider, 2019) and mitigate environmental risks (Rolnick et al., 2019; Vinuesa et al., 2019). However, the use of ADMS is coupled with ethical challenges. An ADMS may be poorly designed, leaving individuals and groups vulnerable to poor quality outcomes, bias and discrimination, and invasion of privacy (Leslie, 2019). Further, ADMS can enable human wrongdoing, reduce human control, and erode human self-determination (Tsamados et al., 2020). At the same time, fear and misplaced concerns could hamper the adoption of well-designed ADMS, thereby leading to significant social opportunity costs (Cookson, 2018). These and other similar ethical challenges cannot be ignored if one wishes to reap the benefits brought by ADMS.

Many governments, research institutes, and NGOs have proposed ethical principles that provide normative guidance to organisations that design and deploy ADMS (Fjeld, 2020).[121] Although differing in terminology, these guidelines tend to converge on five principles: beneficence, non-maleficence, autonomy, justice, and explicability (Floridi & Cowls, 2019). In parallel, numerous organisations have adopted ethics principles of their own (de Laat, 2021). Notable examples include Google (2018), Microsoft (2019), and IBM (Cutler et al., 2018). Collectively, these efforts constitute a step in the right direction. However, the adoption of (and subsequent adherence to) ethics principles remains voluntary (Cath et al., 2018) and often unchecked. Moreover, the industry lacks both incentives and useful tools to translate abstract principles into verifiable criteria (Morley et al., 2020).

Legislation has only recently begun to change this picture. As I discussed at length in Chapter 5, the *Artificial Intelligence Act* (AIA), published by the European Commission in 2021, was the first comprehensive legislative framework for ADMS proposed by any major global economy. Since then, many other countries and regions have followed suit. For instance,

[121] Recent and influential contributions include the European Commission's *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019), the IEEE's principles for *Ethically Aligned Design* (IEEE, 2019), and the OECD's *Recommendation of the Council on Artificial Intelligence* (OECD, 2019).

the US Senate and House are currently considering the *Algorithmic Accountability Act of 2022* (AAA) (Office of US Senator Ron Wyden, 2022). Yet whether and when this bill will pass into law remains uncertain.

As I showed in Chapter 2 of this thesis, academic researchers and technology providers have already done much to bridge the gap between principles and practice in ADMS governance. Researchers have developed translational tools like model cards (Mitchell et al., 2019) and datasheets (Gebru et al., 2018); industry associations have drafted standardised protocols and reporting guidelines for the use of ADMS (Cruz Rivera et al., 2020; Liu et al., 2020); and private companies are increasingly subjecting themselves to *ethics-based auditing* (EBA) (Brown, 2021; Koshiyama et al., 2022). All these efforts serve the overarching purpose of providing organisations with the governance mechanisms needed to ensure that the ADMS they design or deploy are legal, ethical, and technically robust.

Still, many implementation challenges remain unsolved. In particular, the lack of a clear *material scope*[122] – that is, to which technological systems the specific ethical and legal considerations may or may not apply – continues to make it difficult to implement and enforce ADMS governance (Kritikos, 2019; Scherer, 2016). As noted by Aiken (2021), efforts to govern ADMS require standardized approaches to classifying the various types of ADMS in use. However, having studied both researchers' and policymakers' conceptions of ADMS, Krafft et al. (2020a) warn that the very possibility of having informed conversations about how to classify ADMS is hampered by conceptual ambiguity.

This observation is confirmed by my own empirical research. For example, my case study of ADMS governance within AstraZeneca (which I reported on in Chapter 4), indicated that one of the main challenges organisations face when attempting to implement EBA procedures is the difficulty to demarcate their material scope. In fact, organisations typically struggle even to produce an inventory of the ADMS they develop or use. Policymakers face similar difficulties. For example, as my analysis of the AIA in Chapter 5 highlighted, the question of how to define the proposed legislation's material scope has proven challenging for European policymakers and remains to be settled.[123] Despite such difficulties, however, both organisations that commit themselves to ethics principles and regulators that develop ADMS

[122] I use the term 'material scope' rather than just 'scope' because a given policy can also have other types of scope (such as 'territorial scope'). This use of the term 'material scope' is standard in the context of the GDPR.
[123] See e.g., Bryson (2022), CDEI (2021), and Bertuzzi (2023b).

governance frameworks inevitably face the question ‘To which systems and processes ought these additional layers of governance apply?.’

Of course, there is no one way to demarcate the material scope of ADMS governance. Different ADMS pose different ethical and legal challenges (Oxborough et al., 2018). Moreover, ADMS are often embedded in larger socio-technical systems (van de Poel, 2020) in which human- and machine-centric processes overlap and co-evolve (Tam et al., 2017). Admittedly, this ontological underdetermination is not unique to the problem of ADMS governance. Grouping things into neat categories seldom works, given the messy and continuous boundaries of the natural world (Cantwell Smith, 2019). But for the purpose of pragmatic inquiry and practical problem solving, things must be sorted so that their grouping can promote successful actions for some specific end (Dewey, 1920). In short, every policy needs to define its material scope.

6.1.2 Scope, methodology, and contributions

In this chapter, I address SQ4, i.e., how can the material scope for EBA be demarcated? However, as I have stressed in the introduction to this chapter, that question is a special case of the broader question of how the material scope of ADMS governance can be demarcated. If an answer to the latter question is found, the answer to SQ4 will follow. Moreover, the literature on ADMS governance is richer than the literature on EBA. For this reason, I will first focus on how the material scope of ADMS governance can be demarcated, before spelling out the implications of my findings for EBA specifically in the concluding remarks (Section 6.9).

To answer SQ4, I follow the methodology for applied research outlined in Section 1.7.2. To recap, I first conducted a *systematised literature review* (Grant & Booth, 2009) to I identify previous attempts to classify ADMS. I searched five databases (Google Scholar, Scopus, SSRN, Web of Science, and arXiv) for relevant articles. Keywords for the search included (‘artificial intelligence’, ‘automated decision-making systems’ OR ‘AI systems’) AND (‘governance’, ‘regulation’, ‘audit’ OR ‘principles’) AND (‘definition’ OR ‘material scope’).

However, not all ADMS governance frameworks are published in academic journals. In a second step, I thus used a snowballing technique (Wohlin, 2014) to track the citations of already included articles and identify relevant reports written by companies, policymakers, and industry associations. A total of 78 documents were included in this *document analysis* (Karppinen & Moe, 2012). In a third step, I analysed how each of these ADMS governance frameworks demarcated their material scope. The purpose of this *conceptual analysis* (Maggetti et al., 2015) was to identify the underlying logic behind different approaches.

To be clear, in this chapter, I do not propose any new model for how to classify ADMS. Instead, my focus is on identifying, describing, and evaluating different ways of classifying ADMS found in the recent literature. Applying this methodology, I find that previous attempts to classify ADMS follow one of three approaches. According to the *binary approach*, systems either are or are not considered ADMS, depending on their intrinsic characteristics. According to the *risk-based approach*, systems are classified into different categories depending on the types of ethical risks they pose. Finally, according to the *multi-dimensional approach*, various aspects – such as context, data input, and decision-model type – need to be considered when classifying systems. Using mental models (Johnson-Laird, 1983), I call these approaches *the Switch*, *the Ladder*, and *the Matrix*, respectively. In the following sections, I discuss each of these models in detail and provide concrete examples.

Before proceeding, two limitations help demarcate the scope of this chapter. First, my focus is intentionally limited to the identification and evaluation of approaches for how to demarcate the material scope of voluntary, ethics-based, ADMS governance frameworks. Doing so helps further my research objective to sharpen and extend the conceptual toolkit available to organisations who wish to audit the design and use of ADMS for alignment with specific ethics principles. My aim in this chapter is therefore to complement – not duplicate – previous work by legal scholars. Readers interested in different legal definitions of ADMS are referred to Schuett (2021) for a good overview and discussion. That said, legal considerations inevitably shape the design of corporate governance frameworks. Throughout this chapter, I therefore make frequent references to the material scope of regulatory proposals like the EU AIA and the US AAA for illustrative purposes.

Second, my review does not encompass abstract definitions of what an ADMS (much less 'artificial intelligence')[124] really *is*. As is well-known, there exists no universally accepted definition of ADMS (Wang, 2019). Discussions concerning the merits of different universal definitions of ADMS remain outside this chapter's scope. Instead, I focus on *classifications* that help organisations implement their ADMS governance frameworks. To quote John Dewey (1957), 'To have an aim is to limit, select, concentrate, and group.' And I have an aim in mind:

[124] Some researchers use 'AI' to refer to specific types of agents, i.e., those displaying some levels of autonomy, adaptability, and problem-solving capacity (Legg & Hutter, 2007). Others take 'AI' to demarcate computational techniques designed to approximate cognitive tasks (*US National Defence Authorization Act*, 2018). Yet others use 'AI' not to describe any technologies at all, but rather the science and engineering of making performant machines (McCarthy, 2007). As mentioned in Chapter 1, this is the main reason why I, in this thesis, prefer to use the term ADMS, which better describes the features of the systems in question.

to unlock the potential of autonomous and self-learning systems to serve as a force for good while managing the ethical challenges they pose.

A final methodological note. When describing and discussing different mental models for how to classify ADMS, I rely on the *method of levels of abstraction* (Floridi, 2008). Abstraction is a method for analysing and understanding complex phenomena that allows for the creation of concepts and objects at different levels of thinking and language (Van Leeuwen, 2014). Only within a level of abstraction (LoA) can comparisons between objects make sense. Note that this is not a relativist approach; a question is always asked for a purpose, and, for that specific purpose, there is an appropriate LoA that can be compared to others in terms of 'fitting' the purpose more or less successfully.

The remainder of this chapter is structured as follows. In Section 6.2, I build on previous work to showcase how ADMS can be classified in many ways, e.g., based on their technical features or the socio-technical contexts in which they are applied. In Section 6.3, I argue that, to establish the material scope of ADMS governance, good classifications of ADMS should be *fit for purpose*, *simple and clear*, and *stable over time*. I then introduce three models for how to classify ADMS. In Sections 6.4, 6.5, and 6.6, I describe and exemplify the Switch, the Ladder, and the Matrix, respectively. In Section 6.7, I evaluate these models according to the criteria set out in Section 6.3. Finally, in Section 6.8, I conclude by discussing how classifying ADMS is an LoA-dependent question. Hence, none of the models discussed in this chapter should be viewed as applicable absolutely, that is, independently of the choice of the LoA deemed to be most appropriate for the given purpose. Instead, I suggest that the models for classifying ADMS outlined in this chapter collectively constitute a useful set of conceptual tools for technology providers or regulators that wish to clarify the material scope of their ADMS governance frameworks.

6.2 Conceptualising automated decision-making systems

In Chapter 1, I introduced ADMS as autonomous and self-learning systems that gather and process data to make or inform decisions that impact individuals, groups, or the natural environment with little or no human intervention. That remains true at a high LoA. Still, to help practitioners understand whether an ADMS governance framework applies in a particular case, it must be complemented by classifications of ADMS at lower LoAs.

The word system in ADMS indicates that I am talking about a class of systems that differs from others, as opposed to some kind of intelligence that differs from human

intelligence (Kostopoulos, 2021). To capture this distinction, a wide range of terms like 'AI systems' (OECD, 2022; Leslie, 2019), 'AI-based systems' (Gasser & Almeida, 2017; Saleiro et al., 2018), 'autonomous systems' (Bryson & Winfield, 2017; IEEE SA, 2020), and 'algorithmic systems' (Ananny & Crawford, 2018; Rahwan, 2018) are often used interchangeably in the existing literature. For the sake of simplicity, I shall use the term ADMS consistently throughout this chapter. In doing so, I follow AlgorithmWatch (2019) and Whittaker et al. (2018), amongst others. But nothing hinges on this terminological choice.

Previous work has shown that ADMS can be classified according to several different dimensions. A distinction is often made between narrow and general ADMS (Russell et al., 2015). While *narrow* ADMS refers to systems that can perform specific tasks, *general* ADMS refers to systems that can perform a broad range of tasks (Goldstein, 2018). Most current ADMS are narrow, although there is a growing body of research on transfer learning (Weiss et al., 2016) and meta-learning (Vanschoren, 2018) explicitly devoted to building more general systems (more on this in Chapter 7).

Another distinction is often made between different computational techniques that underpin ADMS. While *symbolic* approaches are based on logic programming and symbol manipulation, *adaptive* methods rely on statistical techniques to solve specific problems without being explicitly programmed to do so (Russell & Norvig, 2015). This latter class includes machine learning (ML) algorithms, such as deep neural networks (Samoili et al., 2020). However, the two approaches are not necessarily mutually exclusive. So-called *hybrid architectures* combine the large-scale learning abilities of neural networks with symbolic knowledge representation (Marcus, 2020).

Within the realm of ML, researchers distinguish between supervised, unsupervised, and reinforcement learning. *Supervised learning* involves inferring a relationship from inputs to outputs, e.g., classifying image labels from pixels or predicting economic demand from time-series data (Hastie et al., 2009). In contrast, *unsupervised learning* is about finding patterns (e.g., clusters or latent variables) hidden in collections of unlabelled data without any predetermined target (Frankish & Ramsey, 2014). Finally, *reinforcement learning* occurs when an agent attempts to maximise rewards by interacting within some structured environment (Sutton & Barto, 2018). Policies are gradually improved through repeated trials, as when AlphaGo (Silver et al., 2016) became the world's greatest master of the ancient Chinese game 'Go' by playing against itself millions of times.

ADMS can also be classified with respect to the type of tasks they attempt to emulate (Feigenbaum & Feldman, 1963). Traditionally, ADMS research has focused on the following

problem domains: *perception*, i.e., the ability to transform sensory inputs into usable information; *reasoning*, i.e., the capability to solve problems; *knowledge*, i.e., the ability to represent and understand the world; *planning*, i.e., the capability of setting and achieving goals; and *communication*, i.e., the ability to understand and produce language (Corea, 2019).

Yet another (complementary) way of classifying ADMS is based on the type of analytics they perform. These include *descriptive analytics* (what happened?), *diagnostic analytics* (why did it happen?), *predictive analytics* (what is going to happen?), *prescriptive analytics* (what should happen?), and *automated analytics* (performing actions) (Corea, 2019).

Different ways of classifying ADMS can be combined. For example, in the original draft of the AIA, ADMS were defined through a combination of the computational techniques that underpin a system and the type of tasks it is designed to perform. More specifically, in Annex 1 to the original draft of the EU AIA, ADMS were defined as:

> *'software that [i] is developed with one or more of the [following] techniques and approaches: (a) Machine learning approaches, […]; (b) Logic- and knowledge-based approaches, […]; and (c) Statistical approaches, […], and [ii] can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.'* (European Commission, 2021b, p.1).

As I argued in Chapter 5, this definition is broad by any standard. Further, the diverse nature of the computational techniques it encapsulates – and the wide range of applications they enable – shows that it is often necessary to consider LoA-dependent factors when classifying ADMS. In practice, ADMS are not isolated technologies but integrated into larger socio-technical systems that encompass organisations, people, infrastructures, and processes (Chopra & Singh, 2018). Put differently, information processing – from the collection of input data to the final decision or classification – typically consists of several interconnected (and iterative) steps performed by human operators and computational systems (Chen & Golan, 2016). Hence, the decisions made by ADMS are never just a reflection of their technical properties but also of the socio-technical environment surrounding their use (Eubanks, 2019).

All technical artefacts (ADMS included) are value-laden insofar as they alter the cost-benefit ratio of the actions undertaken by humans and thus influence their decision-making (Danaher, 2012). Moreover, some ADMS can adapt their behaviour based on external inputs and evolve over time. This ability of ADMS to learn, i.e., to update continuously their internal decision-making logic, is one of the reasons why it is difficult to assign accountability when harm occurs (Burrell, 2016). It is this combination of relative autonomy and learning skills that

underpin both beneficial and problematic uses of ADMS. To determine the level of risk ADMS pose, it is therefore necessary to take several factors into account, including *data access*, i.e., the extent to which a specific system has complete and accurate knowledge about its environment; *model stability*, i.e., the extent to which it may alter its own control structure to perform its task; and *goal freedom*, i.e., the extent to which its goals are known and stable.

To summarise, previous work in the field suggests that separating ADMS from other systems is a LoA-dependent, multi-variable problem. As I will demonstrate in Sections 6.4 – 6.6, this problem can be approached in different ways. Before discussing these different approaches in greater detail, let us explore the needs of an effective classification system and establish criteria for the same.

6.3 Criteria for good classifications of ADMS

To implement ADMS governance, the concept 'ADMS' must be operationalised for three reasons. First, organisations are under constant pressure to innovate. By having a clearly defined material scope for ADMS governance, organisations can take care not to unduly burden systems or projects from which no ADMS-specific risks arise (AIEIG, 2020).

Second, governance is most effective when rules and norms are applied fairly, transparently, and consistently (Hodges, 2015). Without a shared understanding of what constitutes ADMS within a specific organisation, systems, and processes (henceforth *use cases*) are likely to be subject to additional scrutiny only on an *ad hoc* basis. Such a procedure undermines the legitimacy of the ADMS governance framework in question and hampers its ability to systematically identify and mitigate risks.

Third, not all ethical risks that organisations face stem from the use of ADMS. The inherent technical opacity of ADMS, for example, is often dwarfed by the opacity stemming from state secrecy or intellectual property rights (Burrell, 2016). Further, human decision-makers can also make mistakes and produce discriminatory outcomes (Kahneman, 2011). As a result, organisations already have processes in place to oversee human decision-making and enforce commitments related to CSR, and data management (e.g., in line with the *General Data Protection Regulation*). Therefore, a well-defined material scope for ADMS governance means that it should complement or refine existing governance structures, not duplicate, or generate inconsistencies within them (Mäntymäki et al., 2022).

What constitutes a good classification of ADMS systems? Drawing on best practices from previous attempts to create *working definitions* within the philosophy of science (Carnap, 1950), I argue that good classifications should be:

1) *Fit for purpose*. Good classifications should help organisations demarcate the material scope of ADMS governance in ways that are neither over- nor underinclusive. A classification is overinclusive when it includes systems that do not require additional oversight with respect to the normative goals of the ADMS governance framework. In contrast, a classification is underinclusive when systems that pose the specific ethical risks the ADMS governance framework seeks to address are not included.

2) *Simple and clear*. Good classifications should be easy to understand and apply in practice. This implies that practitioners should be able to determine, with little effort, how to classify a specific system. Ideally, the classification should be based on conditions that are discrete, i.e., which are either met or not. Finally, usefulness also implies that even non-experts should be able to apply the classification. This is because implementing ADMS governance requires the active participation of a wide range of staff across an organisation, including people who lack technical training and skills.

3) *Stable over time.* Good classifications should be resilient. Since computer science is a rapidly progressing field, ADMS' technical features and potential applications are subject to constant change. Hence, good classifications should not be based on elements that are likely to become obsolete too quickly.

Taken together, these criteria require that good models for classifying ADMS systems should help organisations specify to which use cases their ethics principles apply, enable practitioners determine to which class of ADMS a particular use case belongs, and provide a stable basis for ADMS governance over time. These criteria also constitute the LoA on which I will evaluate the strengths and weaknesses of different models for classifying ADMS in Section 6.7.

Before going any further, it is worth reiterating a point I made in the introduction of this thesis, namely, that *how ADMS are classified is an integral part of the design of ADMD governance frameworks*. For example, classifications that only establish minimum thresholds for what constitutes ADMS demand flexible governance frameworks that can handle a wide range of use cases in proportionate and effective ways. In contrast, fine-grained classifications afford layered ADMS governance frameworks that can specify both the risks and potential remedies associated with different use cases.

Building on this insight, it is possible to cluster pairs of *classifications of ADMS systems* and *ADMS governance frameworks* into three types. Using mental models, I have chosen to call these *the Switch*, *the Ladder*, and *the Matrix*, respectively. These are (of course) ideal types. In practice, many organisations use a combination of these approaches to demarcate the material scope of ADMS governance. Nevertheless, as I will show in the next sections, the Switch, the Ladder, and the Matrix are based on different logics. Hence, there is merit in describing, exemplifying, and evaluating them separately.

6.4 The Switch

Most ADMS governance frameworks include *working definitions* of ADMS. The aim thereby is to capture the most relevant features of the systems under investigation in a single sentence or paragraph, providing policymakers and practitioners with a simple rule of thumb for when to apply the framework. Some of these working definitions are too abstract to help establish a material scope. However, others are also useful for the practical purpose of implementing ADMS governance frameworks without complicating matters. Consider the approach taken by the IEEE. In 2020, the IEEE published its *Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being*.[125] In it, the following working definition is provided:

> *'[An ADMS is][126] a semi-autonomous or autonomous computer-controlled system programmed to carry out some tasks with or without limited human intervention capable of decision making by independent inference and successfully adapting to its context.'* (IEEE, 2020, p.18)

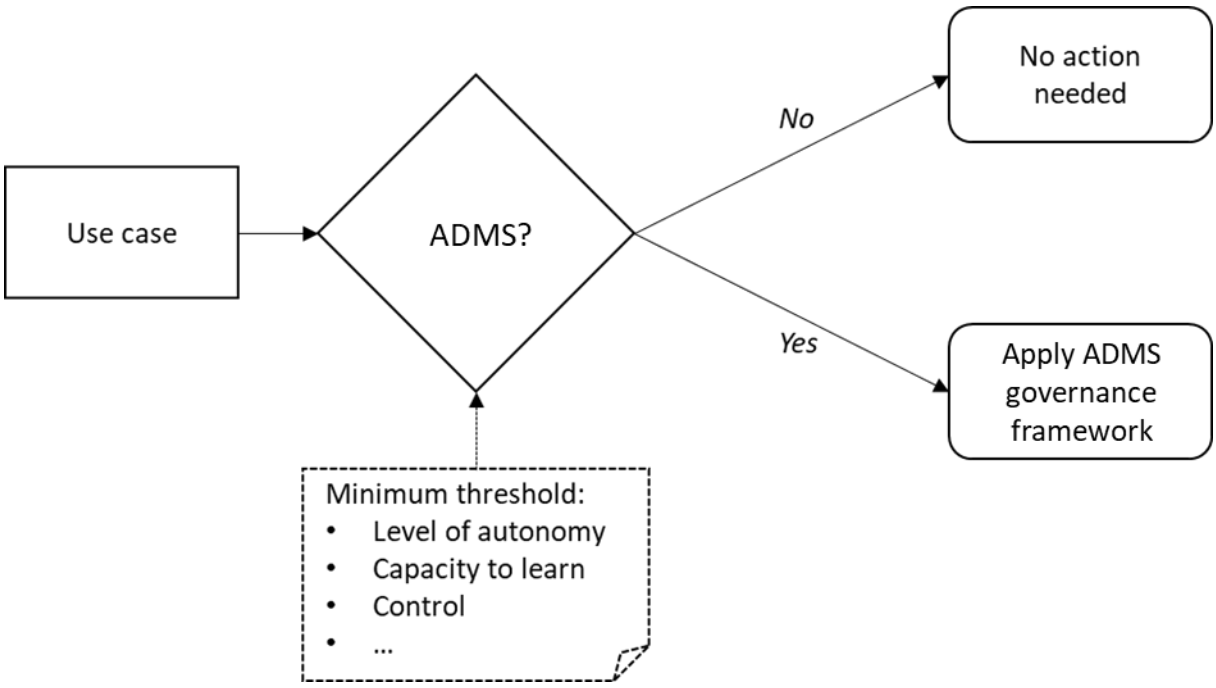
This working definition highlights two central features of ADMS: the level of *autonomy* and the ability to *adapt*. Of course, both are a matter of degree. In some cases, ADMS act with complete autonomy, whereas in other cases, they *only* provide recommendations to a human operator who has the final say (Cummings, 2004). Although it is a simplification, the IEEE's working definition is based on features that are directly linked to the specific ethical concerns posed by ADMS. It also enables practitioners to determine what is *not* within the ADMS

[125] The IEEE uses different definitions of ADMS in different contexts. Here, I am talking explicitly about the definition included in IEEE 7010-2020, which was the first industry standard published in the field of ADMS.
[126] Here, the IEEE does not use the term ADMS but rather Autonomous and Intelligent Systems (A/IS).

governance framework's scope. For example, this definition does not cover expert systems that structure information for the convenience of human decision-makers.

The logic behind the IEEE's working definition of ADMS can be abstracted into what I call the Switch. The Switch is a model for binary classifications: something either *is* or *is not* an ADMS. To establish such a threshold, the Switch consists of one or more *essential requirements*. These requirements can concern technical features (i.e., referring to what the system *is*) and functional aspects (i.e., referring to what the system *does*). Under ideal circumstances, simple yes/no questions are enough to determine whether a specific system satisfies the relevant requirement(s). Essential requirements are individually necessary and jointly sufficient. For my purposes, this means that an ADMS governance framework should apply to any use case that meets the requirements constituting the Switch. Figure 12 illustrates the logic behind the Switch approach.

Figure 12. The Switch – a binary approach to classifying ADMS with the use of thresholds.



The model presented above is a significant simplification. It does not describe exhaustively the process of determining the material scope for a specific ADMS governance framework. Nor does the model necessarily reflect the IEEE's intention behind their working definition of ADMS. Nevertheless, the model helps sketch the logic behind high-level, binary approaches to classifying systems and use cases for the purpose of ADMS governance.

The IEEE's working definition is a good example because it allows for meaningful evaluation of specific use cases. However, not all high-level working definitions of ADMS can

serve the function of the Switch as described above. For instance, in their report *Automating Society*, AlgorithmWatch (2019) takes a more holistic approach:

> *'[An ADMS] … is a socio-technological framework that encompasses a decision-making model, an algorithm that translates this model into computable code, the data this code uses as an in-put – either to "learn" from it or to analyse it by applying the model – and the entire political and economic environment surrounding its use.'* (AlgorithmWatch, 2019, p.9)

Taking a socio-technical systems approach, the working definition of an ADMS provided by AlgorithmWatch has merits. When technical subsystems are targeted separately, essential dynamics of the system as a whole may be lost or misunderstood (Di Maio, 2014). However, my aim here is not to compare the relative merits of different definitions of ADMS. Instead, I wish to exemplify how some high-level definitions (the IEEE's included) can help organisations demarcate the material scope of the ADMS governance frameworks they seek to implement. In contrast, working definitions of ADMS that take '*the entire political and economic environment'* into account do not.

In conclusion, the Switch is an intuitive model to classify systems and use cases for the purpose of ADMS governance. Its strength lies in the fact that it is easy for practitioners to remember and for organisations to communicate internally and externally. However, it is also a coarse approach that is likely to result in material scopes that are either under- or overinclusive. Hence, classifications based on the Switch are feasible only in combination with flexible ADMS governance frameworks that, following an initial assessment, escalate only those use cases that demand further scrutiny.

6.5 The Ladder

A central function of ADMS governance frameworks is to put mechanisms in place that ensure accountability for ADMS and their outcomes, both before and after their implementation (AI HLEG, 2019). Because ADMS may exacerbate existing risks and introduce new ones, ADMS governance is closely linked to risk management, i.e., processes that allow different risks to be identified, understood, and managed (Leslie, 2019). According to ISO 31000 risk management guidelines (2018), *risk* is the effect of uncertainty on objectives. So understood, risks can be ethical, legal, or technical.

Faced with constant pressures to reduce uncertainty, most technology providers already have risk management systems in place (Currie, 2019). Hence, the use of ADMS does not necessarily require a complete overhaul of existing governance structures but rather an

awareness of how ADMS may increase, or complicate the detection of, risks as they manifest themselves in unfamiliar ways (Lee et al., 2020). In short, implementing ADMS governance entails adopting measures to mitigate the ethical risks posed by a specific system in a manner proportionate to the magnitude of those risks.

Building on this rationale, a growing number of proposals have advocated a risk-based approach to ADMS governance and hence to the classification of ADMS systems (Krafft et al., 2020b).[127] Most notable amongst these proposals is the draft AIA, which takes an explicitly risk-based approach (European Commission, 2021a). However, to exemplify the logic behind the approach, I will focus my analysis on the recommendation of the German Data Ethics Commission (DEK). The reason for this is that the DEK, as we shall see, outlined the risk-based approach to classifying ADMS in an outright and pedagogical manner.

In 2018, the DEK called for a risk-based approach to ADMS governance that would range from *no regulation* for the most innocuous ADMS to *a complete ban* for the most dangerous ones (DEK, 2018). In between these two extremes, the DEK defined three intermediary risk levels for which the use of ADMS is generally allowed but subjected to increasingly stringent governance requirements. Table 2 summarises the five-level classification of ADMS for the purpose of ADMS governance as proposed by the DEK.[128]

Table 2. The DEK's five-level classification of ADMS, based on their potential for harm.

Level	***Potential for harm***	***Implications for ADMS governance***
1	Applications with *zero* or *negligible* potential for harm	No specific measures
2	Applications with *some* potential for harm	Measures such as formal and substantive requirements or monitoring procedures
3	Applications with *regular* or *significant* potential for harm	Additional measures such as ex-ante approval procedures
4	Applications with *serious* potential for harm	Additional measures such as a live interface for 'always-on' oversight by supervisory institutions
5	Applications with an *untenable* potential for harm	Complete or partial ban of an ADMS

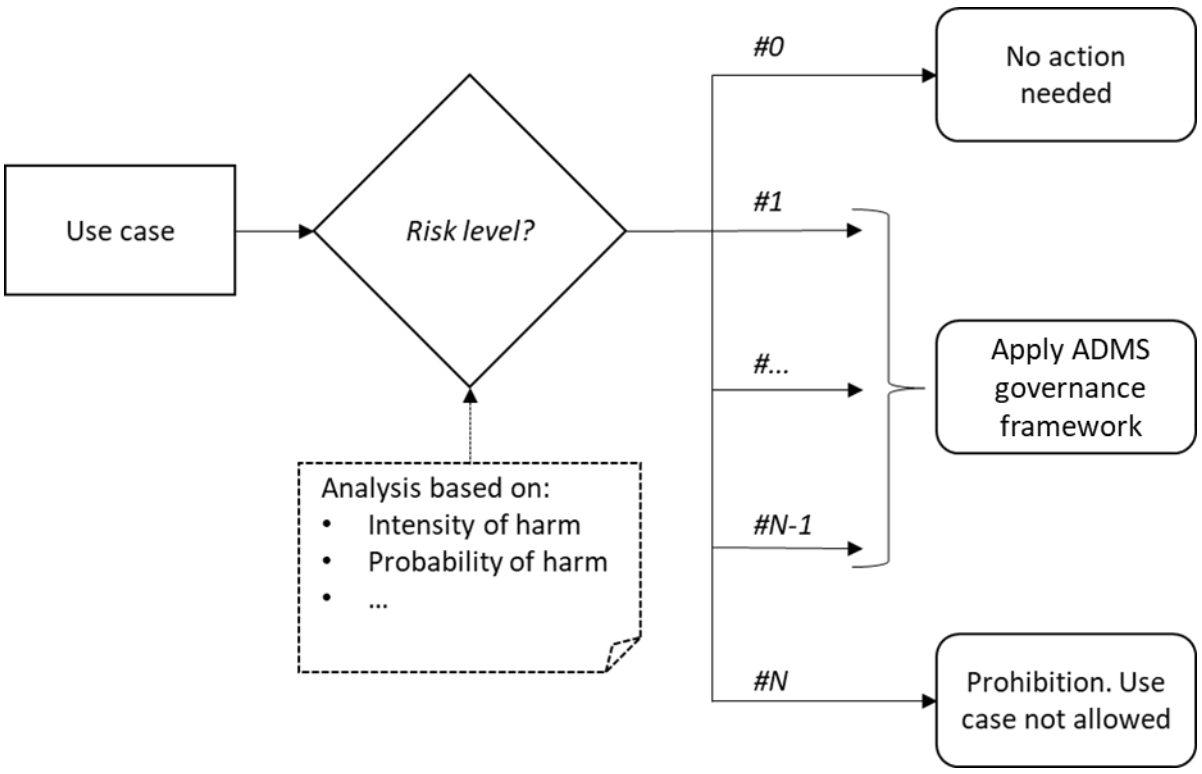
The *potential for harm* is the key variable underpinning this classification of ADMS. The DEK suggested that the potential for harm can be determined by looking at the *likelihood* that harm will occur and the *severity* of that harm (DEK, 2018). Admittedly, determining the potential

[127] In the US, most well-known example is the *NIST AI risk management framework* (NIST, 2022b).
[128] The DEK uses the term 'algorithmic systems.'

for harm is itself a non-trivial undertaking. I will revisit this question later in this section. Here, it is sufficient to note that international standards have long been used to codify risk assessments for many socio-technical issues, from cybersecurity to data privacy (ISO, 2019).

The risk-based approach taken by the DEK (and by the EU) can be abstracted into a model for classifying ADMS for governance purposes. I have chosen to call this model the Ladder. The idea behind the Ladder is to classify use cases into different risk levels. Depending on the ethical risks posed by a specific ADMS, different levels of governance apply. Both the number of steps in the Ladder and the methods used to determine the risks posed by a specific system can vary between different contexts. However, the basic principle remains simple: the greater the potential for harm, the more far-reaching the prescribed interventions. Figure 13 illustrates the generalisable logic behind the Ladder approach.

Figure 13. The Ladder – a risk-based approach to classifying ADMS.



Apart from the DEK and the European Commission, many other organisations have proposed risk-based approaches to ADMS governance. For example, the AI Ethics Impact Group (AIEIG) has developed a framework to help organisations operationalise ADMS governance. The main idea behind their framework is to create a (standardised) context-independent labelling of ADMS inspired by the energy efficiency label (AIEIG, 2020). For their proposed 'Ethics Label', AIEIG assumes a set of ethics principles, including accountability,

transparency, and reliability. Each principle is then broken down into criteria that define when values are honoured or violated. Significantly, these criteria rest on observables that can be monitored and quantified. This level of detail allows the AIEIG to be more specific than the DEK about the level of governance required for ADMS at each risk level. For example, the AIEIG distinguishes between the need for outcome accountability, objective-based accountability, and process accountability for ADMS at different intermediary risk levels.

But how are individual use cases classified into different risk levels? As mentioned above, the most common way to conceptualise risk is to combine the severity of harm and the likelihood of it occurring (Krafft et al., 2020b). These two elements have previously been used to construct risk-ladders that allow for matching regulatory provisions to different risks, such as financial (MacNeil & O'Brien, 2010) or environmental risks (Black & Baldwin, 2012). The severity of harm depends on the task performed by the ADMS and on what the possible outcomes are. For instance, ADMS used for consumer recommendations may have a lower effect on human welfare than those used for recruitment or medical interventions. In contrast, the likelihood of harm can be thought of in terms of probability. However, there is often uncertainty when evaluating real use cases, and exact probabilities are seldom known. Hence, it is important to remember that the purpose of the Ladder, as outlined above, is not to identify hard thresholds between different risk levels but rather to provide practical guidance on how to classify ADMS in ways that are fit for purpose, simple and clear, and stable over time.

A further tension relates to the conflicting incentives actors within organisations may have when self-determining the risks associated with specific systems or use cases (The Government Office for Science, 2014). For example, executives facing pressure to cut costs may want to avoid that red flag are being raised too often. Similarly, individual managers or developers may prefer to handle the risks involved in a project locally, rather than escalating issues that might either kill the project or subject it to more administration and oversight.

To summarise, the Ladder is a purposeful and highly resilient model for classifying ADMS for governance purposes. By focusing on the risks posed by specific use cases, the Ladder presents a technology-agnostic framework that addresses the ethical risks posed by autonomous, self-learning systems. Its proven format also facilitates the integration of ADMS governance frameworks into existing organisational structures and processes. However, the Ladder requires practitioners to make an upfront risk assessment of individual use cases. This extra burden – combined with the uncertainty associated with the concept of risk – means that, compared with the Switch, the Ladder may be more costly and difficult to operationalise.

6.6 The Matrix

As discussed in Section 6.2, ADMS – although hard to define – typically share several characteristics. These characteristics include the ability to *perceive* the environment through input data, *process information* to interpret the data, *make decisions* and based on the data to achieve pre-defined goals (Samoili et al., 2020). Further, because ADMS are embedded in conditions that include and exceed them (Reddy et al., 2019), their effects on society are not determined solely by their design. For example, the extent to which an ADMS' behaviour is perceived as ethical also depends on the input data, which could be incomplete or biased (Kim, 2017), and the task for which the system is used (Lauer, 2020). Hence, the impact an ADMS has on its environment is not always intuitive and may not be consistent over time. Rather, every individual use case poses a unique set of ethical challenges.

To deal with this complexity, many organisations use multiple dimensions to classify ADMS (Wilson et al., 2020). For example, the OECD (2022) has recently published a framework designed to help organisations classify ADMS. The purpose of this framework is to help organisations and policymakers assess and classify ADMS systems according to their potential ethical implications. As a starting point, the OECD's framework requires organisations to account for four dimensions when classifying ADMS:

1) *Context*: the socio-economic environment in which the ADMS is deployed, notably the sector and the potential impact the system may have,
2) *Data and input*: the data used by the ADMS to build a representation of its environment,
3) *Computational model*: the real-world processes in the system's internal environment that constituting the core of the ADMS, and
4) *Task and output*: resulting actions taken by the ADMS to influence its environment.

Each of these four key dimensions is broken down into subdimensions (see Table 3, below).

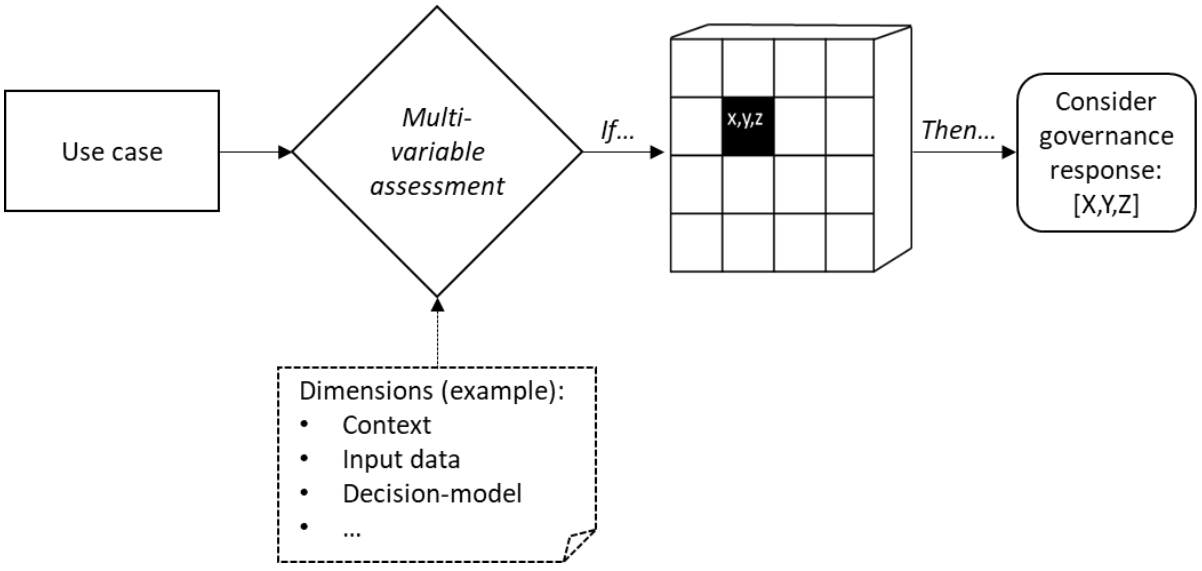
Table 3. Subdimensions of the OECD's framework for classifying ADMS.

1. Context	***2. Data and input***	***3. Computational model***	***4. Task and output***
Sector	Collection	Model type	Task
Scale	Structure	Model properties	Actions
Impact	Type	Acquisition of	Composite system
Critical	Quality	capabilities	Considerations
User	Collection		

Some of the subdimensions are binary. The *user* of the system, for instance, is either an expert or a non-expert. Other subdimensions are categorical. One example here would be the *input data*, which may be structured, unstructured, or semi-structured. Yet other subdimensions, like explainability (Watson & Floridi, 2020), demand further clarification regarding their purpose and target audience and thus allow for free-form entries. As a whole, the framework attempts to capture all the most relevant aspects of ADMS to allow for effective governance.

The approach taken by the OECD represents a model to classify ADMS that I have chosen to call the Matrix.[129] In essence, the Matrix refers to classifications of ADMS that take several different dimensions into account to identify the specific ethical risks associated with a particular use case. The number and scope of these dimensions can vary between different Matrix classification schemes. At least, these dimensions include the *technical characteristics* a system has, which *data* it can access, and the *context* in – and the purpose for which – the system is applied. For each combination in the Matrix, an ADMS governance framework can specify which preventive measures are needed to ensure that a particular system is designed and deployed in ethical ways. Figure 14 illustrates the logic behind this approach.

Figure 14. The Matrix - a multi-dimensional classification of ADMS.



[129] Strictly speaking, a *matrix* is a special case of a *tensor*, i.e., a tensor of rank two. However, since higher order tensors are often represented as matrices, e.g., by column-binding data from later time points, I am not overly concerned with this distinction and use the term *matrix* to refer to multidimensional spaces of arbitrary rank.

Due to this structure, the Matrix represents the most comprehensive, but also the most complicated, model for how to classify ADMS. To illustrate how quickly the model grows in complexity, let us consider how the context of an ADMS can be assessed. The impact of a specific ADMS may depend on several interrelated factors, including the breadth of deployment of the system, its maturity, the stakeholders with which it interacts, and the purpose for which it is applied. In theory, a broader deployment is coupled with greater risks than a narrower one, and a more mature ADMS poses fewer risks than a less mature one. Further, the use of ADMS poses different ethical challenges in different sectors. Healthcare, for instance, is often considered a critical sector since even minor disruptions may have severe consequences for the health and safety of human beings. However, not all ADMS in a critical sector are critical. For example, a hospital's administrative time tracking systems need not necessarily be considered a critical system. This shows how evaluations in one dimension inevitably have implications for other dimensions in the Matrix.

In terms of input, good data governance demands that data sets are complete, accurate, and appropriate, whilst not being collected or processed in ways that infringe on individual privacy or IP rights (ICO, 2018). However, concerning data management and information processing, the Matrix does need to not offer any new tools per se. Rather, it incorporates tools like model cards (Mitchell et al., 2019) and datasheets (Gebru et al., 2018) into a structured process for responding to the specific challenges associated with a given ADMS. The key to this approach is the link between a particular use case's characteristics and the proposed remedies. To give an example, evaluating whether decisions made by ADMS are unfair or discriminatory typically use one, or a combination, of three distinct methods: *pre-processing*, *in-processing*, and *post-processing* (Clavell et al., 2020). Understanding about the way data is collected and structured informs pre-processing methods. In-processing, by contrast, depends on the model itself. Finally, post-processing techniques keep the model as it is but adjust the outcomes so that they correspond better to some predefined normative benchmark.

To summarise, the Matrix is the most comprehensive approach to classifying ADMS for governance purposes. The fact that it attempts to take 'all relevant information into account' means that the Matrix can help organisations that design and deploy ADMS understand which precautions are necessary for a particular use case. At the same time, the Matrix require substantial upfront investments, especially in terms of practitioners' time. Moreover, the different dimensions according to which ADMS are evaluated sometimes overlap and influence each other non-linearly. Hence, the Matrix is not a simple model to implement. Finally, its all-encompassing approach also makes the Matrix less resilient to changes over time. In short,

classifications based on the Matrix hold great promise to inform policy decisions, but they may not necessarily help organisations define the material scope of ADMS governance.

6.7 Discussion

In this section, I will evaluate the three models for classifying ADMS introduced in this chapter according to the criteria in 6.3, i.e., that good classifications should be (i) fit for purpose, (ii) simple and clear, and (iii) stable over time.

To begin with, a significant advantage of the Switch is that it is simple and clear. This is because a few essential requirements are easy to bear in mind yet still provide a basis for identifying the systems to which a given ADMS governance framework applies. Classifications of ADMS based on the Switch are thus easy to grasp and use, even for non-experts. Yet the Switch may not always be fit for purpose, since all binary approaches to classifying ADMS – which draw sharp lines across continuous spectrums – struggle to strike the right balance between over and under inclusiveness. At first glance, it may appear better to err on the side of caution and employ an overinclusive Switch. Doing so would increase the probability of identifying and managing high-risk use cases. However, care must also be taken not to cause unjustifiable administrative burdens.[130] Finally, the extent to which a Switch is stable over time depends on which elements are included in the essential requirements. While requirements that refer to a particular design structure (e.g., neural networks) or a specific use case (e.g., facial recognition) are likely to change over time, essential requirements that refer to capabilities would be more permanent (Schuett, 2021).

Compared with the Switch, the Ladder has several advantages. First, it builds on well-established mechanisms like risk assessments. This procedural continuity helps organisations integrate ADMS governance frameworks into their existing governance structures and QMS (Raji et al., 2020). Second, the Ladder is fit for purpose insofar as it demands that organisations assess the specific ethical challenges associated with the technical systems they design and deploy. This is important given that, even where they are technically similar, the consequences of the decisions made by ADMS may differ considerably depending on the concrete setting in which they are applied (Krafft et al., 2020a). Finally, the fact that it is technology-agnostic

[130] As a case in point, the Centre for Data Innovation recently estimated that the EU AIA will cost the European economy €31 bn over the next five years and reduce R&D investments by up to 20 percent (Mueller, 2021). How precise such numbers are is hard to say. Still, they serve as a healthy reminder that not only underinclusive but also overinclusive definitions of the material scope of ADMS governance are associated with real costs and risks.

makes the Ladder a highly stable model for classifying ADMS. Governance frameworks that classify ADMS according to the Ladder are thus well equipped to handle both rapid technological innovation and social change. But these advantages come at a cost: it can be very challenging to assess all the ethical risks posed by a specific use case in practice.

The Matrix is fit for purpose insofar as it informs the policy considerations associated with different types of ADMS. Classifications based on the Matrix also help organisations identify which precautionary measures are appropriate when designing or implementing a specific ADMS. However, the Matrix is not a simple model for classifying ADMS. It does not help practitioners determine with little effort whether a particular use case should be subjected to ADMS governance. Instead, the Matrix front-loads both the administrative burden – and the process of ethical deliberation – to the initial stages of software development processes. Finally, classifications based on the Matrix may not be stable over time. Many subdimensions consist of variables whose categories or ranges may change due to future technological advances. Computer science research is a quickly moving landscape and speculating about what technical breakthroughs may occur is beyond the scope of this chapter. The point is that any model that attempts to exhaust all possible combinations of available technologies, on the one hand, and potential areas of application, on the other, will struggle to stay relevant and useful.

It should be re-emphasised that different ways of classifying ADMS can be combined and used by the same organisation at various stages of the governance process. An example of this possibility is found in the first draft of the AIA. As a first step, the AIA used a Switch to demarcate its material scope. As mentioned in Section 6.2, the list of computational techniques coverer by the AIA is broad and include everything from logic-based to statistical approached. In a second step, the AIA relies on the Ladder, i.e., on a risk-based approach, to identify those ADMS that need to be subjected to additional oversight and transparency obligations.[131]

In practice, overlaps such as the one described above are common for at two reasons. First, because the Switch, the Ladder, and the Matrix have different (and complementary) strengths and weaknesses, organisations can tailor their material scope by combining the three models for classifying ADMS in different ways. Second, the three models compose a hierarchy of complexity in which each constitutes a special case of the previous: the Switch can be viewed as a bivalent Ladder, and the Ladder as a rank-one Matrix. Despite these overlaps, treating the

[131] This is not the place to discuss the merits and shortcomings of the material scope used in the EU AIA. Readers interested in such a discussion are referred to Bryson (2022).

three models for classifying ADMS as conceptually distinct is useful since it enables a comparison of their respective affordances and constraints. In this chapter, I have therefore chosen to identify each model with the minimum level of complexity required to describe its operations – well aware of the fact that edge cases could strain the typology.

6.8 Concluding remarks

All governance mechanisms need to define their material scope, and EBA is no exemption. In Chapter 1, I introduced EBA as a structured procedure whereby an entity's behaviour is assessed for consistency with relevant ethics principles. To be feasible and effective, EBA procedures presuppose clarity about which systems or processes, exactly, these ethics principles apply to. However, the fact that there exists no universally accepted definition of ADMS need not be a problem. As I have argued in this chapter, it is less important to define what an ADMS is in abstract terms and more important to establish processes for identifying those systems or processes that require additional layers of governance. This brings me to:

SQ4 How can the material scope of EBA be demarcated?

In this chapter, I have shown that there are at least three different ways of classifying ADMS for governance purposes: *the Switch*, i.e., a binary approach according to which systems either are or are not considered ADMS depending on their intrinsic characteristics; *the Ladder*, i.e., a risk-based approach that classifies systems according to the ethical risks they pose; and *the Matrix*, i.e., a multi-dimensional approach that accounts for various aspects, like input data, decision-model, and decision task, when classifying ADMS. Each of these models (as well as any combination of them) can be used to demarcate the material scope of EBA procedures.

Each of these models comes with its own sets of strengths and weaknesses. The Switch is simple and clear – and can thus be easily communicated and applied. At the same time, classifications of ADMS based on the Switch are often under- or overinclusive. In contrast, the Ladder provides a technology-neutral model for classifying ADMS that is fit for purpose and stable over time. However, although the risk-based approach on which the Ladder is built facilitates the integration of ADMS governance into existing governance structures, it also adds complexity that makes it more difficult to implement than the Switch. Finally, the Matrix offers the most comprehensive model for how to classify ADMS. As such, it is well-suited to inform policy decisions. On the downside, classifications based on the Matrix add significant administrative burdens on organisations and practitioners and are also less stable over time.

Two further conclusions can be drawn from this discussion. First, there is a three-way trade-off between how fit for purpose a model for classifying ADMS is, how simple it is to apply, and how stable it is over time. When to use the Switch, the Ladder, or the Matrix to demarcate the material scope of EBA procedures thus remains a question to be evaluated locally, case by case. Second, how ADMS are classified is an integral part of the design of EBA procedures. EBA procedures that do not demarcate their material scope are incomplete, and hence unlikely to be effective in identifying and mitigating the ethical risks ADMS pose.

Some may remain sceptical about whether it is necessary to classify ADMS at all. They may rightly point to the fact that ADMS are just a variation of other systems, with which they share many characteristics and ethical risks. However, human organisation is made possible by conceptual representations of the world taken at a relatively *high level of abstraction* rather than the way the world *really is*. To quote (Dewey, 1920) once more, 'A classification is not a bare transcript of some finished and done-for arrangement pre-existing in nature. It is rather a repertory of weapons for attack upon the future and the unknown.'

Following Dewey's impetus, I hope that this chapter may serve as a map for those who seek to design or implement EBA procedures. My hope rests on the old idea that new insights can be gained through a multiplicity of perspectives. Mental models help us organise information to grasp complex phenomena, and the many-model approach helps illuminate the blind spots inherent in each model (Page, 2018). By drawing on the mental models outlined in this chapter – i.e., the Switch, the Ladder, and the Matrix – I hope that technology providers and auditors will be better equipped to classify ADMS and implement EBA procedures.

As I have shown in this Chapter, how ADMS are classified have direct implications on the feasibility and effectiveness of different governance mechanisms. One example on this is provided by the ongoing policy dialogue concerning the EU AIA. It is less than two years since the European Commission published (2021) the first draft of the AIA, and yet recent technological developments – including the emergence of large language models (LLMs) and other *foundation models* (Bommasani et al., 2021) – have already put pressure on policymakers to revise the material scope of the proposed legislation (FLI, 2022). The same holds true for EBA procedures. Most existing EBA procedures have been designed to audit ADMS used for specific purposes in predictable environments. In Chapter 7, I address that gap by outlining a blueprint for how to audit ADMS with highly general capabilities.

CHAPTER 7

ETHICS-BASED AUDITING OF LARGE LANGUAGE MODELS: A THREE-LAYERED APPROACH

Abstract

Large language models (LLMs) represent a major advance in natural language processing and computer science research. At the same time, the widespread use of LLMs is coupled with significant ethical and social challenges. Previous research has pointed towards *ethics-based auditing* (EBA) as a promising governance mechanism to help identify and mitigate the ethical and social risks posed by automated decision-making systems (ADMS). However, existing EBA procedures fail to address the governance challenges posed by LLMs, which display emergent capabilities and are adaptable to a wide range of downstream tasks. In this chapter, I address that gap by outlining a novel blueprint for how to audit LLMs. Specifically, I propose a three-layered approach, whereby *governance audits* (of technology providers that design and disseminate LLMs), *model audits* (of LLMs after pre-training but prior to their release), and *application audits* (of applications based on LLMs) complement and inform each other. I show how EBA, when conducted in a structured and coordinated manner on all three levels, can be a feasible mechanism for identifying and managing some of the ethical and social risks posed by LLMs. However, it is important to remain realistic about what EBA can be expected to achieve. Therefore, I discuss the limitations not only of my three-layered approach but also of the prospect of auditing LLMs at all. Ultimately, this chapter seeks to sharpen and extend the conceptual toolkit available to organisations who wish to audit the design and use of ADMS with highly general capabilities for alignment with specific ethics principles.

Note

This chapter is based on a peer-reviewed journal article published in *AI and Ethics* (see Mökander et al, 2023b).[132] I have lightly edited the text to harmonise the vocabulary used in the chapter to fit the overarching framing of this thesis. I have also shortened parts of the original manuscript to minimise overlaps with the other chapters included in this thesis.

[132] The article was co-authored with Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Please see Appendix 3, 8, and 9 for authorship statements.

7.1 Introduction

7.1.1. Background

Ethics-based auditing (EBA) is a governance mechanism that technology providers and policymakers can use to identify and mitigate risks associated with *automated decision-making systems* (ADMS) (Brundage et al., 2020; Sandvig et al., 2014). Operationally, EBA is characterised by a structured process whereby an entity's behaviour is assessed for consistency with relevant principles. Previous work on EBA has focused on developing procedures to assess ADMS used for specific tasks in predictable environments for adherence with technology providers' organisational values or sector-specific norms. For example, researchers have developed procedures for how to audit ADMS used in recruitment (Kazim et al., 2021), online searches (Robertson et al., 2018), and medical diagnostics (Liu et al., 2022; Oakden-Rayner et al., 2022). However, the capabilities of ADMS tend to become ever more general.

In a recent article, Bommasani et al. (2021) coined the term *foundation models* to describe ADMS that can be adapted to a wide range of downstream tasks. While foundation models are not new from a technical perspective,[133] they differ from other ADMS as they are effective across many different tasks and display emergent capabilities when scaled. The rise of foundation models also reflects a shift in how ADMS are designed, since these models tend to be trained by one actor and subsequently adapted for different applications by a plurality of other actors (Bommasani & Liang, 2021). From an EBA perspective, foundation models pose significant challenges. To begin with, it is difficult to assess the risks ADMS pose independent of the context in which they are deployed. Moreover, how to allocate responsibility between technology providers and downstream developers when harms occur remains unresolved. Taken together, the capabilities and training processes of foundation models have outpaced the development of procedures to assess whether these align with prespecified ethics principles.

7.1.2 Scope and contributions

In this chapter, I address SQ5, i.e., what could blueprints for feasible and effective EBA procedures look like for ADMS with highly general capabilities? In doing so, I focus on large language models (LLMs), a subset of foundation models, for reasons I explain below.

---

[133] Foundation models are based on deep neural networks and supervised learning. Their recent rise has been enabled by the development of new transformer architectures (Vaswani et al., 2017); the increase in compute resources (Smith-Goodson, 2022); and the availability of large-scale datasets (Luccioni & Viviano, 2021).

LLMs start from a source input, called the prompt, to generate the most likely sequences of words, code, or other data (Floridi & Chiriatti, 2020). Historically, different model architectures have been used in natural language processing (NLP) (Rosenfeld, 2000). However, most recent LLMs are based on deep neural networks trained on a large corpus of texts. Examples of such LLMs include GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), LaMDA (Thoppilan et al., 2022), and Gopher (Rae et al., 2022). Once an LLM has been pre-trained, it can be adapted (with or without fine-tuning)[134] to support various applications, from spell-checking (Hu et al., 2021) to creative writing (Hsieh, 2019).

Developing EBA procedures for LLMs is an important and timely task for two reasons. First, LLMs pose many ethical challenges, including the perpetuation of harmful stereotypes, the leakage of personal data, the spread of misinformation, and plagiarism (Bender et al., 2021; Perez et al., 2022b; Shelby et al., 2022). The urgency of addressing those challenges makes developing procedures to audit LLMs along different normative dimensions (such as privacy, bias, safety, etc.) a critical task in and of itself (Liang et al., 2022). Second, LLMs can be considered proxies for other foundation models. Consider CLIP (Radford et al., 2021), a vision-language model trained to predict which text caption accompanied an image, as an example. CLIP too displays emergent capabilities, can be adapted for multiple downstream applications, and faces similar governance challenges as LLMs. The same holds true for text2image models such as DALL·E 2 (Ramesh et al., 2022). The process of developing EBA procedures for LLMs is thus likely to offer transferable lessons on how to audit other foundation models.[135]

The main contribution offered in this chapter is a novel blueprint for how to audit LLMs. Specifically, I propose a three-layered approach, whereby *governance audits* (of technology providers that design and disseminate LLMs), *model audits* (of LLMs after pre-training but prior to their release), and *application audits* (of applications based on LLMs) complement and inform each other. The key message I stress is that to be feasible and effective, audits conducted on the governance, model, and application levels must be combined into a structured and coordinated EBA procedure. To the best of my knowledge, the blueprint for how to audit LLMs outlined in this chapter is the first of its kind, and I hope it will inform both

[134] To fine-tune LLMs for specific tasks, an additional dataset of in-domain examples can be used to adapt the final layers of a pre-trained model.

[135] Following Jonathan Zittrain (2014), I define 'generative technologies' as technologies that allow third-parties to innovate upon them without any gatekeeping. Colloquially, 'generative AI' sometimes refers to systems that can output content (e.g., images, text, audio), but that is not how I use the term in this chapter.

technology providers' and policymakers' efforts to design and implement EBA procedures for ADMS with highly general capabilities.

In the process of introducing and discussing the three-layered approach, I offer two secondary contributions in this chapter. First, I derive seven claims about how LLM auditing procedures should be designed to be feasible and effective in practice. Second, I identify the conceptual, technical, and practical limitations associated with auditing LLMs. Together, these secondary contributions lay a groundwork that other researchers and practitioners can build upon when designing new, more refined, LLM auditing procedures in the future.

My efforts tie into an ongoing policy formation process. Technology providers like OpenAI, DeepMind, Microsoft, and Anthropic have highlighted the need for new governance mechanisms to address the social and ethical challenges that LLMs pose (Ganguli et al., 2022a; Peyrard et al., 2021). Individual parts of my proposal (e.g., those related to model evaluation (Chowdhery et al., 2022) and red teaming (Perez et al., 2022a))[136] have already started to be implemented across the industry, although not in a structured and transparent manner. Policymakers, too, are interested in ensuring that societies benefit from LLMs while managing the associated risks. Examples of proposals to regulate ADMS include the AIA (European Commission, 2021a) and the US Algorithmic Accountability Act of 2022 (Office of US Senator Ron Wyden, 2022). My blueprint for how to audit LLMs neither seeks to replace existing best practices for training and testing LLMs nor to foreclose forthcoming regulations. Rather, it complements them by showing how EBA – when conducted in a structured and coordinated manner – can help identify and mitigate the ethical risks LLMs pose.

A further remark is needed to narrow down this chapter's scope. My three-layered approach concerns the *procedure* of ethics-based LLM audits and answers questions about *what* should be audited, *when*, and according to *which criteria*. Of course, when designing holistic EBA procedures, several additional considerations exist, e.g., *who* should conduct the audit and *how* to ensure post-audit action (Raji et al., 2022). While such considerations are important, they fall outside the scope of this chapter. How to design an institutional ecosystem to audit LLMs is a non-trivial question that I have neither the space nor the capacity to address here. That said, the policy process required to establish EBA procedures for ADMS with highly

[136] A 'red team' is a group of people authorised to emulate an adversarial attack on an ADMS to identify and exploit its vulnerabilities (NITS, 2023).

general capabilities will likely be gradual and involve negotiations between numerous actors, including technology providers, policymakers, and civil rights groups.

The remainder of this chapter proceeds as follows. Section 7.2 highlights the ethical risks LLM pose and establishes the need to audit them. Section 7.3 describes the methodology used in this chapter. Section 7.4 reviews previous literature on EBA, discusses the properties of LLMs that undermine existing EBA procedures, and derives seven claims for how feasible and effective EBA procedures for LLM should be designed. Section 7.5 outlines a blueprint for how to audit LLMs. Specifically, a three-layered approach that combines governance, model, and application audits is proposed. The section explains why these three types of audits are needed, what they entail, and the outputs they should produce. Section 7.6 discusses the limitations of my three-layered approach and demonstrates that any attempt to audit LLMs will face several conceptual, technical, and practical constraints. Section 7.7 discusses the implications of my findings for researchers, policymakers, and industry practitioners. Finally, Section 7.8 concludes by generalising the three-layered approach outlined in this chapter into a blueprint for how to audit ADMS with highly general capabilities.

7.2 The need to audit LLMs

This section summarises previous research on LLMs. It situates my research in relation to recent technological and societal developments and stresses the need for EBA procedures that capture the ethical risks LLMs pose.

7.2.1 The opportunities and risks of LLMs

Although LLMs represent a major advance in computer science research, the idea of building text-processing machines is not new. Since the 1950s, NLP researchers and practitioners have been developing software that can analyse, manipulate, and generate natural language (Joshi, 1991). Until the 1980s, most NLP systems used logic-based rules to enable machine translation and speech recognition (Hirschberg & Manning, 2015). More recently, the advent of deep learning, advances in neural architectures like transformers, growth in computational power and the availability of internet-scraped training data have revolutionised the field (Chernyavskiy et al., 2021). Further advances in instruction-tuning and reinforcement learning from human feedback have improved model capabilities to predict user intent and respond to natural language requests (Bai et al., 2022; Ouyang et al., 2022; Stiennon et al., 2020).

LLMs' core training task is to produce the most likely continuation of a text sequence (y Arcas, 2022). Consequently, LLMs can be used to recognise, summarise, translate, and

generate texts (Suzgun et al., 2022). Exactly when a language model becomes 'large' is a matter of debate – referring to either more trainable parameters (Villalobos et al., 2022), a larger training corpus (Hoffmann et al., 2022) or a combination of these. Notably, LLMs require fewer in-domain labelled examples than traditional deep learning systems to perform well on unseen tasks (Bowman, 2023). This means that LLMs can more easily be adapted for various downstream applications, such as diagnosing medical conditions (Rasmy et al., 2021), generating code (Chen et al., 2021), and translating languages (Wang et al., 2021b).

LLMs bring new opportunities. Because LLMs perform well on a wide range of task, they can be used to automate specific repetitive workflows (Kojima et al., 2022). Moreover, a scaling law has been identified whereby the training error of an LLM falls off as a power of training set size, model size or both (Kaplan et al., 2020). This means that scaling LLMs further can result in emergent capabilities and increased performance on a wide array of tasks (Srivastava et al., 2022). Finally, while some pre-trained models are protected by paywalls or siloed within companies, many LLMs are accessible via open-source libraries such as HuggingFace, allowing non-experts to use it in their applications (Kirk et al., 2021).

Alongside such opportunities, however, the use of LLMs is coupled with ethical challenges. As recent controversies surrounding ChatGPT (Azaria, 2022) have shown, LLMs are prone to give biased or incorrect answers to user queries (Borji & Ai, 2023). More generally, a recent article by Weidinger et al. (2021) suggests that the ethical challenges associated with LLM can be clustered into six broad risk areas:

1) *Discrimination*: LLMs can introduce representational and allocational harms by perpetuating social stereotypes and biases,
2) *Information hazards*: LLMs may compromise privacy by leaking private information and inferring sensitive information,
3) *Misinformation hazards:* LLMs producing misleading information can lead to less well-informed users and erode trust in shared information,
4) *Malicious use*: LLMs can be co-opted by users with bad intent, e.g., to generate personalised scams or large-scale fraud,
5) *Human-computer interaction harms*: users may overestimate the capabilities of LLMs that appear human-like and use them in unsafe ways, and
6) *Automation and environmental harms:* training and operating LLMs require lots of computing power, incurring high environmental costs.

Each of these risk areas constitutes a vast and complex field of research. Providing a comprehensive overview of each field's nuances is beyond this chapter's scope. Instead, I take Weidinger et al.'s summary of the ethical and social risks associated with LLMs as a starting point for pragmatic problem-solving.

7.2.2 The governance gap

From a governance perspective, LLMs pose both methodological and normative challenges. LLMs are typically developed in two stages. Firstly, a model is pre-trained using self-supervised learning on a large, unstructured text corpus scraped from the internet. Pre-training captures the general language representations. Secondly, the weights or behaviours of this pre-trained model are adapted on a far smaller dataset of labelled, task-specific, examples. Although performance is predictable at a general level, performance on specific tasks can be unpredictable (Ganguli et al., 2022a). That makes it methodologically difficult to assess LLMs independent of the context in which they will be deployed (Bommasani et al., 2021).

Further, even well-functioning LLMs force technology providers and policymakers to face hard questions, such as who should have access to these technologies and for which purposes (Shevlane, 2022). Of course, the challenges posed by LLMs are not necessarily distinct from those associated with classical NLP or other deep-learning-based systems. However, LLMs' widespread use and generality make those challenges deserving of urgent attention. For all these reasons, analysing LLMs from ethical perspectives requires innovation in risk assessment tools, benchmarks, and frameworks (Tamkin et al., 2021).

Several governance mechanisms designed to ensure that LLMs adhere to predefined ethics principles have already been piloted (Avin et al., 2021). Some are technically oriented, including the pre-processing of the training data, the fine-tuning of LLMs on data with desired properties, and procedures to test the model at scale pre-deployment (Perez et al., 2022a). Others seek to address the ethical risks LLMs pose through sociotechnical mitigation strategies, like more diverse developer teams (PAI, 2020) and human-in-the-loop protocols (Wang et al., 2021c). Yet others seek to ensure transparency in development processes, e.g., through model cards (Derczynski et al., 2023), datasheets (Gebru et al., 2021), system cards (MetaAI, 2023), or the watermarking of system outputs (Kirchenbauer et al., 2023).[137]

[137] A watermark is a hidden pattern in a text that is imperceptible to humans but makes it identifiable as synthetic.

To summarise, while LLMs have shown impressive performance across a wide range of tasks, they also pose significant ethical risks. The question of how LLMs should be governed has thus attracted much attention, with proposals ranging from structured access protocols (Shevlane, 2022) to hard regulation prohibiting the use of LLMs for specific purposes (Hacker et al., 2023). However, the effectiveness and feasibility of these proposals have yet to be substantiated by empirical research. Moreover, given the multiplicity and complexity of the risks LLMs pose, policy responses will need to be multifaceted and incorporate several complementary governance mechanisms. As of now, technology providers and policymakers have only started experimenting with different governance mechanisms, and how LLMs should be governed remains an open question (Engler, 2023).

7.2.3 Calls for audits

Against this backdrop, EBA should be understood as one of several governance mechanisms different stakeholders can employ to assess LLMs for adherence with predefined ethics principles. As I have shown in the previous chapters of this thesis, EBA is not a hypothetical idea but a tangible policy option that has been proposed by researchers and policymakers alike.

For instance, when coining the term foundation models, Bommasani et al. (2021) also suggested that 'such models should be subject to rigorous testing and *auditing* procedures.' Similarly, in an open letter concerning the risks associated with LLMs and other foundation models, OpenAI's CEO Sam Altman stated that 'it's important that efforts like ours submit to *independent audits* before releasing new systems' (Altman, 2023). Finally, the European Commission is considering classifying LLMs and other foundation models as 'high-risk ADMS' (Helberger & Diakopoulos, 2023).[138] This would imply that technology providers designing LLMs have to undergo 'conformity assessments with the involvement of an independent third-party', which – as I argued in Chapter 5 – are audits by another name.

Despite widespread calls for LLM auditing, central questions concerning *how* LLMs can and should be audited have yet to be explored. This chapter addresses that gap by outlining a blueprint for how to audit LLMs. The main argument I advance can be summarised as follows. What auditing means varies between different academic disciplines and industry contexts. However, three strands of auditing research and practice are particularly relevant with respect to ensuring good governance of LLMs. The first stems from IT audits, whereby auditors

[138] It is still uncertain how the EU AIA should be interpreted with respect to LLMs. The current formulation states that models that may be used for high-risk applications should be considered high-risk (Bertuzzi, 2023).

assess the adequacy of technology providers' software development processes and quality management procedures (Senft & Gallegos, 2009). The second strand stems from model testing and verification within the computer sciences, whereby auditors assess the properties of ADMS (Dai & Berleant, 2019). The third strand stems from product certification procedures, whereby auditors test consumer goods for legal compliance and technical robustness before they go to market (Voas & Miller, 2016). As I argue throughout this chapter, it is necessary to combine auditing tools and procedural best practices from each of these three strands to identify and manage the ethical risks LLMs pose.

7.2.4 Addressing initial objections

Before proceeding further, two objections to the prospect of auditing LLMs should be anticipated and responded to. First, one may argue that there is no need to audit LLMs per se and that auditing procedures should be established at the application level instead. However, this objection presents a false dichotomy: while some risks LLMs pose can only be addressed at the application level, others are best managed upstream. Ultimately, technology providers are responsible for taking precautions regarding reasonably foreseeable risks during the product life cycle stages that they do control. The same logic underpins the EU's AI liability directive (European Commission, 2022). For this reason, I propose that application audits should be complemented with governance audits of the organisations that develop LLMs.

Second, one may contend that designing EBA procedures for LLM is difficult. I agree and would add that this difficulty has both practical and conceptual components. Different stages in the LLM development lifecycle overlap in messy ways (Gururangan et al., 2020). For example, open-source LLMs are continuously re-trained and re-uploaded on collaborative platforms post-release. That creates practical problems concerning when and where audits should be mandated. Yet the conceptual challenges run even more deeply. For instance, what constitutes disinformation and hate speech are contested questions (O'Neill, 2021). Despite widespread agreement that LLMs should be 'truthful' and 'fair', such notions are hard to operationalise. Because there exists no universal condition of validity that applies equally to all kinds of utterances (Kasirzadeh & Gabriel, 2023), it is hard to establish a normative baseline against which LLMs can be audited.

However, these difficulties are not reasons for abstaining from developing EBA procedures for LLMs. Instead, they are healthy reminders that it cannot be assumed that one single EBA procedure will capture all LLM-related ethical risks or be equally effective in all contexts (Steed et al., 2022). The insufficiency and limited nature of EBA as a governance

mechanism is not an argument against its complementary usefulness. In this chapter, I attempt to show that EBA can be a feasible and effective governance mechanism for managing the ethical risks LLMs pose. However, before outlining the details of my blueprint for auditing LLMs, something should be said about how it was arrived at.

7.3 Methodology

As discussed in Chapter 1, I take a pragmatist stance when exploring how auditing procedures can be designed so they are feasible and effective in practice. According to the pragmatist tradition, research should not only be grounded in real-world problems but also solution oriented (Salkind, 2010). The research conducted in this chapter is thus *applied* insofar as it both evaluates and proposes policy responses to the real-world problem of how to audit LLMs.

At a high LoA, applied research concerns the evaluation of different governance mechanisms in relation to a desired outcome (Haas & Springer, 1998). A further mark of quality in applied research is that questions are answered in ways that are actionable (Legg & Hookway, 2020). This means that researchers must at times go beyond an evaluation of existing options to prescribe new solutions. While there is no guarantee that the best course of action will be found, researchers can ensure rigour by systematically building on previous research and by incorporating input from different stakeholders.

Mindful of those considerations, the following methodology was used to develop the blueprint for how to audit LLMs outlined in Section 7.5. Note that while the four steps below exhaust the range of research activities that went into this study, the sequential presentation is a simplification. In reality, the research process was iterative, with several of the steps overlapping both thematically and chronologically.

Firstly, I mapped existing EBA procedures designed to identify and mitigate the risks associated with ADMS. In doing so, I used the list of EBA procedures generated through my *systematised literature review* (Grant & Booth, 2009) in Chapter 3 as a starting point. However, I updated it using the same databases and keywords for the search to also include articles and reports that had been published after my original literature review was conducted.[139]

Second, I conducted a conceptual gap analysis between the affordances of existing EBA procedures and the governance challenges LLMs pose. Through this *conceptual analysis* (Maggetti et al., 2015), I identified several properties of LLMs – including generativity and

[139] See Section 3.1 for details regarding the databases and keywords used for the systematised literature review.

emergence – that undermine the feasibility and effectiveness of existing EBA procedures. This resulted in seven key claims about how auditing procedures should be designed to capture the full range of risks posed by LLMs. Those claims are presented and discussed in Section 7.4.

Third, I created a draft blueprint for how to audit LLMs by identifying the smallest set of EBA procedures that satisfied my seven key claims. In practice, not all EBA procedures are equally effective in identifying the risks posed by LLMs. Besides, some EBA procedures serve similar functions. This step thus consisted of reducing the theoretical space of possible EBA procedures into a limited set of activities that are jointly sufficient to identify the risks LLMs pose, practically feasible to implement, and have a justifiable cost-benefit ratio. [140]

Finally, I refined my blueprint by triangulating findings from different sources (Frey, 2018). This step included comparing my draft blueprint with other procedures to audit self-learning software systems, such as NIST's (2023) *AI Risk Management Framework* and ISACA's (2019) *COBIT Framework.* It also included onboarding co-authors with complementary skills and expertise. While I have a dual background in engineering and social science, my co-authors come from the fields of computer science, law, and philosophy respectively. My co-authors suggested improvements to my draft blueprint and added technical depth to the analysis. The three-layered approach for how to audit LLMs outlined in Section 7.5 is thus the result of a broad, iterative, and collaborative effort.

With those methodological remarks out of the way, I now turn to review previous work on EBA. The thereby is to explore the merits and limitations of existing EBA procedures when applied to LLMs.

7.4 Seven claims about auditing LLMs

As I demonstrated in Chapter 3 of this thesis, a wide range of EBA procedures have already been developed. However, as we will see, not all EBA procedures are equally effective in identifying and mitigating the risks posed by LLMs. In this section, I introduce several conceptual distinctions to highlight the properties of LLMs that undermine the feasibility and effectiveness of existing EBA procedures. The result of this conceptual analysis is a list of seven claims about how to audit LLMs.

[140] The first three steps discussed in this section corresponds to the methodology for applied research outlined in Section 1.7.2.

7.4.1 The merits and limitations of existing EBA procedures

To start with, it is useful to distinguish between *compliance audits* and *risk audits*. The former compares an entity's actions or properties to predefined standards. The latter asks open-ended questions about how a system works to identify and manage risks. When conducting risk audits of LLMs, auditors can draw on well-established standards for ADMS risk management (ISO, 2023; NIST, 2022) and guidance on how to assess and evaluate ADMS (ICO, 2020; VDE, 2022). In contrast, compliance audits require a normative baseline against which ADMS can be evaluated. However, LLM research is a quickly developing field in which standards have yet to emerge. Moreover, the fact that LLMs are adaptable to many downstream applications (Ganguli et al., 2022a) undermines the feasibility of EBA procedures designed to ensure compliance with sector-specific norms. This implies that EBA procedures focusing on compliance alone are unlikely to provide adequate assurance for LLMs (Claim 1).

Further, it is useful to distinguish between *external* and *internal audits*. While the former is conducted by independent third-parties, the latter is conducted by internal auditors that report directly to an organisation's board (IIA, 2022). External audits address concerns regarding accuracy in self-reporting. However, they are constrained by limited access to processes and personnel (Raji et al., 2020). For internal audits, the inverse is true: while having all necessary access, they run an increased risk of collusion between the auditor and the auditee. Without third-party accountability, technology providers may also ignore audit recommendations that threaten their business interests (Slee, 2020). The risks stemming from misaligned incentives are especially stark for technologies with rapidly increasing capabilities and for companies facing strong competitive pressures (Naudé & Dimitri, 2020). Both conditions apply to LLMs. From this follows that external audits will be required to not only identify the ethical risks LLMs pose but also to hold technology providers accountable in case of irregularities (Claim 2).

A further distinction made in the EBA literature is that between *adversarial* and *collaborative* audits. Adversarial audits are conducted by independent actors to assess the properties or impact ADMS have – without privileged access to their source code or technical design specifications (Blocki et al., 2013). Collaborative audits see technology providers and external auditors working together to assess and improve the processes that shape future ADMS design and safeguards (Berghout et al., 2023). While the former aims to expose harms, the latter seeks to provide assurance. Previous research has shown that

audits are most effective when technology providers and independent auditors collaborate towards the common goal of identifying and managing risks (Power, 1997). This implies that, to be feasible and effective in practice, procedures to audit LLM require active collaboration between technology providers and independent auditors (Claim 3).

Moving on, it is also useful to distinguish between *governance audits* and *technology audits*. The former focuses on the organisation designing ADMS and include assessments of software development and quality management processes, incentive structures, and the allocation of roles and responsibilities (Senft & Gallegos, 2009). The latter focuses on assessing ADMS' properties, e.g., by reviewing the model architecture, checking its consistency with predefined specifications, or repeatedly querying it to understand its inner workings (Metaxa et al., 2021). Some LLM-related risks can be identified and mitigated at the application level. However, others are best addressed upstream, e.g., those concerning the sourcing of training data. This implies that, to be feasible and effective, EBA procedures designed to assess and mitigate the risks posed by LLMs must include elements of both governance and technology audits (Claim 4).

However, both governance audits and technology audits have limitations. Governance audits are limited because it is not possible to anticipate upfront all the risks that emerge as ADMS interact with complex environments over time (Steed et al., 2022). Further, not all ethical tensions stem from technology design alone, as some are intrinsic to given tasks (Danks & London, 2017). While these limitations were discussed already in chapter 3, LLMs introduce additional challenges for technology audits, which have historically focused on assessing ADMS designed to fill specific functions in well-defined contexts, like improving image analysis in radiology (Mahajan et al., 2020) or detecting corporate fraud (Zerbino et al., 2018). Because LLMs enable many downstream applications, such traditional EBA procedures are not equipped to capture the full range ethical risks they pose. While existing best practices in governance auditing appear applicable to organisations designing or deploying LLMs, that is not true for technology audits. In short: the methodological design of technology audits will require significant modifications to identify and assess LLM-related risks (Claim 5).

Previous work on technology audits distinguishes between *functionality*, *model*, and *impact audits* (Mittelstadt, 2016). Functionality audits focus on the rationale underpinning ADMS by asking questions about intentionality, e.g., what is this system's purpose (Kroll, 2018)? Model audits review ADMS internal decision-making logic. For sub-symbolic systems like LLMs, this entails asking how the model was designed, what

data it was trained on, and how it performs on different benchmarks. Finally, impact audits investigate the types, severity, and prevalence of effects an ADMS has on individuals, groups, and the environment (OECD, 2022). These approaches are not mutually exclusive but rather complementary. Still, technology providers that design and disseminate LLMs have limited information about the future deployment of their systems by downstream developers and end-users. This implies that model audits will play a key role in identifying and communicating LLMs' limitations, thereby mitigating downstream harms (Claim 6).

Finally, within technology audits, it is important to distinguish between *ex-ante* and *ex-post audits*, which take place before and after a system is deployed, respectively. Considerable literature already exists on ex-ante auditing techniques such as red teaming (Ganguli et al., 2022b), model fooling (Xu et al., 2018), and functional testing (Röttger et al., 2021). While useful, ex-ante audits cannot capture all the risks associated with systems that continue to 'learn' by updating their internal decision-making logic (Dignum, 2017). This limitation applies to all ADMS but is particularly relevant for LLMs that display emergent capabilities (Wei et al., 2022).[141] This means that ex-posts audits are needed too. Ex-post audits can be further divided into *snapshot audits* (which occur regularly) and *continuous audits* (which monitor performance over time). Most existing EBA procedures are snapshots.[142] Like ex-ante audits, however, snapshots are unable to provide meaningful assurance regarding LLMs as they display emergent capabilities and, in some cases, can learn as they are fed new data. As a result, LLM auditing procedures must include elements of continuous ex-post monitoring to meet their regulatory objectives (Claim 7).

7.4.2 Summary of claims

Much can be learned from existing EBA procedures. However, LLMs display several properties that undermine the feasibility and effectiveness of such procedures. Specifically, LLMs are adaptable to a wide range of downstream applications, display emergent capabilities, and can, in some cases, continue to learn over time. This means that neither functionality audits (which hinge on the evaluation of the purpose of a specific application) nor impact audits (which hinge on the ability to observe a specific system's

[141] Emergence implies that an entity can have properties its parts do not individually possess, and that randomness can generate orderly structures (Corning, 2010).

[142] The post-market monitoring mandated by the proposed EU AIA (European Commission, 2021a) is a rare example of continuous auditing.

actual impact) alone can provide adequate assurance against the social and ethical risks LLMs pose. It also means that ex-ante audits must be complemented by continuous post-market monitoring of outputs from LLM-based applications.

In this section, I have built on these and other insights to derive and defend seven claims about how auditing procedures should be designed to account for the governance challenges LLMs pose. To summarise:

- **Claim 1**: EBA procedures focusing on compliance alone are unlikely to provide adequate assurance for LLMs.
- **Claim 2:** External audits are required – both to identify the ethical risks LLMs pose and to hold technology providers accountable in case of irregularities of incidents.
- **Claim 3.** To be feasible and effective in practice, procedures to audit LLM require active collaboration between technology providers and independent auditors.
- **Claim 4.** EBA procedures designed to assess and mitigate the risks posed by LLMs must include elements of both governance and technology audits.
- **Claim 5.** The methodological design of technology audits will require significant modifications to identify and assess LLM-related risks.
- **Claim 6.** Model audits that identify and communicate LLMs' limitations will play a key role in informing their redesign and in mitigating downstream harms.
- **Claim 7.** LLM auditing procedures must include elements of continuous ex-post monitoring to meet their regulatory objectives.

These seven claims provided my starting point when designing the three-layered approach for auditing LLMs that will be outlined in the next section. However, I maintain that these claims are more general and could serve as guardrails for the design of procedures to audit other ADMS with highly general capabilities too.

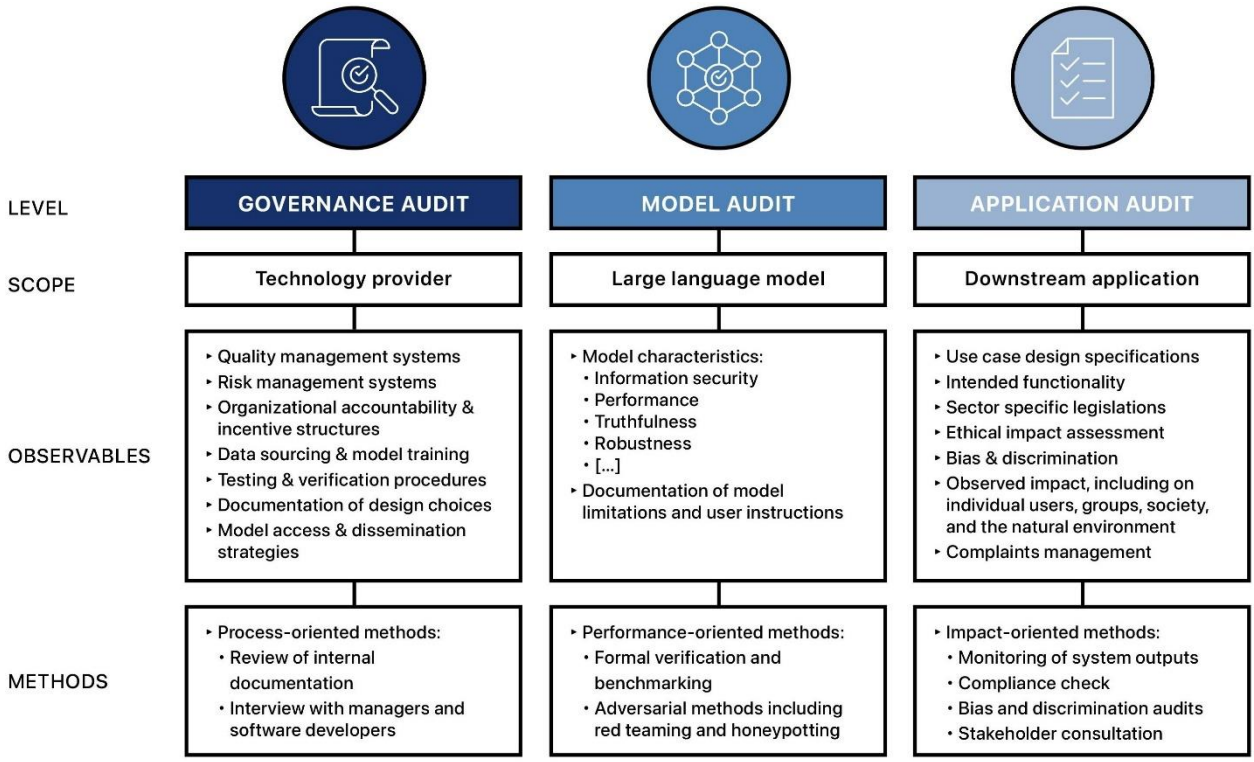
7.5 Auditing LLMs: A three-layered approach

This section outlines a blueprint for how to audit LLMs that satisfies the seven claims listed in Section 7.4. While there are many ways to design EBA procedures that satisfy those claims, the three-layered approached proposed in this section focuses on a limited set of activities that are jointly sufficient to identify LLM-related risks, practically feasible to implement, and have a justifiable cost-benefit ratio.

7.5.1 A blueprint for LLM auditing

Audits should focus on three levels. First, technology providers designing LLMs should undergo *governance audits* that assess their software development processes, accountability structures and QMS. Second, LLMs should undergo *model audits* that assess their capabilities and limitations after initial training but before deployment for specific applications. Third, downstream applications using LLMs should undergo continuous *application audits* that assess the ethical alignment of their intended functions and observable impact over time. Figure 15 illustrates the logic of this three-layered approach.

Figure 15. Blueprint for how to audit LLMs: A three-layered approach.



Four clarifications are needed to flesh out this blueprint. First, governance, model and application audits only provide effective assurance when coordinated. This is because the audits conducted at the three levels have complementary affordances. For example, LLM audits must include of both process- and performance-oriented auditing (Claim 4). In my three-layered approach, governance audits are process-oriented, whereas model and application audits are performance-oriented. Moreover, feasible and effective LLM auditing procedures must include elements of continuous ex-post assessments (Claim 7). These elements are incorporated at the application level of my three-layered approach. But these are just two

examples. As I discuss what governance, model, and applications audits entail, I also highlight how they, when combined, satisfies all seven claims listed in Section 7.4.

Second, while the governance, model, and application audits are all individually necessary, their boundaries overlap and can be drawn in multiple ways. For example, the collection and pre-processing of training data ties into software development practices. Hence, reviewing procedures for obtaining and curating training data is part of governance audits. However, the characteristics LLMs display may reflect biases in their training data (Gehman et al., 2020). Reviewing such data is, therefore, necessary during model audits too (Song & Shmatikov, 2019). Nevertheless, the conceptual distinction between governance, model and application audits remains useful when identifying varied risks that LLMs pose.

Third, it is theoretically possible to add further layers to procedures for auditing LLMs. For example, downstream developers could also be made subject to governance audits. But such audits would be difficult to implement, given that many decentralised actors build applications on top of LLMs. The combination of governance, model, and application audits, I argue, strikes a balance between covering a sufficiently large part of the development and deployment lifecycle to identify LLM-related risks, on the one hand, and being practically feasible to implement, on the other.

Finally, audits on all levels should be external (Claim 2) yet collaborative (Claim 3). This implies that independent third-parties not only seek to verify claims made by technology providers but also work together with them to identify and mitigate risks LLMs pose. As mentioned in the introduction, the question of who should conduct the audits falls outside the scope of this chapter. That said, reasonable concerns about how independent collaborative audits really are can be raised regardless of who is conducting the audit. In Section 7.6, I discuss this and other limitations.

With those clarifications in mind, I will now present the details of my three-layered approach. The following subsections discuss governance, model, and application audits, focusing on why each is needed, what each entails, and what outputs each should produce.

7.5.2 Governance audits

Technology providers designing LLMs should undergo governance audits that assess their organisational procedures, incentive structures, and management systems. Evidence shows that such features influence the design and deployment of technologies (Brundage et al., 2020). Moreover, previous research suggests that risk-mitigation strategies work best when adopted consistently and with executive-level support (Floridi & Strait, 2020). Technology providers

are responsible for identifying the risks associated with their LLMs and are uniquely well-positioned to manage some of those risks. It is thus crucial that their organisational procedures and governance structures are adequate.

Governance audits have a long history in areas like IT governance (Senft & Gallegos, 2009) and safety engineering (Dobbe, 2022; Leveson, 2011). Tasks include assessing internal governance structures, product development processes, and QMS to promote transparency and procedural regularity, ensure that appropriate risk management systems are in place, and spark ethical deliberation throughout the software development lifecycle (Berghout et al., 2023). Governance audits can also improve accountability, since publicising their results prevents companies from covering up undesirable outcomes and incentivises better behaviour (Engler, 2021). Governance audits thus incorporate elements of both compliance audits, regarding completeness and transparency of documentation, and risk audits, regarding the adequacy of the risk management system (Claim 1).

Specifically, governance audits of technology providers should focus on three tasks:[143]

1) *Reviewing the adequacy of organisational governance structures* to ensure that software development processes follow best practices and QMS capture LLM-specific risks. While technology providers have in-house quality management experts, confirmation bias may prevent them from recognising critical flaws. Involving external auditors addresses that issue (Bauer, 2016). Nevertheless, governance audits are most effective when auditors and technology providers collaborate to identify risks (Chopra & Singh, 2018). Therefore, it is important to distinguish accountability from blame at this stage.

2) *Creating an audit trail of the LLM development process* to provide documentary evidence of the development of an LLM's capabilities, including information about its intended purpose, design specifications, as well as how it was trained and tested. This includes the structured use of model cards (Mitchell et al., 2019), system cards (MetaAI, 2023) and datasheets (Gebru et al., 2021) to document how the datasets used to train and validate LLMs were sources, labelled, and curated. The creation of such audit trails serves several related purposes. Stipulating design specifications upfront facilitates checking system adherence to jurisdictional requirements downstream (Falco et al., 2021). Moreover, information concerning intended use cases should inform licensing agreements with

[143] Governance audits could examine many tasks, and prioritization may vary depending on the sector and jurisdiction. Hence, the three tasks I propose are merely a minimum baseline.

downstream developers (Contractor et al., 2022), thereby restricting the potential for harm through malicious use. Finally, requiring providers to document and justify their design choices sparks ethical deliberation by making trade-offs explicit.

3) *Mapping roles and responsibilities within organisations that design LLMs* to facilitate the allocation of accountability for system failures. LLMs' adaptability downstream does not exculpate technology providers from all responsibility. Some risks are 'reasonably foreseeable.' In the adjacent field of ML image recognition, a study found that gender classification systems were less accurate for darker-skinned females than lighter-skin males (Buolamwini & Gebru, 2018). After the release of these findings, technology providers speedily improved the accuracy of their ADMS, suggesting that the problem was not intrinsic, but resulted from inadequate risk management. Mapping the roles and responsibilities of different stakeholders improves accountability and increases the likelihood of impact assessments being structured rather than ad-hoc, thus helping identify and mitigate harms proactively.

The results of governance audits should be tailored to different audiences. The primary audience is the management of the LLM provider. Auditors should provide a full report that transparently lists and discusses the vulnerabilities of existing governance structures. Such reports may recommend actions, but taking actions remains the provider's responsibility. Usually, such audit reports are not made public. However, some evidence obtained during governance audits can be curated for two secondary audiences: law enforcers and developers of downstream applications. In some jurisdictions, hard legislation may demand that technology providers follow specific requirements. For instance, as described in Chapter 5, the AIA required providers to register high-risk ADMS with a centralised database (European Commission, 2021a). In such cases, reports from governance audits can help providers demonstrate adherence to legislation. Reports from governance audits also assist developers of downstream applications to understand an LLM's intended purpose, risks, and limitations.

To conclude this discussion about governance audits, it is useful to reflect about how they contribute to relieving some of the ethical risks LLMs pose. As mentioned in Section 7.2, Weidinger et al. (2021) listed six broad risk areas: discrimination, information hazards, misinformation hazards, malicious use, human-computer interaction harm, and automation and environmental harms. Governance audits address some of these directly. By assessing the adequacy of the governance structures surrounding LLMs, including licencing agreements (Contractor et al., 2022), governance audits help reduce the risk of malicious use. Further, some

information hazards stem from the possibility of extracting sensitive information from LLMs (Carlini et al., 2021). By reviewing the process whereby training datasets were sourced, labelled, and curated, as well as the strategies and techniques used during the model training process – such as differential privacy (Dwork, 2006) or secure federated learning (Kaissis et al., 2020) – governance audits can minimise the risk of LLMs leaking sensitive information. However, for most of the risks posed by LLMs, governance audits have only an indirect impact. Risks areas like discrimination, misinformation hazards, and human-computer interaction harms are better addressed by model and application audits.

7.5.3 Model audits

Before deployment, LLMs should be subject to model audits that assess their capabilities and limitations (Claim 6). Model audits share some features with governance audits. For instance, both take place before an LLM is deployed. However, model audits do not focus on technology providers' organisational procedures but on LLMs' characteristics. Specifically, model audits seek to assess LLMs' capabilities and limitations and communicate these to relevant stakeholders. These two tasks use similar methodologies, but they target different audiences.

The first task – identifying capabilities and limitations – aims to support organisations developing LLMs with benchmarks that inform internal model retraining efforts (Bharadwaj et al., 2021). The second task – communicating capabilities and limitations – aims to inform the design of applications built on top of LLMs. Such communication can take different forms, e.g., *interactive model cards* (Crisan et al., 2022) and *information about the initial training dataset* (Jernite et al., 2022), to help downstream developers adapt the model appropriately.

In Section 7.4, I argued that the way technology audits are being conducted requires modifications to address the governance challenges LLMs pose (Claim 5). Evaluating an LLM's characteristics independent of an intended use case is challenging but not impossible.[144] Auditors can use two distinct approaches. The first involves identifying and assessing intrinsic characteristics. For example, the training dataset can be assessed for completeness and consistency without reference to specific use cases. However, it is often expensive and technically challenging to interrogate large datasets (Paullada et al., 2021). The second approach is indirect and involves testing an LLMs performance across multiple downstream

[144] A wide range of tools and methods to evaluate LLMs already exists. For an overview, see Liang et al. (2022).

use cases, linking the test results to different model characteristics, and assessing the aggregated results using different weighting techniques.

However, selecting the characteristics to focus on during model audits remains challenging. Given such audits' purpose, I recommend examining model characteristics that are: (i) *socially and ethically relevant*, i.e., can be linked to the ethical risks posed by LLMs; (ii) *predictably transferable*, i.e., impact the observable properties of downstream applications; and (iii) *operationalisable*, i.e., can be assessed with the available tools and methods.

Keeping those criteria in mind, I posit that model audits should focus on the performance, robustness, information security and truthfulness of LLMs. Other characteristics may also meet the criteria listed above. Hence, the four characteristics discussed in this section are just examples to illustrate the role of model audits. The list of relevant model characteristics can be amended when developing specific EBA procedures. With those caveats out of the way, I now proceed to discuss how different characteristics can be assessed during model audits:

1) *Performance*, i.e., how well the LLM functions on various tasks. Standardised benchmarks can help assess an LLM's performance by comparing it to a human baseline. For example, *GLUE* (Wang et al., 2018) aggregates LLM performance across multiple tasks into a single reportable metric. Such benchmarks have been criticised for overestimating performance over a narrow set of capabilities. Therefore, it is crucial to evaluate LLMs' performance against many benchmarks. Sophisticated tools and methods have already been proposed for that purpose, including *SuperGLUE* (Wang et al., 2019), which is more challenging and 'harder to game' with narrow LLM capabilities, and *BIG-bench* (Srivastava et al., 2022), which assess LLM's performance on tasks that appear beyond their current capabilities. These benchmarks are particularly relevant for model audits because they were developed to evaluate pre-trained models without task-specific fine-tuning.

2) *Robustness*, i.e., how well the model reacts to unexpected prompts or edge cases. In ML, robustness indicates how well an algorithm performs when faced with new, potentially unexpected, input data. LLMs lacking robustness introduce two distinct risks. First, the risk of critical system failures if, for example, an LLM performs poorly for individuals unlike those represented in the training data (Sohoni et al., 2020). Second, the risk of adversarial attacks (Garg & Ramakrishnan, 2020). Therefore, researchers and developers have created tools and methods to assess LLMs' robustness, including adversarial methods like red teaming (Ganguli et al., 2022), evaluation toolkits like the *Robustness Gym* (Goel et al., 2021), benchmark datasets like *ANLI* (Nie et al., 2020), and open-source platforms for

testing like *Dynabench* (Kiela et al., 2021). Particularly relevant for assessing robustness during model audits is *AdvGLUE* (Wang et al., 2021a), which evaluates LLMs' vulnerabilities to adversarial attacks in different domains using a multi-task benchmark. By quantifying robustness, AdvGLUE facilitates comparisons between LLMs. However, robustness can be operationalised in different ways, e.g., group robustness, which measures a model's performance across different sub-populations (Zhang & Ré, 2022). Therefore, model audits should employ multiple tools and methods to assess robustness.

3) *Information security*, i.e., how difficult it is to extract training data from the LLM. Several LLM-related risks can be understood as 'information hazards', including the risk of compromising privacy by leaking personal data (Weidinger et al., 2021). Adversarial agents can perform *training data extraction attacks* to recover personal information like social security numbers (Carlini et al., 2021). However, not all LLMs are equally vulnerable to such attacks. The memorisation of training data can be minimised through differentially private training techniques (McMahan et al., 2018), but their application generally reduces accuracy and increases training time (Jayaraman & Evans, 2019). Promisingly, it is possible to assess the extent to which an LLM has unintentionally memorised unique training data using metrics such as *exposure* (Carlini et al., 2019). Testing strategies, like exposure, can be employed at the model level, although that requires auditors to have access to the LLM and its training corpus. Still, assessing LLMs' information security during model audits does not address all information hazards because some risk of correctly inferring sensitive information about users can only be audited on an application level.

4) *Truthfulness*, i.e., to what extent the LLM can distinguish between the real world and possible worlds. Some LLM-related risks stem from their tendency to provide false or misleading information (Weidinger et al., 2021). This is because statistical methods struggle to distinguish between factually correct versus plausible but factually incorrect information. That problem is exacerbated by the fact that many LLM training practices, like imitating human text on the web or optimising for clicks, are unlikely to create truthful systems (Evans et al., 2021). Model audits should therefore assess LLMs for truthfulness. Such audits should focus on evaluating overall truthfulness, not the truthfulness of an individual statement. However, that does not preclude focusing on multiple aspects, e.g., how frequent falsehoods are on average, and how bad worst-case falsehoods are. One benchmark that measures truthfulness is *TruthfulQA* (Lin et al., 2022), which generates a percentage score using 817 questions spanning 38 application domains, including medicine

and physics. When evaluating an LLM with the help of TruthfulQA, auditors would get a percentage score on how truthful the model is. However, even a strong performance on TruthfulQA does not imply that an LLM will be truthful in a specialised domain. Nevertheless, such benchmarks offer helpful tools for model audits.

These four characteristics pertain to pre-trained LLMs. However, model audits should also review training datasets. It is well-known that training data gaps or biases create models that perform poorly on different datasets (Nejadgholi & Kiritchenko, 2020). Training LLMs with biased or incomplete data can cause representational and allocational harms (Caliskan et al., 2017). A recent European Parliament report (EPRS, 2022) thus discussed mandating third-party audits of datasets. Technology providers should prepare for such proposals potentially becoming legal requirements by including audits of datasets as part of model audits.

Despite these technical and legal considerations, training datasets are often collected with little curation, supervision, or foresight (Jo & Gebru, 2020). While curating 'unbiased' datasets may be impossible, disclosing how a dataset was assembled can suggest its potential biases (Dodge et al., 2021). Model auditors can use existing tools that interrogate biases in LLMs' pre-trained word embeddings, such as the metrics *DisCo* (Webster et al., 2020), *SEAT* (May et al., 2019) or *CAT* (Nadeem et al., 2021). So-called *data statements* (Bender & Friedman, 2018) also provide developers and users with the context required to understand a specific LLM's potential biases. The availability of such tools is encouraging. Yet it is important to remain realistic about what model audits can achieve. Model audits do not ensure that LLMs are ethical in any global sense. Instead, they contribute to better precision in claims about an LLM's capabilities and inform the design of downstream applications.

Some of the characteristics tested for during model audits correspond directly to the ethical risks LLMs pose. For example, model audits entail evaluating LLMs according to characteristics like information security and truthfulness, which correspond to information hazards and misinformation hazards, respectively, in Weidinger et al.'s taxonomy. However, model audits – as outlined in this section – only focus on a few characteristics of LLMs. That is because the criterion of *operationalisability* sets a high bar: not all risks associated with LLMs can be addressed at the model level. Consider discrimination as an example. Model audits can expose the root causes of some discriminatory practices, such as biases in training datasets that reflect historic injustices. However, what constitutes unjust discrimination is context-dependent and varies between different applications and jurisdictions. That problematises saying anything about risks like unjust discrimination on a model level (Hancox-

Li & Kumar, 2021). However, that observation does not argue against model audits but for complementary approaches like application audits, as discussed next.

7.5.4 Application audits

Products and services built using LLMs should undergo application audits that assess their intended functions and observable impact. Unlike governance and model audits, application audits focus on actors employing LLMs for specific downstream applications. Such audits are well-suited to ensure compliance with not only hard legislation but also with sector-specific standards and organisational ethics principles. Application audits have two components: *functionality audits*, which evaluate applications using LLMs based on their intended goals, and *impact audits*, which evaluate applications based on their impacts on individual, groups, and the natural environment.

During *functionality audits*, auditors should check whether the intended purpose of a specific application is (i) ethical in and of itself and (ii) aligned with the intended use of the LLM on which it is built. This first check is for compliance with not only applicable regulations but also organisational ethics principles and sector specific codes of conduct. The purpose of such a check is straightforward: if an application is unlawful or unethical, the performance of its LLM component is irrelevant and the application should not be put on the market.

The second check within functionality audits aims to address the risks stemming from developers overstating or misrepresenting a specific application's capabilities (Raji et al., 2022). In doing so, functionality audits account for output from audits on other levels. During governance audits, technology providers are obliged to define intended and disallowed use cases of LLMs. During model audits, LLMs' limitations are documented. Using such information, functionality audits ensure that downstream applications are aligned with a given LLM's intended use cases and known limitations. Functionality audits thus combines the elements of compliance and risks audit needed to provide assurance for LLMs (Claim 1).

During *impact audits*, auditors disregard both the purpose and design of an application and focus only on how it impacts individuals, groups, and society. The idea behind impact audits is simple: every system can be understood in terms of its inputs and outputs (Kroll, 2018). However, implementing impact audits is notoriously hard. To begin with, ADMS and their environments co-evolve in non-linear ways (Lauer, 2020). The link between an ADMS intended purpose and its impact may be neither intuitive nor consistent over time. Moreover, it is difficult to track impacts stemming from indirect causal chains (Dafoe, 2017). Establishing which direct and indirect impacts are considered ethically relevant thus remains a context-

dependent question which must be resolved on a case-by-case basis. The application must be redesigned or terminated if the impact is considered unacceptable.

Impact audits should include both pre-deployment (ex-ante) assessments and post-deployment (ex-post) monitoring (Claim 7).[145] The former leverages either empirical evidence or plausible scenarios, depending on how well-defined the application is and how predictable the environments in which it will operate is. For example, applications can be tested in *sandbox environments* (Truby et al., 2022) that mimic real-world environments and allow developers and policymakers to understand their potential impact prior to market deployment.[146] However, real-world environments often differ from training and testing environments in unforeseen ways (Zinda, 2021). Hence, pre-deployment assessments of LLM-based applications must also use analytical strategies to anticipate the application's impact, e.g., *ethical impact assessments* (Mantelero, 2018; Selbst, 2021) and *ethical foresight analysis* (Floridi & Strait, 2020).

While pre-deployment impact assessments are necessary, they are not sufficient. Capturing the full range of potential harms from LLM-based applications also requires EBA procedures to include elements of continuous oversight (again, see Claim 7). Ex-post auditing can be done in different ways, e.g., periodically reviewing the output from an application and comparing it to relevant standards. Such procedures can also be automated, e.g., by using oversight programs that continuously monitor and evaluate system outputs and alert or intervene if they transgress predefined tolerance spans (Etzioni & Etzioni, 2016).

Taken together, application audits seek to ensure that ex-ante testing and impact assessments have been conducted following existing best practices; that post-market plans have been established to enable continuous monitoring of system outputs; and that procedures are in place to mitigate different types of failure modes. By focusing on individual use cases, application audits are well-suited to alert stakeholders to risks that require much contextual information to be addressed. This includes risks related to discrimination and human-computer interaction harms in Weidinger et al.'s taxonomy. Application audits help identify such risks in several ways. For example, quantitative assessments linking inputs and outputs give a sense of what kinds of language an LLM is propagating (Karan & Šnajder, 2019). Moreover, qualitative assessments (e.g., those based on ethnographic methods) provide insights into users' lived experiences of interacting with an LLM (Marda & Narayan, 2021).

---

[145] This structure mirrors the 'conformity assessments' and 'post-market monitoring plans' proposed in the AIA.
[146] Sandboxes have proven safe harbours in which to biases in ADMS can be detected (Akpinar et al., 2022).

The above examples show that holistic EBA procedures must incorporate elements of application audits. This is because many ethical risks are difficult to define in any global sense (Delobelle et al., 2022). For example, several studies have documented situations in which LLMs propagate toxic language (Nozza et al., 2021), but the interpretation of toxicity and the materialisation of its harms vary across cultural, social, or political groups (Costello et al., 2019). Sometimes, 'detoxifying' an LLM may be incompatible with other goals and potentially suppress texts written about or by marginalised groups (Welbl et al., 2021). Moreover, certain expressions might be acceptable in one setting but not in another. In such circumstances, the most promising way forward is to audit not LLMs themselves but downstream applications, thus ensuring that each application's outputs adhere to contextually appropriate conversational conventions (Kasirzadeh & Gabriel, 2023).

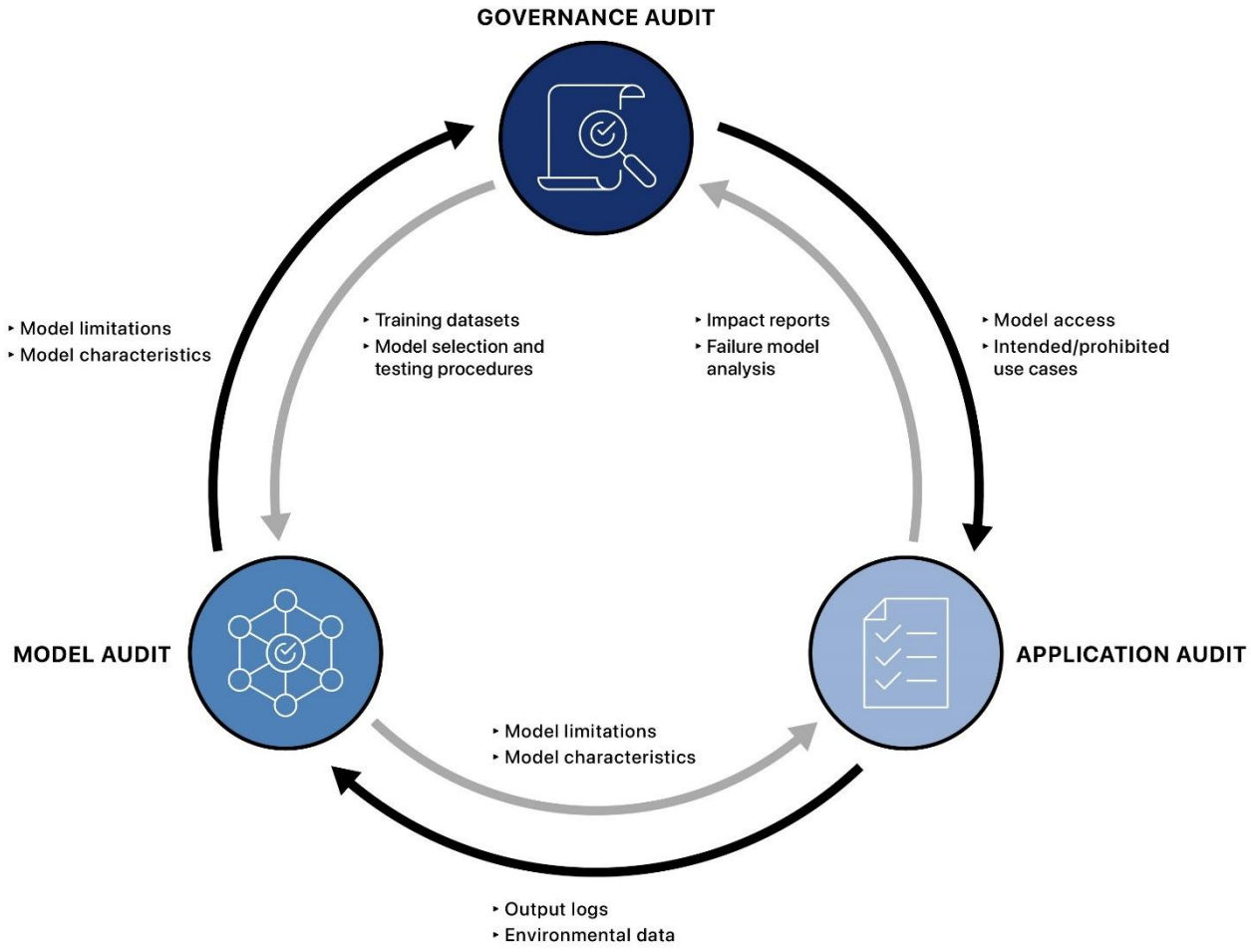
Two further observations help conclude this discussion about application audits. First, while governance audits and model audits should be obligatory for all technology providers designing and disseminating LLMs, application audits can be employed more selectively. One way is to only subject LLMs used for specific high-risk tasks to application audits. Second, although application audits may form the basis for certification (Cihon et al., 2021), auditing does not equal certification. Certification requires predefined standards against which a product or service can be audited and institutional arrangements to ensure the certification process's integrity (Yanisky-Ravid & Hallisey, 2019). However, the results from application audits should be published (at least in summary form) even when not related to certification. This is because publishing audit results incentivise companies to correct behaviour, inform enforcement actions, and help cure informational asymmetries in technology regulation.

7.5.5 Connecting the dots

To identify and mitigate the ethical risks LLMs pose, governance, model, and application audits must be connected into a structured process. In practice, this means that outputs from audits on one level become inputs for audits on other levels. Model audits produce reports summarising LLMs' properties and limitations, which should inform application audits that verify whether a model's known limitations have been considered in the design of downstream applications. Similarly, application audits produce output logs documenting the impact that different applications have in applied settings. Such logs should inform LLMs' continuous redesign. Finally, governance audits must check the extent to which technology providers' software development processes and QMS include mechanisms to incorporate feedback from

application audits. Figure 16 illustrates how governance, model, and application audits are interconnected in the three-layered approach for auditing LLMs.

Figure 16. Outputs from audits on one level become inputs for audits on other levels.



Each step in the three-layered approach should be conducted by independent third-party auditors (Claim 2). However, three caveats are required here. First, it need not be the same organisation that conducts audits on all three levels. Conducting different types of audits require different competencies: governance audits require an understanding of corporate governance and soft skills like stakeholder communication; model audits are highly technical and require knowledge about evaluating ML models; and application auditors typically need domain-specific expertise. All these competencies may not be found within one organisation.

Second, the level of access required to conduct audits vary between the three levels. At the governance level, auditors require the highest level of access, including to information about how and why an LLM was developed. This corresponds to what Koshiyama et al. 2022 refer to as ‘white-box auditing’, and it implies privileged access to facilities, documentation, and personnel. White-box auditing requires that nondisclosure and data-sharing agreements are

in place, which adds to the logistical burden of governance audits. At the model level, auditors also require privileged access, but only to LLMs and their training datasets. In Koshiyama et al.'s typology, this corresponds to 'grey box auditing', whereby auditors can manipulate model parameters and task objectives when assessing LLMs. Finally, 'black box auditing' is sufficient at the application level. This implies that auditors can rely on publicly available information and quantitative assessments of input-output relationships to verify technology providers' claims about their LLM-based applications.

Third, since institutional arrangements vary between jurisdictions and sectors, the best option may be to leverage the capabilities of institutions operating within a specific geography or industry to perform various elements of governance, model, and application audits. For example, medical devices are already subject to testing and certification procedures before being launched. Hence, application audits for new medical devices incorporating LLMs could be integrated with such procedures. In part, this is already happening. The FDA (2021) has proposed a regulatory framework for modifying ML-based software as a medical device. The point is that different independent auditors can perform the three different types of audits and that different institutional arrangements may be preferable in different jurisdictions or sectors.

7.6 Limitations and avenues for further research

This section highlights three limitations to the blueprint for how to audit LLMs outlined in this chapter: one conceptual, one institutional and one practical. First, model audits pose conceptual problems related to construct validity. Second, an institutional ecosystem to support independent third-party audits has yet to emerge. Third, not all LLM-related social and ethical risks can be practically addressed on the technology level. I consider these limitations in turn, discuss potential solutions, and provide directions for future research.

7.6.1 Lack of methods and metrics to operationalise normative concepts

One bottleneck to developing effective EBA procedures is the difficulty of operationalising normative concepts (Jacobs & Wallach, 2021). The problem is rooted in construct validity, i.e., the extent to which a given metric accurately measures what it is supposed to (Smith, 2014). In the blueprint for how to audit LLM outlined in this chapter, construct validity problems primarily arise from attempts to operationalise characteristics like performance, robustness, information security and truthfulness during model audits.

Consider truthfulness as an example. LLMs do not require a model of the real world. Instead, they compress vast numbers of conditional probabilities by picking up on language

regularities (Sobieszek & Price, 2022). Therefore, LLM have no reason to favour any reality but can select from various possible worlds, provided each is internally coherent (Reynolds et al., 2021).[147] However, different epistemological positions disagree about the extent to which this way of sensemaking is unique to LLMs. Simplifying to the extreme, *realists* believe in objectivity and the singularity of truth, at least insofar as the natural world is concerned (Hacking, 1983). In contrast, *relativists* believe that truth and falsity are products of context-dependent conventions and assessment frameworks (Rorty, 2021). Numerous compromise positions can be found on this spectrum. However, tackling pressing social issues cannot await the resolution of long-standing philosophical disagreements. Indeed, courts settle disagreements daily based on pragmatist operationalisations of concepts like truth and falsehood in keeping with the pragmatic maxim that theories should be judged by their success when applied practically to real-world situations (Legg & Hookway, 2020).

That said, developing metrics to capture the essence of normative concepts is difficult and entails many well-known pitfalls. Reductionist representations of normative concepts generally bear little resemblance to real-life considerations, which tend to be highly contextual (Lee et al., 2022). Moreover, different operationalisations of the same normative concept (like 'fairness') cannot be satisfied simultaneously (Friedler et al., 2021). Finally, the quantification of normative concepts can itself have undesired consequences (Islam & Greenwood, 2021). As Goodhart's Law reminds us, a measure ceases to be a good metric once it becomes a target.

The operationalisation of characteristics like performance, robustness, information security and truthfulness discussed in Section 7.5 is subject to the above limitations. Resolving all construct validity problems may be impossible, but some ways of operationalising normative concepts are better than others for the purpose of auditing LLM. Consequently, an important avenue for further research is developing new methods to operationalise normative concepts in ways that are verifiable and maintain high construct validity.

7.6.2 Lack of an institutional ecosystem

A further limitation of my blueprint for how to audit LLMs is that it does not decisively identify who should conduct the three audits it recommends. This is a limitation, since any auditing procedure will only be as good as the institution delivering it (Boddington, 2017). However, I have left the question open for two reasons. First, different institutional ecosystems intended

[147] LLMs favour the statistically most likely reality given their training data. However, any training data constitute a reduction of reality that supports some interpretations but obscures others (Cantwell Smith, 2019).

to support audits and conformity assessments of ADMS are currently emerging in different jurisdictions and sectors (Minkkinen et al., 2021). Second, the three-layered approach is flexible enough to be adopted by any external auditor. Hence, the feasibility and effectiveness of the blueprint outlined in this chapter do not hinge on the question of institutional design.

That said, the question of who audits whom is important, and the lack of an institutional ecosystem remains a limitation for any attempt to audit LLMs. Without clear institutional arrangements, claims that an LLM-based application has been audited are difficult to verify. Further research could usefully investigate the feasibility and effectiveness of different institutional arrangements for conducting and enforcing the three types of audits proposed.

7.6.3 Not all risks from LLMs can be addressed on the technology level

The three-layered approach outlined in this chapter has been designed to contribute to good governance. However, it cannot fully eliminate the risks associated with LLMs for three reasons. First, most risks cannot be reduced to zero (NIST, 2002). Hence, the question is not whether residual risks exist but how severe they are. Second, some risks stem from deliberate misuse, creating an offensive-defensive asymmetry wherein responsible actors constantly need to guard against all possible vulnerabilities while malicious agents can cause harm by exploiting a single vulnerability (Merwijk, 2022). Third, as I will expand on below, not all risks associated with LLMs can be addressed on the technology level.

Weidinger et al. (2021) list over 20 risks associated with LLMs divided into six broad risk areas. In Section 7.5, I highlighted how the three-layered approach helps identify and mitigate some of these risks: governance audits help protect against risks associated with malicious use; model audits help identify and manage information and misinformation hazards; and application audits help protect against discrimination as well as experiential harms. Of course, these are just examples. Audits at each level contribute, directly or indirectly, to addressing many different risks. However, not all the risks listed by Weidinger et al. are captured by the blueprint for how to audit LLMs outlined in this chapter.

Consider ‘automation harm’ as an example. Increasing the capabilities of LLMs to complete tasks threatens to undermine creative economies (Du Sautoy, 2019). While highly potent LLMs may remove the basis for some professions that employ people today – such as translators or copywriters – that is not necessarily a failure of the technology. The alternative of building less capable LLMs is counterproductive since abstaining from technology usage generates significant social and economic opportunity costs (Floridi et al., 2018).

The problem is not change per se but its speed and how the fruits of automation are distributed (Frey, 2019). Hence, problems related to changing economic environments may be better addressed through political reform rather than EBA. For this reason, it is important to remain realistic about what EBA can achieve and not fall into the trap of overpromising when introducing new governance mechanisms (Sloane, 2021). Yet the fact that EBA procedures cannot address all risks associated with LLMs does not diminish their merits. Instead, it points towards another important avenue for further research: how can and should social and political reform complement technically oriented mechanisms in holistic efforts to govern LLMs?

7.7 Implications for researchers, policymakers, and technology providers

Some of the features that make LLMs attractive also create significant governance challenges. The potential to adapt LLMs to a wide range of downstream applications undermines system verification procedures that presuppose well-defined demand specifications and predictable operating environments. My analysis in Section 7.4 thus concluded that existing EBA procedures are not well-equipped to assess whether the QMS put in place by technology providers and downstream developers are sufficient to ensure good governance of LLMs.

In this chapter, I have attempted to bridge that gap by outlining a blueprint for how to audit LLMs. In Section 7.5, I introduced a three-layered approach, whereby governance, model, and application audits inform and complement each other. During *governance audits*, technology providers' QMS are evaluated for completeness and adequacy. During *model audits*, LLMs' capabilities and limitations are assessed for performance, robustness, information security, and truthfulness. Finally, during *application audits*, products and services built on top of LLMs are first assessed for adherence with predefined ethics principles and subsequently evaluated based on their impact on users, groups, and the natural environment.

To conduct governance, model, and application audits, auditors can leverage a wide range of existing tools and methods, such as impact assessments, model evaluation, and red teaming. That said, the feasibility and effectiveness of my three-layered approach hinge on two factors. First, only when conducted in a combined and coordinated fashion can governance, model, and application audits enable different stakeholders to manage LLM-related risks. Hence, audits on the three levels must be connected in a structured process. Second, audits on all three levels must be conducted by an independent third-parties. The case for independent audits rests on concerns about not only the misaligned incentives that technology providers may face but the rapidly increasing capabilities of LLMs (Ziegler et al., 2022).

However, even when implemented under ideal circumstances, audits will not solve all tensions or protect against all risks of harm associated with LLMs. The limitations of my approach discussed in Section 7.6 are thus worth reiterating. To begin with, the feasibility of model audits hinges on the construct validity of the metrics used to assess characteristics like robustness and truthfulness. Further, my blueprint for how to audit LLMs does not specify who should conduct the audits it posits. An ecosystem of actors capable of implementing it has yet to emerge. Finally, not all risks associated with LLMs arise from processes that can be addressed through auditing. Some tensions are inherently political and require continuous management through public deliberation and political reform.

Researchers can contribute to overcoming these limitations by focusing on two of the avenues for further research I have identified. The first is to develop new methods and metrics to operationalise normative concepts in ways that are verifiable and maintain a high degree of construct validity. The second is to disentangle further the sources of different types of risks associated with LLMs. Such research would advance my understanding of how political reform can complement technically oriented mechanisms in holistic efforts to govern LLMs.

Policymakers can facilitate the emergence of an institutional ecosystem capable of carrying out audits of LLMs, for example by creating standardised evaluation metrics and reporting formats (Keyes et al., 2019), facilitating knowledge sharing (Epstein et al., 2018), and incentivise demonstrable achievements (Floridi et al., 2018). Policymakers should also consider updating existing and proposed ADMS regulations in line with the three-layered approach to audit LLMs outlined in this chapter. Take the AIA as an example. While the conformity assessments and post-market monitoring plans proposed by the European Commission mirror application audits, the AIA does currently not contain mechanisms akin to governance and model audits. Without amendments, such regulations are unlikely to generate adequate safeguards against the risks associated with LLMs.

My findings most directly concern technology providers, as they are primarily responsible for ensuring that LLMs are legal, ethical, and technically robust. So, what ought technology providers do? To start with, they should subject themselves to governance audits and the LLMs they design to model audits. That would create a demand for independent auditing bodies and help spark methodological innovation in governance and model audits. Mid-term, Technology providers should also demand that products and services built on top of their LLMs undergo application audits. That could be done through structured access procedures, whereby permission for using an LLM is conditional on such terms. Long-term,

like-minded technology providers should consider establishing, and funding, an independent industry body that conducts or commissions governance, model, and application audits.

It is worth ending this discussing with some words of caution. The blueprint for how to audit LLMs outlined in this chapter is not intended to replace existing governance mechanisms but to complement and interlink them by strengthening procedural transparency and regularity. Rather than being adopted wholesale by technology providers and policymakers, I hope that my three-layered approach can be adopted, adjusted, and expanded to meet the governance needs of different stakeholders and contexts.

7.8 Concluding remarks

Previous work on EBA has focused on developing procedures to audit ADMS that are used for specific tasks in predictable environments. In this chapter, I have outlined a blueprint for how to audit LLMs, which are adaptable to a wide range of downstream tasks. To the best of my knowledge, this is the first blueprint for how to audit ADMS with highly general capabilities published by any researcher, company, or regulator.

I have chosen to focus on LLMs specifically because they have broad societal impacts and are already widely employed today. However, the characteristics of LLM that undermine existing EBA procedures – including generativity and emergence – are not unique to LLMs but apply to all foundation models (Mondal et al., 2023; Muller et al., 2022). My analysis in this chapter thus provides a foundation for answering:

SQ5 What could blueprints for feasible and effective EBA procedures look like for ADMS with highly general capabilities?

A tentative answer to this question can be formulated as follows. To identify and mitigate the risks posed by ADMS with highly general capabilities, EBA procedures would need to integrate elements of both process-oriented assessments (of technology providers that design ADMS) and technology-oriented assessments (of ADMS – both before and after market deployment). In this chapter, I have provided an example of such methodological integration by outlining a three-layered approach, whereby *governance audits* (of technology providers), *model audits* (of ADMS after pre-training but prior to their release), and *application audits* (of applications based on ADMS) complement and inform each other. While further layers can be added, the integration of governance, model, and application audits serves as a baseline for designing feasible and effective EBA procedures for ADMS with highly general capabilities.

Further, it is not enough to conduct governance, model, and application audits in isolation. To identify and mitigate the ethical risks ADMS with highly general capabilities pose, governance, model, and application audits must be connected into a structured process. This means that outputs from audits on one level become inputs for audits on other levels. By doing so, EBA procedures provide the feedback loops necessary to address the governance challenges posed by highly general ADMS. For example, by assessing whether a downstream application aligns with the intended (or allowed) use of a specific ADMS, application audits build on information provided by governance and model audits to mitigate harms before they occur. Similarly, by ensuring that technology providers account for the known limitations and observable impact of ADMS, governance audits build on information provided by model and application audits to identify and mitigate risks upstream in the development process.

That said, the long-term feasibility and effectiveness of this blueprint for how to audit ADMS with highly general capabilities may also be undermined by future developments. For example, governance audits make sense when only a limited number of actors have the ability and resources to train and disseminate ADMS. Hence, the democratisation of ADMS capabilities – either through the reduction of entry barriers or a turn to business models based on open-source software – would challenge this status quo (Rao, 2020). Similarly, if ADMS become more fragmented or personalised (Kirk et al., 2023), there will be many user-specific instantiations of a single ADMS which would make model audits more complex to standardise. As a result, the blueprint for how to audit ADMS with highly general capabilities outlined in this chapter is only a useful starting point, it will need to be continuously revised in response to the changing technological and regulatory landscape.

CHAPTER 8

CONCLUSION

8.1 Synopsis

I opened this thesis by discussing the ethical opportunities and challenges associated with *automated decision-making systems* (ADMS). In Chapter 1, I illustrated through real-world examples that the use of ADMS generates benefits, such as increased efficiency and consistency, and risks, e.g., those related to data privacy and discriminatory outcomes (Tsamados et al., 2021). I also highlighted how technology providers and policymakers are experimenting with new governance mechanisms to manage those risks, with proposals ranging from ethical impact assessments to an outright ban of ADMS for some technologies and use cases. I further showed that many researchers have identified *ethics-based auditing* (EBA) as a promising yet underexplored governance mechanism that various actors can employ to assess whether the design and use of ADMS align with predefined ethics principles (e.g., Sandvig et al., 2014; Brundage et al., 2020; Brown et al., 2021).

Despite a growing literature on the topic, I argued that key questions regarding EBA remain unanswered by empirical research. In Chapter 2, I demonstrated that previous research has focused on *proposing* that ADMS should be audited for alignment with ethics principles (Diakopoulos, 2015; Kim, 2017), *developing* EBA procedures and tools (LaBrie & Steinke, 2019; Raji et al., 2020), or *conducting* EBA of ADMS (Robertson, 2018; Tolan et al., 2019). These and other early works have made important contributions. However, they have left central theoretical and practical questions unexplored. What are the affordances and constraints of EBA as a governance mechanism? What challenges do organisations face when implementing EBA procedures? And how can EBA complement other approaches to managing the ethical risks ADMS pose? These questions must be addressed to inform policymakers' and technology providers' ongoing efforts to design and implement EBA procedures.

The knowledge gap in the literature on EBA has – as I argued in Chapter 2 – both conceptual and empirical components. Without a shared understanding of what EBA is, let alone widely used standards for how it should be conducted, claims that an ADMS has been

audited are hard to verify and assess (Costanza-Chock et al., 2022). Further, the merits and limitations of specific EBA procedures are best studied in applied settings. The lack of empirical case studies in the field has thus left policymakers, researchers, and industry practitioners who design auditing procedures unable to anticipate and address the challenges organisations face when implementing EBA of ADMS.

This thesis set out to investigate whether and how EBA can help organisations design and deploy ADMS in ways that align with their organisational values. Pursuing that objective, I formulated two overarching research questions (RQs):

RQ1 What are the limitations of EBA as a governance mechanism for identifying and mitigating the ethical risks posed by ADMS?

RQ2 How can EBA procedures be designed to effectively identify and mitigate the ethical risks posed by ADMS while being feasible to implement?

In this thesis, I have tackled different aspects of these RQs. I have critically examined previous work on EBA to assess its theoretical affordances and constraints (Chapter 3), conducted an industry case study and presented new observational data to understand the practical limitations of EBA (Chapter 4), and explored how EBA can complement legislative approaches to govern ADMS by analysing the role of auditing in the proposed EU AIA (Chapter 5). I have also developed recommendations for how policymakers, auditors, and industry practitioners can demarcate the material scope of EBA procedures (Chapter 6) and design EBA procedures to manage the ethical risks posed by ADMS with highly general capabilities (Chapter 7).

In this final chapter, I will briefly summarise how the five substantive chapters of this thesis – while addressing their own subsidiary research questions (SQs) – have contributed to answering my RQs. My aim is to synthesise the findings presented in the previous chapters into an overarching thesis. Hence, I will not introduce any new data but instead focus on the bigger picture that emerges when interpreting the results of this research in aggregate.

The chapter is structured as follows. In Section 8.2, I summarise the contribution of previous chapters, demonstrating how each sheds light on the RQs at a conceptual, descriptive, or applied level. In Section 8.3, I synthesise my findings, arguing that despite theoretical and practical limitations, EBA procedures can – if properly designed and implemented – help organisations manage the ethical risks ADMS pose. I conclude that to be feasible and effective, EBA procedures must satisfy five conditions: procedural transparency and regularity, conceptual clarity, continuous monitoring, methodological integration, and operational

independence. In Section 8.4, I spell out the implications of my findings for my four target audiences: academic researchers, auditors, industry practitioners, and policymakers involved in designing and implementing EBA procedures. In Section 8.5, I highlight my research's limitations and discuss promising avenues for future research. I close with some concluding remarks in Section 8.6.

8.2 Summary of chapters and contributions

Each chapter of this thesis has contributed to answering the RQs in different ways. The first two did so indirectly. Chapter 1 introduced the RQs and established the academic and social relevance of this thesis. Chapter 2 reviewed previous research and provided a detailed analysis of both the scholarly and historical contexts within which the RQs should be interpreted. Despite being integral to this thesis, these chapters did not contain any theoretical advances or new empirical data. Hence, this section focuses on my five core chapters (3–7), which offer new theoretical and empirical insights and, in aggregate, form the basis for addressing my RQs.

It is worth restating that I approached my RQs on three levels: conceptual, descriptive, and applied. The *conceptual level* concerns what EBA is and how it works. I focused on that level in Chapter 3, in which my answers relied on a systematised literature review and theory synthesis. The *descriptive level* concerns how organisations integrate EBA with existing governance structures and the challenges they face in the process. To answer these questions, I conducted a longitudinal industry case study in Chapter 4, leveraging qualitative research methods like participant observation and interviews. Finally, the *applied level* concerns how EBA procedures can be designed to be feasible and effective. Chapters 5–7 made contributions at the applied level as they moved beyond evaluating existing policy options to propose new solutions. Given that the different chapters are based on journal articles with slightly different audiences, this structure was introduced in Chapter 1 and is re-emphasised here to help readers assess each chapter's contributions in its proper context.

8.2.1 Chapter 3. Ethics-based auditing of automated decision-making systems

Chapter 3 provided the thesis's conceptual foundation, introduced key concepts, and addressed SQ1: what are the affordances and constraints of EBA as a mechanism to address the ethical risks posed by ADMS? It offered three novel contributions by providing a theoretical framework for how EBA contributes to good governance (Section 3.5), proposing seven criteria for successfully designing EBA procedures (Section 3.6), and demonstrating that

existing EBA procedures are subject to several theoretical and practical constraints (Section 3.7). Let us consider these in turn.

The affordances of EBA as an ADMS governance mechanism can be summarised thus. EBA helps organisations demonstrate that the ADMS they design and deploy adhere to predefined ethics principles, relieves human suffering by anticipating and mitigating harms before they occur, and improves public trust in technology by promoting procedural transparency and regularity. This articulation can be contrasted with previous work, which has alluded only vaguely to how EBA contributes to good governance. Its value lies in forming a baseline against which the effectiveness of specific EBA procedures can be evaluated.

However, the affordances of EBA as an ADMS governance mechanism are potential and not guaranteed. This brings us to the second contribution of Chapter 3: deriving seven criteria for how to successfully design EBA procedures. For example, EBA procedures should be *collaborative*, i.e., enable constructive collaboration between auditors and technology providers, *continuous*, i.e., monitor and evaluate ADMS over time, and *drive re-design*, i.e., provide feedback that informs ongoing improvements to ADMS. However, it should be noted that these criteria were posited based on theory synthesis alone. Hence, they feed into – but do not exhaust – the conditions EBA procedures must satisfy to be feasible and effective (see Section 8.3.2), which also accounts for my findings at the descriptive and applied levels.

The third contribution of Chapter 3 was to demonstrate that EBA is subject to several theoretical and practical constraints. For example, it is difficult to anticipate and quantify the impact ADMS will have during ex-ante assessments, and auditors' struggle to secure the access they need to evaluate an ADMS. The taxonomy of constraints associated with EBA is, I argue, the most important contribution of Chapter 3. A new industry is emerging of private companies offering EBA services. However, these companies seldom acknowledge the limitations of such services. In Section 8.4, I thus propose that claims that an ADMS has been audited should be accompanied by a disclosure of the auditing procedure and a discussion about what confidence we can have in the assessment.

8.2.2 Chapter 4. Operationalising corporate governance through ethics-based auditing

Chapter 4 presented the main body of empirical research conducted for this thesis. It addressed SQ2: how do organisations integrate EBA with existing governance structures, and what challenges do they face in the process? To answer this SQ, I observed and analysed AstraZeneca's internal activities over 12 months as it prepared for and underwent EBA in collaboration with a third-party auditor. Using qualitative research methods, I generated

knowledge about the organisational context in which EBA procedures must be integrated to impact the design and deployment of ADMS.

I made two novel contributions in Chapter 4. First, I provided a descriptive account of how and why an R&D-driven company like AstraZeneca uses EBA to operationalise its commitment to high-level ethics principles. Second, I described the practical challenges and tensions involved in conducting EBA in a real-world setting. Both contributions fill important gaps in the existing literature. Previous research in the field contains few case studies.[148] To the best of my knowledge, the case study in Chapter 4 was the first in which an independent researcher studied a technology provider's internal activities before, during, and after EBA.

My observational data indicated that technology providers have strong incentives to subject themselves to EBA. In AstraZeneca's case, the motivating factors included the need to manage financial, legal, and reputational risks, the competitive pressures forcing technology providers to continuously improve their QMS, and the desire of individuals to design and use ADMS responsibly (see Section 4.3). Given the permanence of these factors, I argued that the demand for new procedures to audit ADMS is likely to continue accumulating even in the absence of forthcoming regulation.

Further, this case study suggested that the main difficulties organisations face when designing or implementing EBA mirror well-known corporate governance challenges. For example, AstraZeneca struggled to harmonise standards across the organisation, demarcate the material scope for ADMS governance, define key performance indicators for ADMS, and act on the results produced by EBA (see Section 4.6). These descriptive findings will form a cornerstone of the analysis when I address my overarching RQs in Section 8.3.

8.2.3 Chapter 5. The role of auditing in the proposed EU AIA

Chapter 5 was the first of three chapters that approached my RQs at the applied level. It addressed SQ3: how can EBA complement legislative approaches to managing the risks posed by ADMS? It did so by exploring the role of auditing play in the EU AIA. The chapter provided an in-depth analysis of the two primary governance mechanisms the AIA proposes: the conformity assessments providers of high-risk ADMS will have to undergo, and the post-

[148] A few articles reporting on challenges and best practices from real-world EBA have been published (e.g., Hasan et al., 2022; Zicari et al., 2021). However, these tend to be written by the auditors rather than by independent academic researchers.

market monitoring plans that providers are expected to establish once the AIA comes into force.[149] Through this analysis, three novel contributions emerged.

First, I argued that the AIA implicitly proposes the establishment of a Europe-wide ecosystem to audit ADMS (Section 5.3). While the AIA only occasionally refers to auditing, the governance mechanisms it proposes have both procedural and substantive affinities with ex-ante and ex-post audits. Moreover, the institutional relationship between what the AIA refers to as 'notified bodies' and 'notifying bodies' mirrors that between auditors and national accreditation bodies. This observation matters both practically, because it helps companies understand what the AIA expects of them, and theoretically, since it anchors the discussion about how to refine the AIA in the vast literature on auditing ADMS.

Second, I identified areas of the AIA in which revisions or clarifications would be helpful (Section 5.7). To do so, I conducted a gap analysis, comparing its provisions with best practices for auditing ADMS. I found that, despite its merits, the AIA can be improved. For instance, it does not specify how to conduct conformity assessments. The gap analysis produced seven recommendations to further strengthen the auditing ecosystem outlined in the AIA, including the need to clarify the proposed legislation's material scope and create standardised evaluation metrics and reporting formats.

Third, I showed that the European Commission encourages technology providers to adopt voluntary codes of conduct and implement EBA procedures (Section 5.6). I argued that it does so for two reasons: to foster the application of the AIA's requirements even for ADMS that are not classified as high-risk and to promote post-compliance ethical behaviour. The main takeaway from Chapter 5 was thus that EBA procedures are compatible with, and complementary to, hard regulations concerning the design and use of ADMS.

8.2.4 Chapter 6. Models for classifying ADMS

Continuing at the applied level, Chapter 6 addressed SQ4: how can the material scope for EBA be demarcated? This question emerged from my empirical research, and its importance became clear as it re-surfaced in multiple studies. For example, in Chapter 4, my observational data revealed that one of the main obstacles AstraZeneca faced when preparing for its 'AI audit' was demarcating its material scope. Similarly, in Chapter 5, my analysis of the role of auditing in the EU AIA suggested that the lack of a clear material scope will undermine the proposed

[149] In Chapter 6, I used the term AI system to reflect the EU AI Act's terminology. However, here I use the term ADMS for in-chapter consistency.

legislation's effectiveness. SQ4 was thus formulated to tackle a real-world problem faced by technology providers and policymakers.

Pursuing a solution-oriented research agenda, I reviewed and compared previous attempts to classify ADMS for the purpose of implementing EBA procedures. Based on that analysis, I offered two contributions to the ADMS auditing literature.

First, I developed a novel taxonomy of models to demarcate the material scope of ADMS governance (Sections 6.4–6.6). According to the *binary approach*, systems either are or are not considered ADMS depending on their intrinsic characteristics. According to the *risk-based approach*, systems are classified into different categories depending on the types of ethical risks they pose. Finally, according to the *multi-dimensional approach*, various aspects – such as context, data input, and decision-model type – need to be considered when classifying systems. I labelled these approaches the Switch, the Ladder, and the Matrix, respectively. In doing so, I provided organisations that design or implement EBA procedures with the vocabulary they need to have an informed discussion about available policy options regarding how to demarcate their material scope.

Second, I demonstrated that the logic according to which ADMS are classified is an integral part of the design of EBA procedures (Section 6.7). This contrasts with previous research, which has predominantly focused on the ontological question of what an ADMS *is*, implying that the question can and should be answered a priori. Taking an explicitly pragmatic stance, my findings suggested a different path: it is less important to define an ADMS in abstract terms and more important to establish processes to classify ADMS in ways that promote successful actions for some specific end.

8.2.5 Chapter 7. Ethics-based auditing of large language models

Chapter 7 also addressed a real-world problem faced by technology providers and policymakers alike: how to audit ADMS with highly general capabilities. Previous research has focused on developing procedures to audit ADMS used for specific tasks. However, the capabilities of ADMS are becoming increasingly general. For example, large language models (LLMs) are adaptable to a wide range of downstream applications, which undermines the effectiveness of EBA procedures designed to ensure compliance with sector-specific norms and regulations. To bridge that gap, Chapter 7 addressed SQ5: what could a blueprint for feasible and effective EBA procedures look like for ADMS with highly general capabilities?

The chapter's main contribution was a novel blueprint for auditing LLMs. I proposed a three-layered approach, wherein *governance audits* (of technology providers that design and

disseminate LLMs), *model audits* (of LLMs after pre-training), and *application audits* (of applications based on LLMs) complement and inform each other (Section 7.5). The tools and procedures to conduct audits at the three levels already exist. However, to provide meaningful assurance, I argued, governance, model, and application audits must be combined into structured and coordinated procedures. This three-layered approach was, to the best of my knowledge, the first-ever published blueprint for auditing LLMs.

In the process of introducing and discussing the three-layered approach, I made two secondary contributions. First, I derived and defended seven claims about how to audit LLMs (Section 7.4). I argued that LLM audits should be external yet collaborative, incorporate elements of both process-oriented and technology-oriented audits, and include continuous monitoring of system outputs. Second, I identified the conceptual, technical, and practical limitations associated with any attempt to audit LLMs (Section 7.6). For instance, I showed that not all ethical risks posed by LLMs can be addressed at the technology level and that construct validity remains a major challenge during model audits. Together, the secondary contributions offered in Chapter 7 provide a foundation that future researchers can use when designing new, more refined, LLM auditing procedures.

The three-layered approach for auditing LLMs holds valuable lessons for how to audit other ADMS with highly general capabilities. I focused on LLMs because they pose a wide range of social and ethical risks with which technology providers and policymakers are currently struggling. However, the features of LLMs that make them difficult to audit – including adaptability and complexity – apply to other ADMS. To conclude, the three-layered approach provides a blueprint not only for how to audit LLMs but also for what feasible and effective EBA procedures could look like more generally.

8.3 Synthesis of findings

In the previous section, I summarised the contributions made in the five substantive chapters of this thesis. While these chapters answered distinct SQs, they also shed light on the overarching RQs in different ways. In this section, I consider my findings in aggregate and explain how they contribute to answering my RQs and addressing my larger research objectives. Rather than repeating the contributions previously discussed, this section highlights the broader conclusions that arise from their integration.

8.3.1 Considerations and challenges

This thesis set out to explore the limitations of EBA as a governance mechanism for managing the ethical risks ADMS pose (RQ1). So, what have we learned in this regard? Taken together, my findings suggest that EBA is subject to a wide range of conceptual, technical, economic, social, and institutional limitations. While these limitations were discussed in-depth in Chapter 3, I will revisit the most important ones here alongside examples from my empirical research.

To begin with, EBA is subject to conceptual limitations that cannot be easily overcome by either technical innovation or policy design. For example, EBA procedures can never fully eliminate but only help identify and mitigate the ethical risks ADMS pose. That is primarily because it is difficult (perhaps impossible[150]) to anticipate the impact an ADMS will have. Although researchers and policymakers accept this limitation in theory, it is not always sufficiently accounted for in their communication. For example, EBA is repeatedly referred to as a means to *ensure* that ADMS are ethical (e.g., European Commission, 2021a; Felländer et al., 2022). Professional service providers may have reasons to market EBA procedures that way. However, based on the conceptual limitations identified in this thesis, researchers and policymakers are advised to use more carefully crafted language when describing the affordances of EBA procedures.

Further, implementing EBA procedures requires high-level ethics principles to be translated into verifiable criteria. However, such efforts face conceptual difficulties. Different ethics principles sometimes conflict and require tradeoffs. In Chapter 4, I described how AstraZeneca uses ADMS to detect treatment response patterns amongst patients receiving specific drugs. The risk that such ADMS produce harmful outcomes must be balanced against the lives they can save. As this example illustrates, it is often not possible to err on the safe side. Moreover, the apparent consensus around ethics principles like fairness and transparency masks disagreements about how these should be interpreted. For example, there exist more than six definitions of fairness, some mutually incompatible (Narayanan, 2018). While EBA can assess how 'fair' an ADMS is according to a specific metric, we should not expect it to resolve these normative tensions. Again, this runs counter to convention. In the ADMS auditing literature, EBA is commonly envisioned as a mechanism to 'ensure fairness and transparency

[150] The extent to which future events are determined by (or can be predicted based on) current material and social conditions have long been a contentious question in both philosophy (Dafoe, 2015) and sociology (Schroeder, 2007). While we have no way of settling that question empirically, pragmatists maintain that the future shaped but not determined by material and social conditions (Dewey, 1922).

in the ADMS that impact us all' (Metaxa et al., 2021). The limitations identified in this thesis suggest that a more realistic function of EBA would be to make visible implicit choices and tensions and arrive at resolutions that, even when imperfect, are at least publicly defensible.

The feasibility and effectiveness of EBA are also subject to technical limitations rooted in the autonomous, complex, and adaptable nature of ADMS. For example, the behaviour of ADMS in test settings is not always indicative of their behaviour in real-world environments (Auer & Felderer, 2018). The most critical technical limitation, however, is the difficulty of operationalising normative concepts during EBA. Take explainability as an example. ADMS draw inferences from large datasets in ways that appear opaque to the human mind (Burrell, 2016). While many tools and methods have been developed to improve the interpretability of complex ADMS, information is invariably lost through reductive explanations. Similar problems related to construct validity exist for other normative concepts.

These technical limitations are well-known. In fact, systems engineers and computer scientists are already developing methods to audit autonomous ADMS (Strenge & Schack, 2020) and tools to assess the interpretability of complex ADMS (Kroll, 2018). However, my findings suggest that the technical limitations of these methods and tools have been overlooked in the design of EBA procedures. Consider LLMs as an example. Previous work has focused on evaluating LLMs based on input-output relationships (Mayson, 2019) or benchmarking their performance on specific tasks (Aspillaga et al., 2020). Such technology-oriented approaches are useful since they help gather evidence about the properties of LLMs. However, the technical limitations identified in this thesis suggest that they need to be complemented with process-oriented audits of how LLMs are designed and deployed.

In addition to conceptual and technical limitations, the feasibility and effectiveness of EBA are also constrained by economic and social factors. For example, EBA is a time-consuming activity. In AstraZeneca's case, external auditors and in-house employees combined invested around 2,000 person-hours throughout the audit. And yet, that was a comparatively light-touch audit which only assessed a limited sample of ADMS-related projects within AstraZeneca. Moreover, EBA is associated with financial costs, including the auditors' fees, the procurement of equipment and licenses, and the hiring and training of in-house staff to support the audit. While quantifying the total cost of EBA is difficult, rough indications can be given. For example, researchers have estimated that providers of high-risk ADMS will spend approximately 10–14% of their development costs to demonstrate adherence to the EU AIA's requirements (Renda et al., 2021; Haataja & Bryson, 2021).

To control these costs, there is a risk that technology providers reduce EBA to a box-ticking exercise. That would limit its effectiveness because, as Chapter 4 demonstrates, it is precisely the manual elements – whereby auditors ask open-ended questions to spark ethical deliberation amongst developers – that are key to proactively identifying and mitigating the risks ADMS pose. Social factors also limit the space for such interactions. My interviews with managers and software developers suggest that they are more motivated to develop new ADMS than to document their work or meet with auditors. Hence, the effectiveness and feasibility of EBA hinge on both the resources available for conducting the audit and the motivations of the actors involved. This points towards a critical gap in the existing literature, which has focused on developing technical tools or step-by-step procedures for auditing ADMS.[151] The findings presented in this thesis, however, suggest that the main bottleneck to implementing EBA procedures is not a lack of tools but that these are not being employed in a rigorous and structured manner due to economic and social factors.

Finally, the effectiveness and feasibility of EBA as an ADMS governance mechanism are limited by institutional constraints. A governance mechanism is only as good as the institution backing it (Boddington, 2017). However, an institutional ecosystem to conduct EBA – and verify claims that ADMS have been audited – has yet to emerge (CDEI, 2021c). Currently, EBA is conducted by a plurality of decentralised actors without standardised reporting formats. This leaves room for malpractices like 'ethics-bluewashing', whereby technology providers make unsubstantiated claims about ADMS to appear more ethical than they are (Floridi, 2019b). Consequently, the European Commission (2021a) has proposed that independent third parties audit high-risk ADMS. Still, the institutional ecosystem sketched in the EU AIA is only intended to support legally mandated audits, not EBA. To address that gap, researchers have proposed different models for structuring EBA, including the creation of platforms for sharing audit reports (Keyes et al., 2019) and the formation of new industry bodies to develop sector-specific evaluation criteria and reporting standards (Falco et al., 2021). In Section 8.3.2, I will build on these proposals when discussing what blueprints for feasible and effective EBA procedures could look like.

[151] There are a few exceptions, which thoroughly discuss the challenges organisations face when attempting to implement EBA (e.g., Hasan et al., 2022; Landers & Behrend, 2022).

In the above discussion, I have highlighted only the most important limitations of EBA as a governance mechanism for ADMS. The full range of limitations identified and discussed in the preceding chapters are summarised in Table 4 below.

Table 4. Summary of EBA's limitations as an ADMS governance mechanism.

Type	***Constraints***
Conceptual	Lack of consensus around high-level ethical principles
	Normative values conflict and require trade-offs
	It is difficult to quantify the externalities of ADMS
	Information is lost through reductionist explanations
	It is difficult to demarcate the material scope of ADMS governance
Technical	ADMS appear opaque and are hard to interpret
	EBA exposes data integrity and privacy risks
	Linear compliance mechanisms are incompatible with agile development
	Tests may not indicate the behaviour of ADMS in real-world environments
Economic and Social	EBA incurs financial and administrative costs
	Audits are vulnerable to adversarial behaviour
	EBA may disproportionately disadvantage specific sectors or groups
	The transformative effects of ADMS challenge notions of human dignity
	Emerging audit frameworks reflect and reinforce existing power relations
	Employees may lack incentives for or interest in conducting EBA
Institutional	There is a lack of institutional clarity about who audits whom
	Auditors may lack the access or information required to evaluate ADMS
	The global nature of ADMS challenges national jurisdictions

In this section, I have answered RQ1: what are the limitations of EBA as a governance mechanism for identifying and mitigating the ethical risks posed by ADMS? In doing so, I have approached the question both conceptually and descriptively by advancing theoretical arguments and presenting new empirical data to support my conclusions. Importantly, however, the limitations highlighted above are not intended to diminish EBA's merits as an ADMS governance mechanism. Instead, they serve a constructive purpose: to design feasible and effective EBA procedures, these limitations must be understood and accounted for.

8.3.2 Paths forward

As discussed in Chapter 1, I adopt a pragmatist stance, according to which research should be grounded in real-world problems and solution-oriented. Hence, in addition to exploring EBA's

limitations, this thesis also explored RQ2: how can EBA procedures be designed to effectively identify and mitigate the ethical risks posed by ADMS while being feasible to implement? Considered in aggregate, my findings suggest that to be feasible and effective, EBA procedures should satisfy five conditions: (i) procedural transparency and regularity, (ii) conceptual clarity, (iii) continuous monitoring, (iv) methodological integration, and (v) operational independence.

Here, I expand on that conclusion and explain how it is supported by my findings. However, it is useful to first reflect on what it means to 'design' an EBA procedure. In Chapter 2, I showed that audits can be structured in many ways and that policy design choices concern not only *what* should be audited, *when*, and according to *which* criteria, but also *who* should conduct the audit and *how* results should be published. The answers to these questions – i.e., how auditing procedures are designed – significantly impact the confidence we can have in the results audits produce. Equipped with that clarification, we can now proceed to answer RQ2.

To start with, EBA should follow structured and transparent procedures because the principal affordances of EBA as a governance mechanism are undermined if audits are conducted in unstructured or opaque ways. In Chapter 3, I demonstrated that EBA contributes to good governance in several ways. It can improve trust between different stakeholders by verifying technology providers' claims about ADMS, provide a basis for holding decision-makers accountable in case of irregularities by mapping organisational roles and responsibilities, and help identify and mitigate harms before they occur by sparking ethical deliberation amongst software developers. However, these affordances are potential and not guaranteed. EBA conducted on an ad-hoc basis is not a robust mechanism for verifying technology providers' claims. Similarly, to provide a basis for accountability, audit reports must be communicated transparently and proactively to relevant stakeholders.

Most existing EBA procedures do not live up to that standard. While previous research stresses the need for technology providers to document how ADMS are designed and deployed, the procedures auditors use to assess technology providers and their ADMS are typically not disclosed. For example, the EBA procedure employed in AstraZeneca's case was never made public since it constituted a trade secret for the company that conducted the audit. This indicates the tension between intellectual property and procedural transparency. Researchers are more transparent about their EBA procedures (see, e.g., Koshiyama et al., 2022; Zicari et al., 2021). However, due to a lack of resources and incentives to publish new findings, researchers typically conduct EBA only irregularly.

In addition to procedural regularity and transparency, feasible and effective EBA procedures require conceptual clarity – with respect to the audit's material scope and normative

baseline. Consider these points in turn. Every policy needs to define its material scope (Schuett, 2021); however, while many organisations have published high-level ethics principles that guide their design and use of ADMS, it often remains unclear to which systems, exactly, these principles apply. This observation has resurfaced in different forms throughout this thesis. In Chapter 4, my case study suggested that technology providers struggle to demarcate the material scope of ADMS governance. In Chapter 5, my analysis found that there is no consensus amongst policymakers on how to define ADMS. In Chapter 6, I addressed that issue head-on, concluding that without a clearly defined material scope, ADMS will only be scrutinised on an *ad hoc* basis. That undermines the legitimacy of EBA and hampers its ability to identify and mitigate ethical risks.

Relatedly, EBA presupposes a normative baseline against which ADMS can be evaluated. Although widely accepted in theory, this point is rarely observed in practice. For example, when conducting EBA, it is not enough to list the ethics principles that should guide the design and deployment of ADMS. Auditors also require guidance on how these principles should be interpreted. Take explainability as an example. What counts as an explanation varies between contexts and hinges on the *target audience* and the *purpose* of the explanation (Larson & Heintz, 2020; Watson, 2021). Hence, for claims that ADMS have been audited for explainability to be verifiable, the baseline must be both clearly defined and operationalisable. In short, feasible and effective EBA procedures presuppose conceptual clarity.

Further, EBA procedures should include elements of continuous monitoring of the outputs of ADMS. This conclusion follows directly from the conceptual limitations of EBA procedures previously discussed. Because autonomous and adaptable ADMS learn and acquire new capabilities as they operate in dynamic environments (Russel & Norvig, 2015), it is difficult to identify and mitigate all risks upfront. Hence, EBA procedures focusing only on ex-ante assessments cannot address all the governance challenges ADMS pose. In Chapter 3, I proposed a pragmatic solution: continuously monitoring the operations of ADMS throughout their lifecycle. This solution has two advantages. First, real-time monitoring of ADMS allows auditors to detect undesired behaviours early and potentially intervene to prevent further harms (Jotterand & Bosco, 2020). Second, information about the outputs of ADMS constitutes valuable feedback that should inform their continuous re-design (Tran & Daim, 2008).

As highlighted in Chapter 2, continuous monitoring is a well-established practice in the governance of safety critical systems. However, in the field of ADMS auditing, researchers and policymakers have only recently begun appreciating its importance (Strenge & Shack, 2020; Minkkinen et al., 2022). The post-market monitoring plans mandated in the EU AIA

constitute a rare and commendable exception. These require technology providers to document and analyse the behaviour of high-risk ADMS throughout their lifecycles and flag any serious incidents or malfunctioning (European Commission, 2021a). Such continuous monitoring is well-suited to address the specific governance challenges ADMS pose and should be incorporated into EBA procedures.

The need for continuous monitoring illustrates a more general point: to be feasible and effective, EBA procedures should cover all stages of the ADMS development and deployment lifecycle. This implies that EBA should combine *process-oriented* audits of technology providers that design and deploy ADMS and *technology-oriented* audits of ADMS. The rationale for this is that process-oriented and technology-oriented audits have distinct yet complementary affordances. They are both individually necessary and individually insufficient to address all the governance challenges ADMS pose.

This conclusion has been foreshadowed in almost every preceding chapter. In Chapter 2, I argued that previous work on EBA can be divided into *narrow* and *broad* approaches. The former is technology-oriented and aims to assess the properties or capabilities of ADMS. The latter is process-oriented and focuses on the adequacy of technology providers' governance structures and QMS. However, although the distinction is analytically useful, there is no conflict between the two approaches. In Chapter 3, I went further and demonstrated that EBA must combine elements of technology-oriented and process-oriented assessments to identify and mitigate the different ethical risks ADMS pose.

In practice, however, the proposal to combine process-oriented and technology-oriented audits is aspirational. Currently, the two approaches constitute distinct strands of research and practice that rarely converse with each other. Technology-oriented audits are common in computer science and social science research; process-oriented audits dominate in disciplines like systems engineering and organisation studies as well as in applied settings. For example, AstraZeneca's EBA process was a governance audit, focusing exclusively on assessing the adequacy of the organisation's software development process and not containing any technical evaluations of ADMS. I concluded that such procedures, while useful to improve technology providers' software development processes, are fundamentally unable to produce verifiable claims about ADMS. Invertedly, while technology-oriented audits can help identify the limitations of ADMS, they do not reveal much about their root causes.

The good news is that the tools and methods to conduct both process and technology-oriented audits already exist.[152] The next step in the evolution of feasible and effective EBA procedures should thus be integrating the two approaches. Importantly, it is insufficient to conduct technology and process-oriented audits in isolation; they must be connected in structured procedures. In Chapter 7, I provided an example of this by demonstrating how process-oriented audits (of technology providers that train and disseminate LLMs) and technology-oriented audits (of LLMs both prior to and after fine-tuning) complement and inform each other. While Chapter 7 focused on LLMs specifically, the methodological integration it proposed can and should be used to audit other ADMS as well.

Finally, EBA should be conducted by independent third-party auditors. Per definition, auditing presupposes operational independence between the auditor and the auditee. That can be achieved in several ways. For example, many organisations have internal auditors who operate independently from line managers and report directly to their boards. As I argued in Chapter 3, both internal and external audits have their own strengths and weaknesses.[153] For instance, it is often easier for internal auditors to secure access to the information and personnel required to conduct an audit. However, taken together, the evidence suggests that external audits are required to address the governance challenges ADMS pose.

There are three reasons for this. First, external auditors' involvement contributes to the objectivity and professionalism of audits (Power, 1997). My case study of AstraZeneca supported that conventional wisdom by showcasing how external auditors challenged the confirmation bias that had prevented in-house experts from recognising critical flaws. Second, specialised knowledge is required to conduct technology-oriented audits of ADMS.[154] External auditors bring expertise that not all companies that design or deploy ADMS can access in-house (Bauer, 2016). Third, the involvement of external auditors increases accountability because they are scrutinised by regulatory bodies and risk losing their licenses if they operate irregularly or unethically (Raji et al., 2022). Given the competitive pressures technology providers face to design and deploy ADMS rapidly, such external accountability and oversight is a prerequisite for good governance.

---

[152] For an overview of these tools and methods, see Sections 2.4, 3.4, and 7.3.

[153] The distinction between internal and external audits is analytically useful. However, it should not be overemphasised. In practice, external auditors often rely on internal auditors to gather information and get access to systems and personnel (Haron et al., 2004).

[154] In Chapter 7, I demonstrated this point when discussing the complexities involved in evaluating the properties and capabilities of LLMs (see Section 7.4.3).

To summarise, considered in aggregate, my findings suggest that to be a feasible and effective governance mechanism for identifying and mitigating the ethical risks ADMS pose, EBA procedures should satisfy five conditions:

1) *Procedural regularity and transparency*, i.e., that EBA is conducted in a structured way, and that the methodology used to conduct the audit and its results are transparently and proactively communicated to relevant stakeholders.

2) *Conceptual clarity* i.e., that the material scope of the EBA is clearly demarcated, and that the normative baseline for the assessment can be operationalised.

3) *Continuous monitoring*, i.e., that EBA procedures incorporate elements of continuous auditing, including the monitoring of the outputs of ADMS throughout their lifecycle.

4) *Methodological integration*, i.e., that EBA integrates process-oriented assessments (of technology providers designing or using ADMS) and technology-oriented assessments (of the properties and capabilities of ADMS) into structured procedures.

5) *Operational independence*, i.e., that EBA is conducted by external auditors who are accountable to policymakers or independent industry bodies.

Of course, I am not the first to highlight each of these conditions. Considered in isolation, each has been defended by different authors in different contexts. However, what the five conditions lack in novelty they make up for in unity. For within them, we find almost all that is needed to design and implement EBA procedures that are feasible and effective in identifying and mitigating the social and ethical risks posed by ADMS. Of course, this conclusion needs to be qualified, and in Section 8.5 I will discuss some important limitations of my work. But first, having answered my two overarching RQs, I now turn to discuss the implications of my findings for different target audiences.

8.4 Implications and policy recommendations

In Chapter 1, I envisioned four main audiences for this work: *academic researchers* studying how ADMS can be governed and audited; *auditors* developing EBA procedures or offering EBA services to technology providers; *industry practitioners* implementing EBA procedures in organisations that design and deploy ADMS; and *policymakers* drafting legislation and guidance on how to govern ADMS. In this section, I consider the implications my findings have for each group.

8.4.1 Academic researchers

In Chapter 2, I showed that EBA is a multidisciplinary field of study, harbouring contributions from computer science, systems engineering, law, social science, media and communication studies, political and moral philosophy, and organisational studies alike. For that diverse community of academic researchers, this thesis offers two main takeaways.

The first is cautionary. The ethical risks ADMS pose are real and pressing, and EBA procedures can indeed help identify and mitigate some of them. However, my research has demonstrated that EBA is subject to both conceptual and practical limitations. Having studied how EBA is implemented in applied settings, I fear that academic researchers – guided by good intentions and the desire to propose concrete solutions – have overstated the merits of EBA and underestimated its limitations. This may lead to resources being wasted on less effective governance initiatives and lend unjustified legitimacy to EBA procedures that provide only a false sense of security. In short, academic researchers proposing EBA as a remedy for the ethical risks posed by ADMS should be careful not to overstate its merits.

The second implication for academic researchers is more constructive. My research has highlighted several ways in which further research can strengthen the feasibility and effectiveness of EBA procedures. These include developing new tools and methods to operationalise normative values during technology-oriented assessments and evaluating different institutional arrangements for structuring independent EBA. In Section 8.5, I will expand on these and other avenues for future research. Here, I wish to emphasise a more general point. In Chapter 2, I showed that while researchers in computer science and the social sciences focus on technology-oriented audits of ADMS, researchers from systems engineering and organisation studies focus on process-oriented audits of technology providers that design or deploy ADMS. I also observed that there is only limited dialogue between the literature produced by these distinct academic communities. That becomes problematic when – as I concluded in Section 8.3 – EBA procedures must incorporate elements of both technology-oriented and process-oriented assessments to be feasible and effective. EBA research would thus benefit from increased cross-disciplinary collaboration and knowledge transfer.

8.4.2 Auditors

My research also has direct implications for auditors that design EBA procedures or offer EBA services. ‘Auditors’ in this context refers to a heterogeneous group that includes professional services firms, startups, and non-profit organisations. Despite having different incentives, all these organisations have developed EBA procedures to help technology providers identify and

mitigate the risks ADMS pose. Auditors are advised to pay specific attention to the conditions feasible and effective EBA procedures must satisfy outlined in Section 8.3.2.

Amongst the best practices that have emerged from my research, three are worth re-emphasising here since they suggest alternatives to the status quo. First, on their own, ex-ante auditing procedures are ill-equipped to address the governance challenges associated with autonomous, complex, and adaptable ADMS. This implies that auditors should spend less time and effort on conducting snapshot audits and more on developing continuous auditing procedures to monitor both technology providers' conduct and the outputs of ADMS over time. Second, EBA procedures are most likely to identify and mitigate risks when conducted in collaboration with technology providers. This means that – while adversarial audits are also important – auditors should develop collaborative EBA procedures that inform the continuous re-design of ADMS. Third, EBA is most effective when conducted transparently and consistently. To improve public trust in ADMS, claims that EBA has been conducted should be accompanied by transparent communication concerning how the audit was conducted, the limitations associated with the employed methodology, and the level of confidence we can have in the audit's results in light of those limitations.

8.4.3 Industry practitioners

Industry practitioners implementing EBA procedures in organisations that design or deploy ADMS constitute another target audience for my research. Typically, this responsibility falls on managers with titles like chief information officer, responsible AI lead, head of IT compliance, or internal audit director.

Industry practitioners implementing EBA procedures should take two main lessons from this thesis. The first concerns how to implement EBA procedures. As illustrated by my case study in Chapter 4, EBA procedures are most effective when integrated into existing governance structures. That is because managers and developers may perceive procedures that duplicate existing structures as unnecessary by the managers and developers expected to implement them. Rather than creating new tools and reporting channels, industry practitioners should explore how EBA procedures can inform, interlink, and revise existing software development processes and QMS, thereby complementing and enhancing them.[155]

[155] This conclusion is supported by established theory: the ethical behaviour of an organisation is embedded in its operating model through processes, roles and responsibilities, incentives, etc. (Crane & Matten, 2016). Consequently, EBA should not be conducted in isolation but linked to organisational governance as a whole.

The second lesson concerns organisations' motivations for implementing EBA. Previous research has emphasised the need to manage financial, legal, and reputational risks as the main drivers for implementing EBA. While supporting those claims, my findings suggest that, from a corporate governance perspective, EBA can fill other functions – like facilitating agenda setting, catalysing internal change, and expanding organisational units' mandates. Industry practitioners implementing EBA should explore and exploit these alternative motivations to secure support and resources from different internal decision-makers.

8.4.4 Policymakers

Finally, my findings have implications for policymakers who mandate EBA procedures as part of larger efforts to govern ADMS. This includes the European Commission, the US's Federal Trade Commission and Government Accountability Office, and the UK's Information Commissioner's Office and Center for Data Ethics and Innovation.

As I have repeatedly stressed, the primary responsibility for demonstrating that ADMS are legal, ethical, and technically robust rests with technology providers. That said, self-regulation brings inherent challenges, and its effectiveness is not independent of the guidance and support policymakers provide (Floridi, 2021a). In fact, my research has indicated that policymakers and regulators can do much to facilitate the emergence, incentivise the adoption, and strengthen the effectiveness of EBA procedures. In what follows, I highlight six recommendations for policymakers that follow from the findings discussed in Section 8.3. Policymakers should consider to:

1) *Create standardised evaluation metrics and reporting formats*. Standardised formats for evaluation and communication help technology providers and users assess and compare different ADMS. While technology providers should be free to pilot different EBA procedures, policymakers can strengthen the synergies between their efforts by standardising metrics and reporting formats.

2) *Facilitate knowledge sharing and the communication of best practices*. As researchers and auditors accumulate knowledge locally, it would be beneficial if they collaborated with a commitment to reproducibility and shared know-how and technical solutions. To support such collaboration,[156] policymakers should provide digital platforms to

[156] For example, the sharing of past failures of ADMS could help mitigate future harms (Brundage et al., 2020).

facilitate exchanges of code and data and incentivise the sharing of best practices for designing and implementing EBA.

3) *Create an independent body to oversee EBA*. While the piloting of different EBA procedures should be encouraged, policymakers should avoid shifting the ultimate enforcement of procedural standards from judicial courts to private actors. The solution here is to create an independent body that authorises the auditors who conduct EBA. This agency's role would not be to evaluate the design of ADMS directly but rather to scrutinise and approve the EBA processes auditors employ.

4) *Incentivise technology providers' adoption of EBA*. The implementation of EBA will be slow if organisations designing or using ADMS perceive that, *for them*, the costs outweigh the benefits. Policymakers should thus reward demonstrable achievements to incentivise the adoption of EBA. This includes monetary incentives, like tax breaks, and immaterial acknowledgements, like publishing lists of technology providers that adhere to specific standards or best practices in their design and use of ADMS.

5) *Ensure accountability*. Policymakers can strengthen public trust in EBA procedures by ensuring accountability, e.g., by imposing sanctions when required. Technology providers publishing inaccurate or misleading information ought to be fined and complicit auditors lose their licenses. Such sanctions are compatible with the voluntary nature of EBA. A parallel can be made to the European food-labelling regulation: while food may contain both non-vegetarian and vegetarian ingredients, mislabelling one for the other is not allowed.[157]

6) *Provide governmental leadership*. Policymakers should lead by example. As ADMS are used throughout the public sector (Levy et al, 2021),[158] the first step should be to subject such ADMS to EBA. Doing so would not only improve accountability in the public sector but also contribute to standardising EBA procedures around which other actors can gather.[159]

---

[157] See Regulation (EU) No 1168 (2011).

[158] Recent reports documenting the public sector's use of ADMS include those produced by the Ada Lovelace Institute (2021) and the ELI (2022).

[159] In most countries, the public sector accounts for 30–50% of the economy (OECD, 2023). Consequently, public sector standards significantly impact how private sector actors operate.

In this section, I have discussed my findings' implications for different audiences. However, it is important to note that the relatively new field of researching EBA lacks established theories and practices. The implications and policy recommendations discussed above are thus provisional and may require revision based on new empirical data or theoretical advancements. Moreover, these recommendations are only intended to inform, not determine, stakeholders' actions. Of course, this is standard practice when interpreting research findings. Still, researchers can aid such processes by acknowledging their research's limitations and explicitly stating which conclusions their findings do not support. In the next section, I will do just that.

8.5 Limitations and directions for future research

In Chapter 1, I introduced limitations to the scope of my research, outlined and discussed my overarching methodological approach, and reflected on relevant ethical considerations. I will not repeat those discussions here. Instead, I will highlight some limitations in my research design and discuss aspects of my RQs that I did not address. My aim in so doing is twofold: to aid readers in interpreting the implications and policy recommendations provided in Section 8.4 and to discuss promising avenues for future research.

8.5.1 Methodological limitations and reflections

The first set of methodological limitations stems from the pragmatist stance I adopted for this thesis. While that stance allowed me to ground my research in real-world problems and move beyond evaluating existing policy options to propose new solutions, it also creates methodological challenges. According to Kaushik and Walsh (2019), these include reliance on subjective judgements, a lack of generalisability, and the potential for oversimplification. Let us consider each in turn.

As opposed to positivistic researchers, pragmatists maintain that knowledge about the world cannot be separated from acting within it. This implies that pragmatist research tends to rely on researchers' *subjective judgements*, which can introduce bias and affect the research's validity (Ramanadhan et al., 2021). Inevitably, my own values and experiences have influenced each step of this research process, from the choice of topic to the presentation of findings. Consequently, I cannot make any claims regarding the research's reproducibility. However, this limitation should be qualified. Throughout the process, I have collected new empirical data, triangulated findings from different sources, sought input from industry practitioners, and collaborated with researchers from different disciplines. Hence, my findings are not so much subjective as relational, meaning that their validity is tied not only to a particular time and place

but also to a specific historical context. My findings concerning the merits and limitations of EBA must thus be interpreted against the backdrop of contemporary societal efforts to ensure that ADMS are designed and used in ways that are ethical, legal, and technically robust.

A related limitation concerns *generalisability*. My case study provided unique observational data about how AstraZeneca conducted EBA and the challenges it faced. However, it is often difficult to determine the extent to which case study data can explain phenomena outside its specific scope (Schaefer, 2016). I readily submit that the specific conditions that shaped the governance of ADMS within AstraZeneca cannot represent the full diversity of experiences different organisations have when implementing EBA procedures. Yet some researchers have gone further, arguing that the purpose of case studies should never be to generalise but only to particularise (Stake, 1995; Thomas, 2010). I reject such reasoning. Pragmatist research has a long history of 'analytical generalisation', whereby researchers generalise by comparing data from case studies to existing theory and by positing logical or empirical connections between the samples that were studied and those that were not (Yin, 2014).[160] Following that tradition, I have used examples from my case study only to illustrate (not prove) more general points whose soundness hinges not only on the data presented throughout this thesis but also on the quality of the reasoning used to extrapolate from it.

When developing new policy responses to specific real-world problems, there is no guarantee that the best solution will be found (Prasad, 2021). There are many reasons for this. For example, it is difficult to isolate variables to establish their causal effects when studying real-world phenomena (Hacking, 1983). Moreover, in the social sciences, it is often hard to separate the normative and empirical elements of research (Habermas, 1996). Pragmatism cannot overcome these difficulties but offers two strategies to mitigate them: instrumentalism and incrementalism. The former suggests that conceptual advances are only valuable insomuch as they are useful (Dewey, 1920), the latter that concrete problems are best addressed piecemeal, allowing new ideas to be tested and honed over time (Lindblom, 1959). I have amply employed both strategies in this thesis. In Chapter 6, I stressed the instrumental value of distinguishing between different approaches to classifying ADMS; in Chapter 7, I emphasised that the blueprint for auditing LLMs should be adopted incrementally and amended depending on context-specific considerations. The same applies to the findings synthesised in this chapter.

[160] Pragmatists thus maintain that generalisations *always* include logical arguments for extending one's claims beyond the data (Steinberg, 2015).

How EBA procedures should be designed is a complex question and the urge to *oversimplify* it must be resisted. My findings should thus be viewed not as objective facts, but as instrumental knowledge stakeholders can draw on to address real-world problems incrementally.

8.5.2 Limitations in scope and directions for future research

A second set of limitations concerns aspects of my RQs that I did not explore or that my research design did not satisfactorily answer. For example, my research has not shed any new light on the lived experiences of individuals and groups who are impacted by ADMS. This is not because that task is unimportant but because my research was conducted at a different level of abstraction. To explore EBA's limitations as an ADMS governance mechanism, I focused on observing and explaining organisational processes and dynamics. That said, research focusing on structures can serve as an important precursor and enabler of research focusing on lived experiences (Frechette et al., 2020; Stephan et al., 2016). Hence, there is scope for future work to extend my research by exploring how implementing different EBA procedures impact the lives of different individuals and groups in society.

Moreover, despite drawing on multiple methods, my research has been limited in geographical reach. For example, my industry case study concerned a Swedish-British company based in the UK; Chapter 5 focused exclusively on the forthcoming European AIA; my bibliography only includes English-language sources. The main reason for this is that language barriers reduced my ability to access and assess research contributions published in non-English journals. This is a severe limitation. Chinese-language research constitutes the fastest-growing body of academic literature on ADMS (Chou, 2022). Moreover, policymakers in different jurisdictions have adopted different approaches to address the governance challenges ADMS pose (Dixon, 2022; Roberts et al., 2021). Therefore, future research could replicate parts of my research with an extended geographical scope. Specifically, it would be useful to explore how the social, economic, and institutional limitations of EBA differ between different cultural contexts and regulatory regimes.

The choice to focus on structures over lived experiences and to restrict the literature review to English-language sources were part of my initial research design. Other substantive choices, however, presented themselves only as my preliminary findings opened new avenues for inquiry. Some of these I decided to pursue. For example, SQ4, on how to demarcate the material scope of EBA, was formulated in response to a real-world challenge revealed by my research at the conceptual and descriptive levels presented in Chapters 3–4. However, resource constraints prevented me from pursuing all possible emerging extensions of my research. While

I have identified different avenues for future research in each substantive chapter, I will re-emphasise two particularly promising topics here.

First, a major bottleneck to developing feasible and effective EBA procedures is the difficulty of operationalising normative concepts like fairness and truthfulness (Jacobs & Wallach, 2021). As my case study illustrated, the lack of standardised evaluation metrics was one of the main challenges faced during AstraZeneca's EBA. The problem is rooted in construct validity, i.e., the extent to which a given metric accurately measures what it is supposed to (Smith, 2014). An important avenue for further research would thus be to develop and evaluate different methods to operationalise normative concepts in ways that are verifiable and maintain high construct validity.

Of course, various tools and methods to measure, visualise, and evaluate the performance of ADMS along different normative dimensions have been developed. However, these generally focus on evaluating computational models, not ADMS operating in applied settings. From an EBA perspective, a clear typology is required that combines (i) available metrics for assessing the performance of (ii) different types of ADMS (e.g., symbolic vs sub-symbolic systems) along (iii) different ethical dimensions (like fairness and truthfulness) for (iv) each step in their lifecycle. Computer scientists Kearns and Roth (2020) took a first step towards creating such a typology by demonstrating how multi-dimensional Pareto frontiers can be used to visualise the normative values embodied in ADMS during the model training stage. To inform the design of feasible and effective EBA procedures, other researchers could extend Kearns and Roth's model by incorporating metrics for further normative dimensions, different types of ADMS, and later stages in the design and deployment lifecycle.

Second, the feasibility and effectiveness of EBA remain constrained by the lack of an institutional ecosystem. Without clear institutional arrangements, claims that an ADMS has been audited are difficult to verify and may even exacerbate harms by contributing to a false sense of security. This observation has been made previously (Costanza-Chock et al., 2022; Engler 2021). However, to the best of my knowledge, no peer-reviewed research has yet systematically explored different institutional arrangements for structuring EBA and holding technology providers accountable for potential breaches of trust.

As I argued in Chapter 2, much can be learned from how audits are structured in other domains. To recap, audits in different contexts are conducted by private service providers, governments, industry bodies, non-governmental organisations, and intergovernmental organisations. Each of these institutional arrangements has its own affordances and constraints. For instance, the constant pressure on private service providers to innovate can be beneficial

given how rapidly new ADMS are developed. However, such providers' reliance on good relationships with the organisations they audit increases the risk of collusion (Duflo et al., 2013). Some researchers have thus called for more government involvement, including an 'FDA for algorithms' (Tutt, 2017). However, the relative merits and limitations of different institutional arrangements to conduct EBA should be evaluated systematically and – when possible – studied empirically in applied settings.

In summary, while this thesis has contributed to an improved understanding of whether and how EBA can help organisations design and deploy ADMS in ways that align with their organisational values, many important questions – like who should conduct the audits and according to which metrics ADMS should be evaluated – remain unanswered. These questions are left to be taken up by future research.

8.6 Concluding remarks

As noted in the introduction, policymakers, researchers, and social advocacy groups have all called for the design and use of ADMS to be audited for alignment with ethics principles. However, a significant discrepancy has remained between the attention EBA has received and the lack of empirical research concerning its feasibility and effectiveness as an ADMS governance mechanism. To help bridge that gap, I set out in this thesis to explore two overarching RQs:

RQ1 What are the limitations of EBA as a governance mechanism for identifying and mitigating the ethical risks posed by ADMS?

RQ2 How can EBA procedures be designed to effectively identify and mitigate the ethical risks posed by ADMS while being feasible to implement?

The answers to these RQs can be summarised as follows. First, as an ADMS governance mechanism, EBA is subject to a wide range of conceptual, technical, economic, social, and institutional limitations. While some of these limitations can be addressed by appropriate policy responses and future technological innovation, others are intrinsic. Policymakers, researchers, and auditors should therefore exercise caution and remain realistic about what EBA can be expected to achieve. Based on the findings presented in this thesis, I propose that EBA should be understood as one governance mechanism that, by contributing to procedural regularity and transparency, can help organisations not solve but continuously manage some of the ethical risks associated with ADMS.

Second, how EBA procedures are designed and implemented matters greatly. Specifically, my findings suggest that to be feasible and effective, EBA procedures should (i) be structured and transparent, (ii) assess a clearly defined material scope according to an equally clearly defined normative baseline, (iii) incorporate elements of both technology-oriented assessments of ADMS and process-oriented assessments of organisations that design and deploy ADMS, (iv) include continuous monitoring of ADMS, and (v) be conducted by independent third-party auditors.

A further message I want to stress from this concluding chapter is the need for increased collaboration and knowledge transfer between different research communities. In Chapter 2, I showed how previous research on EBA can be divided into *narrow* and *broad* approaches. The former is technology-oriented and focuses on assessing the outputs of ADMS for different input data. The latter is process-oriented and focuses on assessing the adequacy of technology providers' QMS. While both strands of research are flourishing, they seldom have dialogue with each other. This becomes problematic when – as I have concluded in this thesis – feasible and effective EBA procedures must incorporate elements of both technology and process-oriented assessments. On the upside, many tools and methods to conduct technology and process-oriented audits have already been developed. The next step in the evolution of EBA as an ADMS governance mechanism should thus be to interlink available tools and methods into structured procedures.

A final remark: the extent to which EBA procedures contribute to good governance of ADMS depends not only on how they are designed and implemented but also on the intent of different stakeholders, including technology providers and end users. An analogy borrowed from Floridi (2014) helps to illustrate this point: the best pipes may improve the flow but do not improve the quality of the water, yet water of the highest quality is wasted if the pipes are rusty or leaky. Like the pipes in the analogy, EBA is not morally good in itself, but it enables moral goodness to be realised if properly designed and combined with the right values. To conclude, EBA procedures can facilitate the delivery of ethically sound outcomes but are not per se sufficient to ensure such outcomes.

BIBLIOGRAPHY

Abebe, R., Hill, S., Vaughan, J. W., Small, P. M., & Schwartz, H. A. (2019). Using search queries to understand health information needs in Africa. *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*, *13*, 3–14. doi.org/10.48550/arxiv.1806.05740

Ada Lovelace Institute. (2021). *Public sector use of data and algorithms*. www.adalovelaceinstitute.org/our-work/programmes/public-sector-data-algorithms/

Ada Lovelace Institute. (2021). *Technical methods for regulatory inspection of algorithmic systems in social media platforms*. https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/

Adams, W. C. (2015). Conducting semi-structured interviews. In *Handbook of Practical Program Evaluation*. John Wiley & Sons, Ltd. doi.org/10.1002/9781119171386.CH19

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., … & Le, Q. V. (2020). Towards a human-like open-domain Chatbot. *ArXiv*. doi.org/10.48550/arxiv.2001.09977

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, *54*, 95–122. doi.org/10.1007/s10115-017-1116-3

Agee, J. (2009). Developing qualitative research questions: A reflective process. *International Journal of Qualitative Studies in Education*, *22*(4), 431–447. doi.org/10.1080/09518390902736512

Ahmed, A. M., & Hammarstedt, M. (2008). Discrimination in the rental housing market: A field experiment on the Internet. *Journal of Urban Economics*, *64*(2), 362–372. https://doi.org/10.1016/j.jue.2008.02.004

AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

AI HLEG. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Aiken, C. (2021). *Classifying AI systems CSET data brief*. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/classifying-ai-systems/

Aizenberg, E., & van den Hoven, J. (2020). Designing for human rights in AI. *Big Data and Society*, *7*(2). https://doi.org/10.1177/2053951720949566

Akpinar, N.-J., Nagireddy, M., Stapleton, L., Cheng, H.-F., Zhu, H., Wu, S., & Heidari, H. (2022). A sandbox tool to bias(stress)-test fairness algorithms. *ArXiv*. doi.org/10.48550/arxiv.2204.10233

Akula, R., & Garibay, I. (2021). Audit and assurance of AI algorithms: A framework to ensure ethical algorithmic practices in artificial intelligence. *International Conference on Human-Computer Interaction*, 1–12. https://doi.org/10.48550/arXiv.2107.14046

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., … Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *ArXiv*. https://doi.org/10.48550/arxiv.2204.14198

Alford, R. R. (1998). *The craft of inquiry: Theories, methods, evidence*. Oxford University Press.

Algorithmic Justice League. (2023). *Unmasking AI harms and biases*. https://www.ajl.org/

AlgorithmWatch, & Bertelsmann Stiftung. (2019). *Automating society: Taking stock of automated decision-making in the EU*. https://algorithmwatch.org/wp-content/uploads/-2019/01/Automating_Society_Report_2019.pdf

Ali, M., Sapiezynski, P., Mislove, A., Rieke, A., Bogen, M., & Korolova, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 199. https://doi.org/10.1145/3359301

Allford, L., & Carson, P. (2015). Safety practice safety, health, and environment audits with selected case histories. In *Loss Prevention Bulletin* (241). www.researchgate.net/publication/307978324

Alshammari, M., & Simpson, A. (2017). Towards a principled approach for engineering privacy by design. In *Privacy Technologies and Policy* (pp. 161–177). Springer. https://doi.org/10.1007/978-3-319-67280-9_9

Altman, S. (2023). *Planning for AGI and beyond*. OpenAI. https://openai.com/blog/planning-for-agi-and-beyond#fn1

Amaro, S. (2021). *Dutch government resigns after childcare benefits scandal*. CNBC. www.cnbc.com/2021/01/15/dutch-government-resigns-after-childcare-benefits-scandal-.html

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Ananny, M., & Karahalios, K. (2014). *Auditing algorithms from the outside: Methods and Implications a proposal for a half-day workshop*. http://mike.ananny.org/papers/ananny-KarahaliosSandvigWilson_auditingAlgorithmsFromTheOutside_2014.pdf

Aragona, B. (2022). *Algorithm audit: Why, what, and how?* (1st ed.). Routledge.

Arjoon, S. (2005). Corporate governance: An ethical perspective. *Journal of Business Ethics*, *61*(4), 343–352. https://doi.org/10.1007/s10551-005-7888-5

Armed Services Committee. (2017). *HR 2810, National Defense Authorization Act for Fiscal Year 2018*. 115th Congress. https://doi.org/H.R.2810

Arvan, M. (2018). Mental time-travel, semantic flexibility, and A.I. ethics. *AI and Society*, 1–20. https://doi.org/10.1007/s00146-018-0848-2

Ashenden, S. K. (2021). Introduction to drug discovery. *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, 1–13. https://doi.org/10.1016/B978-0-12-820045-2.00002-7

Ashenden, S. K., Deswal, S., Bulusu, K. C., Bartosik, A., & Shameer, K. (2021). Data types and resources. *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, 27–60. doi.org/10.1016/B978-0-12-820045-2.00004-0

Aspillaga, C., Carvallo, A., & Araujo, V. (2020). Stress test evaluation of transformer-based models in natural language understanding tasks. *Proceedings of the 12th Conference on Language Resources and Evaluation*, 11–16. https://doi.org/10.48550/arXiv.2002.06261

AstraZeneca. (2020). *AI governance framework*. AstraZeneca Data and AI Ethics. www.astrazeneca.com/sustainability/ethics-and-transparency/data-and-ai-ethics.html

AstraZeneca. (2020). AstraZeneca annual report & form 20-F information 2020. In *Issues in Science and Technology* (Vol. 25, Issue 4). www.astrazeneca.com/annualreport2020

AstraZeneca. (2020). *AstraZeneca data and AI ethics*. www.astrazeneca.com/sustainability/ethics-and-transparency/data-and-ai-ethics.html

AstraZeneca. (2021). *Data science & artificial intelligence: Unlocking new science insights*. https://www.astrazeneca.com/r-d/data-science-and-ai.html#UsingAI

AstraZeneca. (2021). *Our therapy areas*. www.astrazeneca.com/our-therapy-areas.html

Auer, F., & Felderer, M. (2018). Shifting quality assurance of machine learning algorithms to live systems. *Software Engineering Und Software Managemen*, 211–212. https://dl.gi.de/bitstream/handle/20.500.12116/21162/B1-64.pdf

Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., Anderljung, M., … & Zilberman, N. (2021). Filling gaps in trustworthy development of AI. *Science*, *374*(6573), 1327–1329. https://doi.org/10.1126/SCIENCE.ABI7176

Ayling, J., & Chapman, A. (2021). Putting AI ethics to work: Are the tools fit for purpose? *AI and Ethics*, *2*(3), 405–429. https://doi.org/10.1007/s43681-021-00084-x

Azaria, A. (2022). *ChatGPT usage and limitations*. HAL. https://hal.science/hal-03913837

BABL AI. (2023). *Boutique consultancy on responsible AI*. https://babl.ai/

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., … & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*. doi.org/10.48550/arXiv.2204.05862

Baird, A., Hantke, S., & Schuller, B. (2019). Responsible and representative multimodal data acquisition and analysis: On auditability, benchmarking, confidence, data-reliance & explainability. *ArXiv*. https://doi.org/10.48550/arXiv.1903.07171

Balas, V. E., Kumar, R., & Srivastava, R. (2020). *Recent trends and advances in artificial intelligence and Internet of Things*. Springer.

Baldassarri, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology*, *43*, 41–73. doi.org/10.1146/ANNUREV-SOC-073014-112445

Baldwin, R., & Cave, M. (1999). *Understanding regulation: Theory, strategy, and practice*. Oxford University Press.

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on Human-Computer Interaction*, *5*(1), 1–34. https://doi.org/10.1145/3449148

Bandy, J., & Diakopoulos, N. (2019). Auditing news curation systems: A case study examining algorithmic and editorial logic in Apple News. *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, *2020*, 36–47.

Barfield, W., & Pagallo, U. (2018). *Research handbook on the law of artificial intelligence*. Edward Elgar Publishing.

Barocas, S., & Selbst, A. D. (2016). Big Data's disparate impact. *California Law Review*, *104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., … & Herrera, F. (2020). Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. doi.org/10.1016/j.inffus.2019.12.012

Barrett, C. (2019). Are the EU GDPR and the California CCPA becoming the De Facto Global Standards for Data Privacy and Protection? *SciTech Lawyer*, *15*(3), 24–29. https://www.proquest.com/docview/2199825726

Bartley, N., Abeliuk, A., Ferrara, E., & Lerman, K. (2021). Auditing algorithmic bias on Twitter. *ACM International Conference Proceeding Series*, 65–73. https://doi.org/10.1145/3447535.3462491

Bartosch, U., Bauberger, S., … & Rehbein, M. (2018). Policy paper on the asilomar principles on artificial intelligence. *Vereinigung Deutscher Wissenschaftler Research Technology Assessment of Digitisation*. www.researchgate.net/publication/329963051

Bass, J. M., Lero, S. B., & Noll, J. (2018). Experience of industry case studies: A comparison of multi-case and embedded case study methods. *Proceedings of the International Workshop on Conducting Empirical Studies in Industry*, 13–20. doi.org/10.1145/3193965.3193967

Bauer, J. (2016). Necessity of auditing artificial intelligence. *SSRN Electronic Journal*, *577*, 1–16. https://doi.org/10.2139/ssrn.3218675

Baum, S. D. (2017). Social choice ethics in artificial intelligence. *AI and Society*, 1–12. https://doi.org/10.1007/s00146-017-0760-1

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, *1*(1), 20–23. https://doi.org/10.1038/s42256-018-0004-1

Bellamy, R. K. E., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., … & Mehta, S. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 1–4. doi.org/10.1147/JRD.2019.2942287

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. In *Transactions of the Association for Computational Linguistics* (Vol. 6, pp. 587–604). MIT Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. doi.org/10.1145/3442188.3445922

BenevolentAI. (2019). *AstraZeneca starts artificial intelligence collaboration to accelerate drug discovery*. https://www.benevolent.com/news/astrazeneca-starts-artificial-intelligence-collaboration-to-accelerate-drug-discovery

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code* (1st ed.). Polity.

Bennett, D. (2002). Health and safety management systems: Liability or asset? *Journal of Public Health Policy*, *23*(2), 153–171. https://doi.org/10.2307/3343192/METRICS

Berghout, E., Fijneman, R., Hendriks, L., de Boer, M., & Butijn, B.-J. (2023). *Advanced digital auditing*. Springer Nature.

Berlin, I. (1988). The pursuit of the ideal. In H. Hardy (Ed.), *The crooked timber of mankind: Chapters in the history of ideas* (pp. 1–20). Princeton University Press.

Berlin, I. (1997). *The proper study of mankind: An anthology of essays*. Chatto & Windus.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal: A field experiment on labor market discrimination. *The American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Bertuzzi, L. (2021). *EU Council presidency pitches significant changes to AI Act proposal*. Euractiv. https://www.euractiv.com/section/digital/news/eu-council-presidency-pitches-significant-changes-to-ai-act-proposal/

Bertuzzi, L. (2023). *EU lawmakers set to settle on OECD definition for Artificial Intelligence*. Euractiv. https://www.euractiv.com/section/artificial-intelligence/news/eu-lawmakers-set-to-settle-on-oecd-definition-for-artificial-intelligence/

Bertuzzi, L. (2023). *AI Act: EU Parliament's crunch time on high-risk categorisation, prohibited practices*. Efficacité et Transparence Des Acteurs Européens. https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-parliaments-crunch-time-on-high-risk-categorisation-prohibited-practices/

Besse, P., Castets-Renard, C., … & Loubes, J.-M. (2018). Can everyday AI be ethical? *Fairness of Machine Learning Algorithms*. doi.org/10.13140/RG.2.2.22973.31207

Bharadhwaj, H., Huang, D.-A., Xiao, C., Anandkumar, A., & Garg, A. (2021). Auditing AI models for verified deployment under semantic specifications. *ArXiv*. doi.org/10.48550/arXiv.2109.12456

Bickman, L., & Rog, D. J. (2008). *The SAGE handbook of applied social research methods* (2nd ed.). SAGE Publications.

Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy and Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Binns, R. (2018). What political philosophy can teach us about Algorithmic Fairness? *IEEE Security & Privacy*, *6*(3), 73–80. https://doi.org/10.1109/MSP.2018.2701147

Binns, R. (2020). Algorithmic decision-making: A guide for lawyers. *Judicial Review*, *25*(1), 2–7. https://doi.org/10.1080/10854681.2020.1732739

Black, J., & Baldwin, R. (2012). When risk-based regulation aims low: Approaches and challenges. *Regulation & Governance*, *6*(1), 2–22. https://doi.org/10.1111/j.1748-5991.2011.01124.x

Blocki, J., Christin, N., Datta, A., Procaccia, A. D., & Sinha, A. (2013). Audit Games. *ArXiv*. https://arxiv.org/abs/1303.0356

Blodgett, S. L., Barocas, S., III, H. D., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/V1/2020.ACL-MAIN.485

BMW Group. (2020). *Seven principles for AI: BMW Group sets out code of ethics for the use of artificial intelligence*. www.press.bmwgroup.com/global/article/detail/T0318411EN

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer.

Boddington, P., Millican, P., & Wooldridge, M. (2017). Minds and machines special issue: Ethics and artificial intelligence. *Minds and Machines, 27,* 569–574. doi.org/10.1007/s11023-017-9449-y

Bommasani, R., & Liang, P. (2021). *Reflections on Foundation Models*. Human-Centered Artificial Intelligence. https://hai.stanford.edu/news/reflections-foundation-models

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., … Liang, P. (2021). On the opportunities and risks of foundation models. *ArXiv*. doi.org/10.48550/arXiv.2108.07258

Bommasani, R., Philosophy, K. A. C., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Advances in Neural Information Processing Systems*, *35*, 3663–3678.

Borji, A. (2023). A categorical archive of ChatGPT failures. *ArXIv*. doi.org/10.48550/arXiv.2302.03494

Borradaile, G., Burkhardt, B., & Leclerc, A. (2020). Whose Tweets are surveilled for the police: An audit of a social-media monitoring tool via log Files. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 570–580. doi.org/10.1145/3351095.3372841

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies* (1st ed.). Oxford University Press.

Bradford, A. (2012). The Brussels effect. *School of Law Printed in USA. Northwestern University Law Review*, *107*(1), 1–68. http://ssrn.com/abstract=2770634

Bradford, A. (2020). *The Brussels effect*. Oxford University Press.

Bradford, L., Aboy, M., & Liddell, K. (2020). COVID-19 contact tracing apps: A stress test for privacy, the GDPR, and data protection regimes. *Journal of Law and the Biosciences*, *7*(1), 1–21. https://doi.org/10.1093/jlb/lsaa034

British Safety Council. (2023). *About the British Safety Council*. https://www.britsafe.org/about-us/introducing-the-british-safety-council/about-the-british-safety-council/

Brown, R. G. (1962). Changing audit objectives and techniques. *The Accounting Review*, *37*(4), 696–703. https://www.proquest.com/docview/1301318804

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, *8*, doi.org/10.1177/2053951720983865

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., … Amodei, D. (2020). Language models are few-shot learners. *34th Conference on Neural Information Processing Systems*. doi.org/10.48550/arxiv.2005.14165

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., … Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *ArXiv*. http://arxiv.org/abs/2004.07213

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.

Bryson, J. (2022). *Europe is in danger of using the wrong definition of AI*. WIRED. https://www.wired.com/story/artificial-intelligence-regulation-european-union/

Bryson, J. J. (2021). The artificial intelligence of the ethics of artificial intelligence. In M. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Bryson, J., & Winfield, B. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, *50*(5), 116–119. https://doi.org/10.1109/MC.2017.154

Buiten, M. C. (2019). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation*, *10*(1), 41–59. https://doi.org/10.1017/err.2019.8

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability, and Transparency*, *1*, 1–15. https://doi.org/10.2147/OTT.S126905

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1). https://doi.org/10.1177/2053951715622512

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, *81*(5), 825–836. https://doi.org/10.1111/puar.13293

Cabrera, Á. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (2019). FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning. *2019 IEEE Conference on Visual Analytics Science and Technology*, 46–56. doi.org/10.1109/VAST47406.2019.8986948

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Calvo, R. A., Peters, D., & Cave, S. (2020). Advancing impact assessment for intelligent systems. *Nature Machine Intelligence*, *2*(2), 89–91. doi.org/10.1038/s42256-020-0151-z

Cancian, F. M. (1993). Conflicts between activist research and academic success: Participatory research and alternative strategies. *The American Sociologist*, *24*(1), 92–106. https://doi.org/10.1007/BF02691947

Carlini, N., Brain, G., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. *USENIX Security Symposium*, *267*. https://www.usenix.org/system/files/sec19-carlini.pdf

Carlini, N., Tramèr, F., Lee, K., Roberts, A., Wallace, E., Jagielski, M., … & Raffel, C. (2021). Extracting training data from large language models. *Proceedings of the 30th USENIX Security Symposium*,

Carnap, R. (1950). *Logical foundations of probability* (Vol. 2). University of Chicago Press.

Carpenter, D. (2014). *Reputation and power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton University Press. doi.org/10.5860/choice.48-3548

Cartwright, N., & Montuschi, E. (2014). *Philosophy of social science: A new introduction*. Oxford University Press.

Cath, C., Cowls, J., Taddeo, M., & Floridi, L. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133). doi.org/10.1098/rsta.2018.0080

Cave, J., Marsden, C., & Simmons, S. (2008). *Options for and Effectiveness of Internet Self- and Co-Regulation*. https://www.rand.org/pubs/technical_reports/TR566.html

Centre for Data Ethics and Innovation. (2021a). *The European Commission's Artificial Intelligence Act highlights the need for an effective AI assurance ecosystem*. https://cdei.blog.gov.uk/2021/05/11/the-european-commissions-artificial-intelligence-act-highlights-the-need-for-an-effective-ai-assurance-ecosystem/

Centre for Data Ethics and Innovation. (2021b). *The need for effective AI assurance*. https://cdei.blog.gov.uk/2021/04/15/the-need-for-effective-ai-assurance/

Centre for Data Ethics and Innovation. (2021c). *The roadmap to an effective AI assurance ecosystem - extended version*. www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem-extended-version

Chamon, M. (2016). *EU Agencies: Legal and political limits to the transformation of the EU administration*. Oxford University Press.

Chasalow, K., & Levy, K. (2021). Representativeness in Statistics, statistics, politics, and machine learning. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 77–89. https://doi.org/10.48550/arxiv.2101.03827

Chen, M., & Golan, A. (2016). What may visualization processes optimize? *IEEE Transactions on Visualization and Computer Graphics*, *22*(12), 2619–2632. doi.org/10.1109/TVCG.2015.2513410

Chen, M., Tworek, J., Jun, H., Yuan, Q., Ponde de Oliveira Pinto, H., Kaplan, J., Edwards, H., Burda, Y., … Zaremba, W. (2021). Evaluating large language models trained on code. In *arXIv*. doi.org/10.48550/arXiv.2107.03374

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). Transformers: "The end of history" for natural language processing? *Machine Learning and Knowledge Discovery in Databases.* 677–693. doi.org/10.1007/978-3-030-86523-8_41/TABLES/5

Chiou, J., Magazzini, L., Pammolli, F., & Riccaboni, M. (2012). *The value of failure in pharmaceutical R&D* (No. 1, Issue 1). www.researchgate.net/publication/254421030

Chopra, A. K., & Singh, M. P. (2018). Sociotechnical systems and ethics in the large. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 48–53. https://doi.org/10.1145/3278721.3278740

Chou, D. (2022). Counting AI research: Exploring AI research output in English- and Chinese-language sources. In *Center for Security and Emerging Technology*. https://cset.georgetown.edu/publication/counting-ai-research/

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., … Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. *ArXiv*. https://doi.org/10.48550/arxiv.2204.02311

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W.W. Norton & Company Ltd.

Christiano OpenAI, P. F., Leike DeepMind, J., Brown Google Brain, T. B., Martic DeepMind, M., Legg DeepMind, S., & Amodei OpenAI, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*. https://doi.org/10.48550/arXiv.1706.03741

Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. *Theory and Society*, *49*(5–6), 897–918. https://doi.org/10.1007/S11186-020-09411-3/METRICS

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., & Baum, S. D. (2021). AI certification: Advancing ethical practice by reducing information asymmetries. *IEEE Transactions on Technology and Society*, *2*(4), 200–209. https://doi.org/10.1109/tts.2021.3077595

Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate governance of artificial intelligence in the public interest. *Information*, *12*(7), 1–30. https://doi.org/10.3390/info12070275

Citron, D., & Pasquale. (2014). The scored society: Due process for automated predictions. *HeinOnline*, *89*(1), 1–34. https://papers.ssrn.com/=2376209

Clavell, G. G., Zamorano, M. M. n., Castillo, C., Smith, O., & Matic, A. (2020). Auditing algorithms: On lessons learned and the risks of data minimization. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 265–271. https://doi.org/10.1145/3375627.3375852

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable automated decision-making: A framework for accountable algorithmic systems. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598–609. https://doi.org/10.1145/3442188.3445921

Coeckelbergh, M. (2020). Artificial Intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Commission nationale de l'informatique et des libertés. (2019). *Privacy impact assessment - Methodology*. https://www.cnil.fr/en/privacy-impact-assessment-pia

Conrad, C. A. (2018). A philosophical and behavioral approach. In *Business ethics* (pp. 171–184). Springer.

Contractor, D., McDuff, D., Haines, J. K., Lee, J., Hines, C., Hecht, B., Vincent, N., & Li, H. (2022). Behavioral use licensing for responsible AI. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 778–788. https://doi.org/10.1145/3531146.3533143

Cookson, C. (2018). *Artificial intelligence faces public backlash, warns scientist*. Financial Times. https://www.ft.com/content/0b301152-b0f8-11e8-99ca-68cf89602132

Corea, F. (2019). AI knowledge map: How to classify AI technologies. In *An Introduction to Data* (pp. 25–29). Springer. https://doi.org/10.1007/978-3-030-04468-8_4

Corning, P. A. (2010). The re-emergence of emergence, and the causal role of synergy in emergent evolution. *Synthese*, *185*(2), 295–317. doi.org/10.1007/s11229-010-9726-2

Cropanzano, R. (2009). Writing nonempirical articles for journal of management: General thoughts and suggestions. *Journal of Management*, *35*(6), 1304–1311. https://doi.org/10.1177/0149206309344118/ASSET/0149206309344118.FP.PNG_V03

Corrigan, C. C. (2022). *Lessons Learned from Co-governance Approaches – Developing Effective AI Policy in Europe*. 25–46. https://doi.org/10.1007/978-3-031-09846-8_3

Cosserat, G. W. (2004). *Modern auditing* (2nd ed.). John Wiley & Sons, Ltd.

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *2022 ACM Conference on Fairness, Accountability, and Transparency*, *22*, 1571–1583. https://doi.org/10.1145/3531146.3533213

Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social group identity and perceptions of online hate. *Sociological Inquiry*, *89*, 427–452. doi.org/10.1111/SOIN.12274

Coston, A., Guha, N., Ouyang, D., Lu, L., Chouldechova, A., & Ho, D. E. (2021). Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for COVID-19 Policy. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 173–184. https://doi.org/10.1145/3442188.3445881

Council of Europe. (2018). *Algorithms and human rights*. www.coe.int/freedomofexpression

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Credo AI. (2023). *The leader in responsible AI*. https://www.credo.ai/

Creswell, J., & Clark, V. (2011). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Crisan, A., Drouhard, M., Vig, J., & Rajani, N. (2022). Interactive model cards: A human-centered approach to model documentation. *ACM International Conference Proceeding Series*, *22*, 427–439. https://doi.org/10.1145/3531146.3533108

Crowe, D. (2020). *Modelling biomedical data for a drug discovery knowledge graph*. Towards Data Science. https://towardsdatascience.com/modelling-biomedical-data-for-a-drug-discovery-knowledge-graph-a709be653168

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., & Sheikh, A. (2011). The case study approach. *BMC Medical Research Methodology*, *11*(1), 1–9. https://doi.org/10.1186/1471-2288-11-100/TABLES/9

Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., Darzi, A., … Rowley, S. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine*, *26*(9), 1351–1363. doi.org/10.1038/s41591-020-1037-7

Cugueró-Escofet, N., & Rosanas, J. M. (2017). The ethics of metrics: Overcoming the dysfunctional effects of performance measurements through justice. *Journal of Business Ethics*, *140*(4), 615–631. https://doi.org/10.1007/S10551-016-3049-2/TABLES/2

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In *Decision making in aviation* (pp. 289–294). Routledge. doi.org/10.4324/9781315095080-17

Currie, N. (2019). Risk based approaches to artificial intelligence. In *Crowe Data Management*. https://www.crowe.com/-/media/Crowe/LLP/folio-pdf/Risk-Approaches-to-AI.pdf

Curry, D. (2023). *ChatGPT revenue and usage statistics*. Business of Apps. https://www.businessofapps.com/data/chatgpt-statistics/

Cutler, A., Pribić, M., & Humphrey, L. (2018). *Everyday ethics for artificial intelligence*. IBM Design for AI. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

D'Agostino, M., & Durante, M. (2018). Introduction: The governance of algorithms. *Philosophy and Technology*, *31*(4), 499–505. https://doi.org/10.1007/s13347-018-0337-z

Dafoe, A. (2015). On Technological Determinism: A Typology, Scope Conditions, and a Mechanism. *Science Technology and Human Values*, *40*(6), 1047–1076. https://doi.org/10.1177/0162243915579283

Dafoe, A. (2017). AI Governance: A research agenda. *American Journal of Psychiatry*, 1–53. https://doi.org/10.1176/ajp.134.8.aj1348938

Dai, W., & Berleant, D. (2019). Benchmarking contemporary deep learning hardware and frameworks: A survey of qualitative metrics. *Proceedings - 2019 IEEE 1st International Conference on Cognitive Machine Intelligence, CogMI 2019*, 148–155. https://doi.org/10.1109/COGMI48466.2019.00029

Danaher, J. (2012). Is technology value-neutral? New technologies and collective action problems. *Techné: Research in Philosophy and Technology*, *26*(1), 31–56. https://doi.org/10.5840/techne2022524159

Danks, D. (2022). Governance via explainability. In *The Oxford Handbook of AI Governance*. Oxford University Press.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence AI and Autonomy Track*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654

Danks, D., & London, A. J. (2017). Regulating autonomous systems: Beyond standards. *IEEE Intelligent Systems*, *32*(1), 88–91. https://doi.org/10.1109/MIS.2017.1

Darke, P., Shanks, G., & Broadbent, M. (1998). Successfully completing case study research: combining rigour, relevance and pragmatism. *Information Systems Journal*, *8*(4), 273–289. https://doi.org/10.1046/J.1365-2575.1998.00040.X

Dash, A., Chakraborty, A., Ghosh, S., Mukherjee, A., & Gummadi, K. P. (2021). When the umpire is also a player: Bias in private label product recommendations on E-commerce marketplaces. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 873–884. https://doi.org/10.1145/3442188.3445944

Dash, A., Mukherjee, A., & Ghosh, S. (2019). A network-centric framework for auditing recommendation systems. *IEEE INFOCOM 2019-IEEE Conference on Computer Communications, April*, 1990–1998. https://doi.org/10.1109/INFOCOM.2019.8737486

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Daten Ethik Kommission. (2018). *Recommendations of the data ethics Commission for the federal government's strategy on artificial intelligence*. BMJ. www.bmi.bund.de/SharedDocs/downloads/EN/themen/it-digital-policy/recommendations-data-ethics-commission.pdf?__blob=publicationFile&v=3

Daten Ethik Kommssion. (2018). *Opinion of the data ethics commission*. BMJ. https://www.bmi.bund.de/SharedDocs/downloads/EN/themen/it-digital-policy/datenethikkommission-abschlussgutachtenkurz.pdf?__blob=publicationFile&v=2

Dawson, M., Burrell, D. N., Rahim, E., & Brewster, S. (2010). Integrating software assurance into the software development life cycle (SDLC) meeting department of defense (DOD) demands. *Journal of Information Systems Technology and Planning*, *3*(6), 49–53. www.academia.edu/22484322

de Laat, P. B. (2021). Companies committed to responsible AI: From principles towards implementation and regulation? *Philosophy and Technology*, *34*(4), 1135–1193. https://doi.org/10.1007/s13347-021-00474-3

De Moor, P., & De Beelde, I. (2005). Environmental auditing and the role of the accountancy profession: A literature review. *Environmental Management*, *36*(2), 205–219. https://doi.org/10.1007/s00267-004-0142-6

Dean, S., Gilbert, T. K., Lambert, N., & Zick, T. (2021). Axes for Sociotechnical Inquiry in AI Research. *IEEE Transactions on Technology and Society*, *2*(2), 62–70. doi.org/10.1109/tts.2021.3074097

Dechert. (2021). *European Commission proposes regulation on artificial intelligence*. News & Insights. https://www.dechert.com/knowledge/onpoint/2021/5/european-commission-proposes-regulation-on-artificial-intelligen.html

Delobelle, P., Tokpo, E. K., Calders, T., & Berendt, B. (2022). Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. *The 2022 Conference of the North American Chapter of the Association for Computational Linguistics,* 1693–1706. doi.org/10.18653/V1/2022.NAACL-MAIN.122

Deloitte. (2020). *Deloitte introduces trustworthy ai framework to guide Organizations in ethical application of technology*. https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-introduces-trustworthy-ai-framework.html

Dennis, L. A., Fisher, M., Lincoln, N. K., Lisitsa, A., & Veres, S. M. (2016). Practical verification of decision-making in agent-based autonomous systems. *Automated Software Engineering*, *23*(3), 305–359. https://doi.org/10.1007/s10515-014-0168-9

Denzin, N. K., & Lincoln, Y. S. (2018). *The SAGE handbook of qualitative research* (5th ed.). SAGE Publications.

Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., … & Mohammad, S. (2023). Assessing language model deployment with risk cards. *ArXiv*. doi.org/10.48550/arXiv.2303.18190

Devos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. (2022). Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3517441

DeVries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does object recognition work for everyone? *ArXiv*. https://arxiv.org/abs/1906.02659v2

Dewey, J. (1920). *Reconstruction in philosophy*. Beacon Press.

Dewey, J. (1922). *Human nature and conduct : an introduction to social psychology*. Allen & Unwin.

Di Maio, P. (2014). Towards a metamodel to support the joint optimization of socio technical systems. *Systems*, *2*(3), 273–296. https://doi.org/10.3390/systems2030273

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411

Diakopoulos, N. (2021). Transparency. In M. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Diamandis, P., & Kotler, S. (2012). *Abundance: The future is better than you think*. Free Press.

Dignum, V. (2017). Responsible autonomy. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, *1*, 5. https://doi.org/10.24963/ijcai.2017/655

Dignum, V. (2020). Responsibility and Artificial Intelligence. In *The Oxford Handbook of Ethics of AI* (Issue November, pp. 213–231). Oxford Handbooks.

Dillon, M., & Griffith, C. J. (2001). *Auditing in the food industry: From safety and quality to environmental and other audits*. CRC Press.

Dixon, R. B. L. (2022). A principled governance for emerging AI regimes: lessons from China, the European Union, and the United States. *AI and Ethics*, 1–18. https://doi.org/10.1007/S43681-022-00205-0

Dobbe, R. I. J. (2022). System safety and artificial intelligence. In *The Oxford Handbook of AI Governance* (p. C67.S1-C67.S18). Oxford University Press.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large WebText corpora: A case study on the colossal clean crawled corpus. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. https://doi.org/10.48550/arXiv.2104.08758

Drudi, D. (2015). The quest for meaningful and accurate occupational health and safety statistics. *Monthly Labor Review*, *2015*(12). https://doi.org/10.21916/MLR.2015.53

Du Sautoy, M. (2019). *The creativity code: Art and innovation in the age of AI*. Belknap Press.

Duflo, E., Greenstone, M., Pande, R., & Ryan, N. (2013). Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India. *The Quarterly Journal of Economics*, *128*(4), 1499–1545. doi.org/10.1093/QJE/QJT024

Dunleavy, P., & Margetts, H. (2015). *Design principles for essentially digital governance*. SCRIBD.

Dunn, M., & Hope, R. A. (2018). *Medical ethics: A very short introduction* (2nd ed.). Oxford University Press.

Durante, M., & Floridi, L. (2022). A legal principles-based framework for AI liability regulation. In *The 2021 Yearbook of the Digital Ethics Lab* (pp. 93–112). Springer International Publishing.

Dwork, C. (2006). Differential privacy. *Automata, Languages and Programming: 33rd International Colloquium*, *5*, 1–19. https://doi.org/10.1007/978-3-540-79228-4_1

Economist Intelligence Unit. (2020). *Staying ahead of the curve – The business case for responsible AI*. https://www.eiu.com/n/staying-ahead-of-the-curve-the-business-case-for-responsible-ai/

ECP. (2018). *Artificial intelligence impact assessment*. https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf

Edwards, L., & Veale, M. (2018). Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"? *EEE Security & Privacy, 16*(3), 46–54. doi.org/10.1109/MSP.2018.2701152

Edwards, R., & Holland, J. (2013). *What is qualitative interviewing?* Bloomsbury Academic Publishing.

Ellemers, N., van der Toorn, J., Paunov, Y., & van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366. doi.org/10.1177/1088868318811759

Engler, A. (2023). *Early thoughts on regulating generative AI like ChatGPT*. Brookings TechTank. https://www.brookings.edu/blog/techtank/2023/02/21/early-thoughts-on-regulating-generative-ai-like-chatgpt/

Engler, A. C. (2021). *Outside auditors are struggling to hold AI companies accountable*. FastCompany. https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue

Epstein, Z., Payne, B. H., Shen, J. H., Hong, C. J., Felbo, B., Dubey, A., Groh, M., Obradovich, N., Cebrian, M., & Rahwan, I. (2018). Turingbox: An experimental platform for the evaluation of AI systems. *IJCAI International Joint Conference on Artificial Intelligence*, 5826–5828. https://doi.org/10.24963/ijcai.2018/851

Erdelyi, O. J., & Goldsmith, J. (2018). Regulating artificial intelligence proposal for a global solution. *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf

Ernst & Young LLP. (2018). *Assurance in the age of AI*. https://assets.ey.com/content/dam-/ey-sites/ey-com/en_gl/topics/digital/ey-assurance-in-the-age-of-ai.pdf

Esposito, E. (2022). *Artificial communication*. MIT Press.

Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology, 18*(2), 149–156. https://doi.org/10.1007/s10676-016-9400-6

Eubanks, V. (2019). *Automating inequality: How high tools profile, police, and punish the poor* (1st ed.). St. Martin's Press.

European Commission. (2019). *Communication: Building trust in human centric artificial intelligence*. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

European Commission. (2020a). *Report on current policy measures and policy opportunities Artificial intelligence-critical industrial applications*. https://doi.org/10.2826/47005

European Commission. (2020b). *White Paper On Artificial Intelligence-A European approach to excellence and trust*. 27. https://commission.europa.eu/system/files/2020-02/-commission -white-paper-artificial-intelligence-feb2020_en.pdf

European Commission. (2021a). Artificial Intelligence Act. *Proposal for Regulation of the European Parliament and of the Council - Laying down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts*. https://eur-lex.europa.eu/-legal-content/EN/TXT/?uri=celex%3A52021PC0206

European Commission. (2021b). *ANNEXES to the proposal for a regulation of the European Parliament and of the Council*. www.eur-lex.europa.eu/legalcontent/EN/TXT/?uri=-celex%3A52021PC0656

European Commission. (2021c). *Commission staff working document: Impact assessment accompanying the proposal for a regulation of the European Parliament and of the Council (artificial intelligence ACT) and amending certain union legislative acts*. https://artificialintelligenceact.eu/wp-content/uploads/2022/06/AIA-COM-Impact-Assessment-1-21-April.pdf

European Commission. (2022). AI liability directive. In *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence*. https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf

European Data Protection Board. (2021). *EDPB/EDPS Joint Opinion 5/2001 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artifiticial intelligence (Artificial Intelligence Act)*. http//edpb.-europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf

European Data Protection Supervisor. (2023). *The History of the General Data Protection Regulation* |. European Data Protection Supervisor. https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en

European Law Institute. (2022). *Guiding principles for automated decision-making in the EU*. www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_-Innovation_Paper_on_Guiding_Principles_for_ADM_in_the_EU.pdf

European Parliament. (2001). Directive 2001/95/EC of the European Parliament and of the Council on general product safety. *Official Journal of the European Communities*, *L11/4*(7), 4–17. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:-2002:011:0004:0017:en:PDF

European Parliament. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council. In *Official Journal of the European Union*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

European Parliament. (2021). *Resolution on the Commission evaluation report on the implementation of the General Data Protection Regulation two years after its application*. www.europarl.europa.eu/doceo/document/TA-9-2021-0111_EN.html

European Parliament. (2022). Digital Services Act. *REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act)*, 1–222.

European Parliamentary Research Service. (2019). *A governance framework for algorithmic accountability and transparency*. https://data.europa.eu/doi/10.2861/59990

European Parliamentary Research Service. (2022). *Auditing the quality of datasets used in algorithmic decision-making systems*. www.europarl.europa.eu/regdata/etudes-/stud/-2022/729541/eprs_stu(2022)729541_en.pdf

European Union. (2020). *EU Regulation on European data governance* (Vol. 0340). https://eur-lex.europa.eu/resource.html?uri=cellar%3A91ce5c0f-12b6-11eb-9a54-

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., Righetti, L., & Saunders, W. (2021). Truthful AI: Developing and governing AI that does not lie. *ArXiv*. https://doi.org/10.48550/arXiv.2110.06674

Fabbri, D., & LeFevre, K. (2011). Explanation based auditing. *Proceedings of the VLDB Endowment*, *5*(1), 1–12. https://doi.org/10.14778/2047485.2047486

Fagerholm, F., Guinea, A. S., Mäenpää, H., & Münch, J. (2014). Building blocks for continuous experimentation. *Proceedings of the 1st International Workshop on Rapid Continuous Software Engineering*, 26–35. https://doi.org/10.1145/2593812.2593816

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., … & Yeong, Z. K. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence 2021 3:7*, *3*(7), 566–571. doi.org/10.1038/s42256-021-00370-7

Falkenberg, L., & Herremans, I. (1995). Ethical behaviours in organizations: Directed by the formal or informal systems? *Journal of Business Ethics*, *14*(2), 133–143. https://doi.org/10.1007/BF00872018

Farber, H. S., Silverman, D., & Wachter, T. M. Von. (2017). Factors determining callbacks to job applications by the unemployed: An audit study. *Russell Sage Foundation Journal of the Social Sciences*, *3*(3), 168–201. https://doi.org/10.7758/rsf.2017.3.3.08

Faught, A. M., Davidson, S. E., Fontenot, J., … & Followill, D. S. (2017). Development of a Monte Carlo multiple source model for inclusion in a dose calculation auditing tool: *Medical Physics*, *44*(9), 4943–4951. https://doi.org/10.1002/mp.12426

Feigenbaum, E. A., & Feldman, J. (1963). *Computers and thought*. McGraw-Hill.

Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., & Heintz, F. (2022). Achieving a data-driven risk assessment methodology for ethical AI. *Digital Society*, *1*(2), 13. https://doi.org/10.1007/s44206-022-00016-0

Ferretti, T. (2021). An institutionalist approach to AI ethics: Justifying the priority of government regulation over self-regulation. *Moral Philosophy and Politics*, *9*(2), 239–265. https://doi.org/10.1515/mopp-2020-0056

Fitzgerald, B., Stol, K. J., O'Sullivan, R., & O'Brien, D. (2013). Scaling agile methods to regulated environments: An industry case study. *Proceedings - International Conference on Software Engineering*, 863–872. https://doi.org/10.1109/ICSE.2013.6606635

Fjeld, jessica. (2020). Principled artificial intelligence. *IEEE Instrumentation and Measurement Magazine*, *23*(3), 27–31. https://doi.org/10.1109/MIM.2020.9082795

Flint, D. (1988). *Philosophy and principles of auditing: An introduction*. Macmillan Education.

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines*, *18*(3), 303–329. https://doi.org/10.1007/s11023-008-9113-7

Floridi, L. (2013). Distributed morality in an information society. *Science and Engineering Ethics*, *19*(3), 727–743. https://doi.org/10.1007/s11948-012-9413-4

Floridi, L. (2014). *The 4th revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Floridi, L. (2016a). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083). https://doi.org/10.1098/rsta.2016.0112

Floridi, L. (2016b). Tolerant paternalism: Pro-ethical design as a resolution of the dilemma of toleration. *Science and Engineering Ethics*, *22*, 1669–1688. doi.org/10.1007/s11948-015-9733-2

Floridi, L. (2017a). A defence of constructionism: Philosophy as conceptual engineering. *Pensamiento*, *73*(276), 271–300. https://doi.org/10.14422/pen.v73.i276.y2017.003

Floridi, L. (2017b). Infraethics–on the conditions of possibility of morality. *Philosophy and Technology*, *30*(4), 391–394. https://doi.org/10.1007/s13347-017-0291-1

Floridi, L. (2017c). The logic of design as a conceptual logic of information. *Minds and Machines*, *27*(3), 495–519. https://doi.org/10.1007/s11023-017-9438-1

Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy and Technology*, *31*(1). https://doi.org/10.1007/s13347-018-0303-9

Floridi, L. (2019a). The green and the blue—Naïve ideas to improve politics in a mature information society. In *The 2018 Digital Ethics Lab Yearbook* (pp. 183–221). Springer.

Floridi, L. (2019b). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy and Technology*, *32*(2), 185–193. doi.org/10.1007/s13347-019-00354-x

Floridi, L. (2021a). The end of an era: From self-regulation to hard law for the digital industry. *Philosophy and Technology*, 619–622. doi.org/10.1007/s13347-021-00493-0

Floridi, L. (2021b). The European legislation on AI: A brief analysis of its philosophical approach. In *Digital Ethics Lab Yearbook* (pp. 1–8). Springer International Publishing.

Floridi, L. (2022). AI as Agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 15. https://doi.org/10.2139/ssrn.4358789

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. In *Minds and Machines* (Vol. 30, Issue 4, pp. 681–694). Springer.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*, 1–13. https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*, 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Floridi, L., & Strait, A. (2020). Ethical foresight analysis: What it is and why it is needed? *Minds and Machines*, *30*(1), 77–97. https://doi.org/10.1007/s11023-020-09521-y

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., … & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). capAI — A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act. *SSRN Electronic Journal*, 1–90. https://doi.org/10.2139/ssrn.4064091

Flyvbjerg, B. (2001). *Making social science matter*. Cambridge University Press.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, *12*(2), 219–245. https://doi.org/10.1177/1077800405284363

Fontanelli, F. (2016). The Court of Justice of the European Union and the illusion of balancing in internet related disputes. In O. Pollicino & G. Romeo (Eds.), *The Internet and Constitutional Law. The protection of fundamental rights and constitutional adjudication in Europe* (pp. 94–117). Routledge.

Food and Drug Administration. (2021). *Artificial intelligence and machine learning in software as a medical device*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Food and Drug Administration. (2022). *Inspection classification database*. www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/inspection-classification-database

ForHumanity. (2023). *Independent audit of AI systems*. https://forhumanity.center/independent-audit-of-ai-systems/

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Frankish, K., & Ramsey, W. M. (2014). *The Cambridge handbook of artificial intelligence*. Cambridge University Press.

Fraser, H. L., & Bello y Villarino, J.-M. (2021). Where residual risks reside: A comparative approach to Art 9(4) of the European Union's proposed AI regulation. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3960461

Frey, B. B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vol. 4). Sage Publications.

Frey, C. B. (2019). *The technology trap: Capital, labor, and power in the age of automation*. Princeton University Press.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im)possibility of fairness. *Communications of the ACM*, *64*(4), 136–143. https://doi.org/10.1145/3433949

Fukuchi, K., Hara, S., & Maehara, T. (2019). Faking fairness via stealthily biased sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(1), 412–419. https://doi.org/10.1609/aaai.v34i01.5377

Furner, J. (2006). Conceptual analysis: A method for understanding information as evidence, and evidence as information. *Archival Science*, *4*, 233–265. doi.org/10.1007/s10502-005-2594-8

Future of Life Institute. (2017). *Asilomar AI principles*. https://futureoflife.org/ai-principles/

Future of Life Institute (2022). *Developments - The Artificial Intelligence Act*. https://artificialintelligenceact.eu/developments/

Frechette, J., Bitzas, V., Aubry, M., Kilpatrick, K., & Lavoie-Tremblay, M. (2020). Capturing Lived Experience: Methodological Considerations for Interpretive Phenomenological Inquiry. *International Journal of Qualitative Methods*, *19*, 1–12. doi.org/10.1177/160940692090-7254

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gaddis, S. M. (2018). *An introduction to audit studies in the social sciences*. Springer International Publishing.

Gallo, V., Strachan, D., Bartoletti, I., Denev, A., & Lavrinenko, K. (2021). *The new EU AI Act | What do financial services firms need to know?* Deloitte Insights. https://ukfinancialservicesinsights.deloitte.com/post/102gxhz/the-new-eu-ai-act-what-do-financial-services-firms-need-to-know

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., … Clark, J. (2022a). Predictability and surprise in large generative models. *ACM International Conference Proceeding Series*, 1747–1764. https://doi.org/10.1145/3531146.3533229

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., … Clark, J. (2022b). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *ArXiv*. https://github.com/anthropics/hh-rlhf

Gao, L., & Huang, R. (2017). Detecting online hate speech using context aware models. *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 260–266. https://doi.org/10.26615/978-954-452-049-6_036

Garg, S., & Ramakrishnan, G. (2020). BAE: BERT-based adversarial examples for text classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 6174–6181. doi.org/10.18653/V1/2020.EMNLP-MAIN.498

Gasser, U., & Almeida, V. A. F. (2017). A layered model for AI governance. *IEEE Internet Computing*, *21*(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

Gasser, U., & Schmitt, C. (2021). The role of professional norms in the governance of artificial intelligence. In M. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Gay, A. S., & New, N. H. (1999). Auditing health and safety management systems: a regulator's view. *Occupational Medicine*, *49*(7), 471–473. https://doi.org/10.1093/occmed/49.7.471

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369. http://arxiv.org/abs/2009.11462

Gibson Dunn. (2023). *New York city proposes rules to clarify upcoming artificial intelligence law for employers*. https://www.gibsondunn.com/new-york-city-proposes-rules-to-clarify-upcoming-artificial-intelligence-law-for-employers/

Gilbert, S., Green, C., & Crewe, D. (2021). *Vaxxers: The inside story of the Oxford AstraZeneca vaccine and the race against the virus*. Hodder & Stoughton.

Gilson, L. L., & Goldberg, C. B. (2015). Editors' comment: So, what is a conceptual paper? *Group and Organization Management*, *40*(2), 127–130. https://doi.org/10.1177/1059601115576425

Given, L. M. (2008). *The SAGE encyclopedia of qualitative research methods*. SAGE Publications.

Goel, K., Rajani, N., Vig, J., Taschdjian, Z., Bansal, M., & Ré, C. (2021). Robustness gym: Unifying the NLP evaluation landscape. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, 42–55. https://doi.org/10.18653/V1/2021.NAACL-DEMOS.6

Goldstein, B. (2018). *A Brief Taxonomy of AI*. LinkedIn. www.linkedin.com/pulse/brief-taxonomy-ai-bernard-golstein?trk=public_profile_article_view

Goodman, B. (2016). A step towards accountable algorithms?: Algorithmic discrimination and the european union general data protection. *29th Conference on Neural Information Processing Systems (NIPS 2016),* 1–7. www.mlandthelaw.org/papers/goodman1.pdf

Goodman, B. (2021). Hard choices and hard limits in artificial intelligence. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 112–120. https://doi.org/10.1145/3461702.3462539

Google. (2018). *Artificial intelligence at Google: Our principles*. https://ai.google/principles/

Google. (2020). *What-if-tool*. https://pair-code.github.io/what-if-tool/index.html

Government of Canada. (2019). *Algorithmic impact assessment tool*. Responsible Use of Artificial Intelligence (AI). www.canada.ca/en/government/system/digital-government/-digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Government Office for Science. (2014). *Innovation: Managing risk, not avoiding it*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/381905/14-1190a-innovation-managing-risk-report.pdf

GRAIL. (2021). *GRAIL announces collaborations with Amgen, AstraZeneca, and Bristol Myers Squibb to evaluate cancer early detection technology for minimal residual disease*. www.grail.com/press-releases/grail-announces-collaborations-with-amgen-astrazeneca-and-bristol-myers-squibb-to-evaluate-cancer-early-detection-technology-for-minimal-residual-disease/

Grand View Research. (2017). *Financial auditing professional services market report, 2025*. https://www.grandviewresearch.com/industry-analysis/financial-auditing-professional-services-market

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine*, *5*(3), 101. https://doi.org/10.1016/S0899-3467(07)60142-6

Green, R. M., & Donovan, A. (2009). The methods of business ethics. In *The Oxford handbook of business ethics*. Oxford University Press.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2122–2131. https://doi.org/10.24251/hicss.2019.258

Grimes, D. A., & Schulz, K. F. (2002). Descriptive studies: What they can and cannot do. *Lancet*, *359*, 145–149. https://doi.org/10.1016/S0140-6736(02)07373-7

Grimpe, C., & Kaiser, U. (2010). Balancing internal and external knowledge acquisition: The gains and pains from R & D outsourcing. *Journal of Management Studies*, *47*(8), 1483–1509. https://doi.org/10.1111/j.1467-6486.2010.00946.x

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205–211. doi.org/10.1136/medethics-2019-105586

Gupta, A., & Mishra, M. (2023). Artificial intelligence for recruitment and selection. *The Adoption and Effect of Artificial Intelligence on Human Resources Management, Part B*, 1–11. https://doi.org/10.1108/978-1-80455-662-720230001

Gupta, K. (2004). *Comtemporary auditing*. McGraw Hill.

Gururangan, S., Marasovi´cmarasovi´c, A., … & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360. doi.org/10.18653/V1/2020.ACL-MAIN.740

Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., & Katyal, V. (2018). *Why we need to audit algorithms*. Harward Business Review. https://hbr.org/2018/11/why-we-need-to-audit-algorithms

Haas, P. J., & Springer, J. F. (1998). *Applied policy research: Concepts and cases*. Routledge.

Haataja, M., & Bryson, J. J. (2021). *What costs should we expect from the EU's AI Act?* (No. 8nzb4). Center for Open Science. https://osf.io/preprints/socarxiv/8nzb4/

Haataja, M., & Bryson, J. J. (2022). Reflections on the EU's AI act and how we could make it even better. In *Competition Policy International* (Vol. 24). https://eur-lex.europa.-eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AF-

Haberler, G. (1936). *The theory of international trade: With its applications to commercial policy*. Macmillan.

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy*. Polity.

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. *ArXiv*. https://doi.org/10.48550/arXiv.2302.02337

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

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

Hale, C. (2017). *What is activist research?* Social Science Research Council. https://items.ssrc.org/from-our-archives/what-is-activist-research/

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, … Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open, 2*, 225–250. https://doi.org/10.1016/J.AIOPEN.2021.08.002

Hancox-Li, L., & Kumar, I. E. (2021). Epistemic values in feature importance methods: Lessons from feminist epistemology. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 817–826. https://doi.org/10.1145/3442188.3445943

Hansen, J. V., & Messier, W. F. (1986). A knowledge-based expert system for auditing advanced computer systems. *European Journal of Operational Research*, *26*(3), 371–379. https://doi.org/10.1016/0377-2217(86)90139-6

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 3323–3331. https://papers.nips.cc/-paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

Haron, H., Chambers, A., Ramsi, R., & Ismail, I. (2004). The reliance of external auditors on internal auditors. *Managerial Auditing Journal*, *19*(9), 1148–1159. https://doi.org/10.1108/02686900410562795/FULL/PDF

Harste, G. (2021). *The Habermas-Luhmann debate*. Columbia University Press.

Hasan, A., Brown, S., Davidovic, J., Lange, B., & Regan, M. (2022). Algorithmic bias and risk assessments: Lessons from practice. *Digital Society, 1*(2), 14. https://doi.org/10.1007/s44206-022-00017-z

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Review, 12*(1). https://doi.org/10.14763/2023.1.1682

High-Level Expert Group on AI. (2019). *Policy and investment recommendations for trustworthy Artificial Intelligence*. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60343

Hill, K. (2020). *Twitter tells facial recognition trailblazer to stop using site's photos*. New York Tmes. https://www.nytimes.com/2020/01/22/technology/clearview-ai-twitter-letter.html?searchResultPosition=11/3https://nyti.ms/2RIvPy5

Hirschberg, J., & Manning, C. D. (2015). *Advances in natural language processing*. *349*(6245), 261–266. https://doi.org/10.1126/SCIENCE.AAA8685

Hodges, C. (2015). Ethics in business practice and regulation. In *Law and corporate behaviour integrating theories of regulation, enforcement, compliance and ethics* (pp. 1–21). Bloomsbury Publishing. https://doi.org/10.5040/9781474201124

Hoffmann, A. L., Roberts, S. T., Wolf, C. T., & Wood, S. (2018). Beyond fairness, accountability, and transparency in the ethics of algorithms: Contributions and perspectives from LIS. *Proceedings of the Association for Information Science and Technology*, *55*(1), 694–696. https://doi.org/10.1002/pra2.2018.14505501084

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de Las, Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., … Sifre, L. (2022). Training compute-optimal large language models. *ArXiv*. http://arxiv.org/abs/2203.15556

Holistic AI. (2023). *AI risk management solutions*. www.holisticai.com/ai-risk-management

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *ArXiv*, *May*. http://arxiv.org/abs/1805.03677

Holweg, M., Younger, R., & Wen, Y. (2022). *The reputational risks of AI*. California Management Review. https://cmr.berkeley.edu/2022/01/the-reputational-risks-of-ai/

Hsieh, K. (2019). *Transformer Poetry: Poetry classics reimagined by artificial intelligence*. Paper Gains Publishing.

Hu, Y., Jing, X., Ko, Y., & Rayz, J. T. (2021). Misspelling Correction with pre-trained contextual language model. *2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 144–149. https://doi.org/10.1109/ICCICC50026.2020.9450253

Hubinger, E. (2019). *Relaxed adversarial training for inner alignment*. AI Alignment Forum. https://www.alignmentforum.org/posts/9Dy5YRaoCxH9zuJqa/relaxed-adversarial-training-for-inner-alignment

Hustedt, C. (2020). From principles to practice - An interdisciplinary framework to operationalise AI ethics. In *AI Ethics Impact Group.* Bertelsmann Stifung. https://doi.org/10.11586/2020013

Ibáñez, J. C., & Olmeda, M. V. (2021). Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI and Society*, *4*. https://doi.org/10.1007/s00146-021-01267-0

Idowu, S. O. (2013). Legal Compliance. In *Encyclopedia of Corporate Social Responsibility* (p. 1578). Springer. https://doi.org/10.1007/978-3-642-28036-8_100980

IEEE Standard Association. (2019). Ethically aligned design. *Intelligent Systems, Control and Automation: Science and Engineering*, *95*, 11–16. doi.org/10.1007/978-3-030-12524-0_2

IEEE Standard Association. (2020). IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being. *IEEE 7010-2020*, *16*, 1–96. https://doi.org/10.1109/IEEESTD.2020.9084219

IEEE Standard Association. (2020). *IEEE standards dictionary online*. http://dictionary.ieee.org

IEEE Standard Association. (2023). *IEEE portfolio of AIS technology and impact standards and standards projects*. The Institute of Electrical and Electronics Engineers. https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/

Iliescu, F.-M. (2010). Auditing IT Governance. *Informatica Economica*, *14*(1), 93–102. https://www.proquest.com/docview/1433236144

Imana, B., Korolova, A., & Heidemann, J. (2023). Having your privacy cake and eating it too: Platform-supported auditing of social media algorithms for public interest. *Proceedings of the ACM on Human-Computer Interaction*, *7*, 1–33. https://doi.org/10.1145/3579610

Information Accountability Foundation. (2019). *Ethical data impact assessments and oversight models* (Issue January). https://www.immd.gov.hk/pdf/PCAReport.pdf

Information Commissioner's Office. (2018). *Guide to the General Data Protection Regulation* (GDPR). https://doi.org/10.1111/j.1751-1097.1994.tb09662.x

Information Commissioner's Office. (2020). *Guidance on the AI auditing framework: Draft guidance for consultation*. https://ico.org.uk/media/about-the-ico/consultations/-2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Information Systems Audit and Control Association. (2018). *Auditing artificial intelligence*. https://transformingaudit.isaca.org/featured-articles/auditing-artificial-intelligence

Institute of Internal Auditors. (2018). The IIA's artificial intelligence Auditing framework. In *Global Perspectives*. https://www.nist.gov/system/files/documents/2021/10/04/GPI-Artificial-Intelligence-Part-III.pdf

Institute of International Finance. (2019). *Machine learning in credit risk report*. https://www.iif.com/Portals/0/Files/content/Research/iif_mlcr_2nd_8_15_19.pdf

International Atomic Energy Agency. (2015). *Quality management audits in nuclear medicine practices.* (No. 33; IAEA Human Health Series).

International Organization for Standardization. (2015). *ISO/IEC 38500:2015 - Information technology — Governance of IT for the organization*. www.iso.org/standard/62816.html

International Organization for Standardization. (2018). *ISO 31000 - Risk management - guidelines*. https://www.iso.org/obp/ui/#iso:std:iso:31000:en

International Organization for Standardization. (2019). *It's all about trust*. ISO News. https://www.iso.org/news/ref2452.html

International Organization for Standardization. (2021). *ISO/IEC TR 24027:2021 - Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making*. https://www.iso.org/standard/77607.html

International Organization for Standardization. (2022). *ISO/IEC 38507:2022 - Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations*. https://www.iso.org/standard/56641.html?browse=tc

International Organization for Standardization. (2023). *ISO/IEC 23894 - Information technology — Artificial intelligence — Guidance on risk management*. https://www.iso.org/standard/77304.html

Islam, G., & Greenwood, M. (2021). The metrics of ethics and the ethics of metrics. *Journal of Business Ethics*, *175*, 1–5. https://doi.org/10.1007/s10551-021-05004-x

Jaakkola, E. (2020). Designing conceptual articles: four approaches. *AMS Review*, *10*(1–2), 18–26. https://doi.org/10.1007/s13162-020-00161-0

Jackall, R. (2010). *Moral mazes: The world of corporate managers* (20th ed.). Oxford University Press.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, *11*(21), 375–385. https://doi.org/10.1145/3442188.3445901

Jager, T., & Westhoek, E. (2023). Keeping control on deep learning image recognition algorithms. *Advanced Digital Auditing*, 121–148. doi.org/10.1007/978-3-031-11089-4_6

Jaiswal, S., Duggirala, K., Dash, A., & Mukherjee, A. (2022). Two-face: Adversarial audit of commercial face recognition systems. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 381-392. doi.org/10.1609/icwsm.v16i1.19300

James, W. (1907). *Pragmatism: A new name for some old ways of thinking*. Longmans, Green, & Co.

Janssen, H. (2020). An approach for a fundamental rights impact assessment to automated decision-making. *International Data Privacy Law*, *10*(1), 76–106. doi.org/10.1098/rsta.2018.0084

Jay, R., Malcolm, W., Parry, E., Townsend, L., & Bapat, A. (2018). *Guide to the General Data Protection Regulation*. Sweet & Maxwell.

Jayaraman, B., & Evans, D. (2019). Evaluating differentially private machine learning in practice. *Proceedings of the 28th USENIX Security Symposium*, 14–16. http://arxiv.org/abs/1902.08874

Jernite, Y., Nguyen, H., Biderman, S., Rogers, A., … & Mitchell, M. (2022). Data governance in the age of large-scale data-driven language technology. *Proceedings of 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2206–2222. doi.org/10.1145/3531146.3534637

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 306–316). doi.org/10.1145/3351095.3372829

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 2019 1:9*, *1*, 389–399. doi.org/10.1038/s42256-019-0088-2

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Joshi, A. K. (1991). Natural language processing. *Science*, *253*(5025), 1242–1249. https://doi.org/10.1126/SCIENCE.253.5025.1242

Jotterand, F., & Bosco, C. (2020). Keeping the "human in the loop" in the age of artificial intelligence: Accompanying commentary for "correcting the brain?". *Science and Engineering Ethics*, *26*(5), 2455–2460. doi.org/10.1007/s11948-020-00241-1

Kahneman, D. (2013). *Thinking, fast and slow* (1st ed.). Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Kak, A., & West, S. M. (2023). Confronting tech power 2023 Landscape. *AI Now Institute*. https://ainowinstitute.org/2023-landscape.

Tweede Kamer (2020). Ongekend onrecht. In *35 510 Parlementaire ondervraging Kinderopvangtoeslag*. www.tweedekamer.nl/sites/default/files/atoms/files/-20201217-_eindverslag_parlementaire_ondervragingscommissie_kinderopvangtoeslag.pdf

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., … & Amodei, D. (2020). Scaling laws for neural language models. *ArXiv*. https://doi.org/10.48550/arXiv.2001.08361

Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard: Translating strategy into action*. Harvard Business School Press.

Karan, M., & Šnajder, J. (2019). Preemptive toxic language detection in Wikipedia comments using thread-level context. *Proceedings of the Third Workshop on Abusive Language Online*, 129–134. https://doi.org/10.18653/V1/W19-3514

Karanasiou, A. P., & Pinotsis, D. A. (2017). A study into the layers of automated decision-making: Emergent normative and legal aspects of deep learning. *International Review of Law, Computers & Technology*, *31*, 170–187. doi.org/10.1080/13600869.2017.1298499

Karppinen, K., & Moe, H. (2012). What we talk about when we talk about Document Analysis. *Trends in Communication Policy Research: New Theories, Methods and Subjects*, 177–193. https://doi.org/10.1080/09528820601138683

Kasirzadeh, A., & Gabriel, I. (2023). In conversation with Artificial Intelligence: Aligning language models with human values. *Philosophy and Technology*, *36*(2), 1–24. https://doi.org/10.48550/arxiv.2209.00731

Kassir, S., Baker, L., Dolphin, J., & Polli, F. (2022). AI for hiring in context: A perspective on overcoming the unique challenges of employment research to mitigate disparate impact. *AI and Ethics*, 1–24. https://doi.org/10.1007/s43681-022-00208-x

Kaushik, S., Choudhury, A., Sheron, P. K., Dasgupta, N., Natarajan, S., Pickett, L. A., & Dutt, V. (2020). AI in healthcare: Time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in Big Data*, *3*(4). doi.org/10.3389/fdata.2020.00004

Kaushik, V., & Walsh, C. A. (2019). Pragmatism as a research paradigm and its implications for social work research. *Social Sciences 2019, Vol. 8, Page 255*, *8*(9), 255. https://doi.org/10.3390/SOCSCI8090255

Kazim, E., & Koshiyama, A. (2020). AI assurance processes. *SSRN Electronic Journal*, 1–9. https://doi.org/10.2139/ssrn.3685087

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns*, *2*(9), 1–18. https://doi.org/10.1016/j.patter.2021.100314

Kazim, E., Denny, D. M. T., & Koshiyama, A. (2021). AI auditing and impact assessment: According to the UK information commissioner's office. *AI and Ethics*, *1*, 301–310. https://doi.org/10.1007/s43681-021-00039-2

Kazim, E., Koshiyama, A. S., Hilliard, A., & Polle, R. (2021). Systematizing audit in algorithmic recruitment. *Journal of Intelligence*, *9*(3), 1–11. doi.org/10.3390/jintelligence9030046

Kearns, M. J., & Roth, A. (2020). *The ethical algorithm: The science of socially aware algorithm design*. Tantor and Blackstone Publishing.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *35th International Conference on Machine Learning, ICML 2018*, *6*, 4008–4016. https://proceedings.mlr.press/v80/kearns18a.html

Keyes, O., Durbin, M., & Hutson, J. (2019). A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. *Conference on Human Factors in Computing Systems*, 1–11. doi.org/10.1145/3290607.3310433

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., … & Williams, A. (2021). Dynabench: rethinking benchmarking in NLP. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4110–4124. https://doi.org/10.18653/V1/2021.NAACL-MAIN.324

Kim, P. (2017). Auditing algorithms for discrimination. *University of Pennsylvania Law Review*, *166*, 189–203.

Kim, T. W., Hooker, J., & Donaldson, T. (2021). Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, *70*, 871–890. https://doi.org/10.1613/JAIR.1.12481

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. *ArXiv*. https://doi.org/10.48550/arXiv.2301.10226

Kirk, H. R., Birhane, A., Vidgen, B., & Derczynski, L. (2022). Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 497--510). aclanthology.org/2022.findings-emnlp.35

Kirk, H. R., Jun, Y., Iqbal, H., Benussi, E., Volpin, F., Dreyer, F. A., Shtedritski, A., & Asano, Y. M. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in Neural Information Processing Systems*, *34*, 2611–2642. doi.org/10.48550/arXiv.2102.04130

Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2023). Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *ArXiv*. https://doi.org/10.48550/arXiv.2303.05453

Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T., & Hale, S. A. (2022). Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics.* doi.org/10.18653/V1/2022.NAACL-MAIN.97

Kleinberg, J. (2018). Inherent trade-offs in algorithmic fairness. *Bstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, *46*(1), 40. https://doi.org/10.1145/3292040.3219634

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences of the United States of America*, (22). doi.org/10.1073/PNAS.2018340118/SUPPL_FILE/PNAS.2018340118.SAPP.PDF

Klinect, J., Murray, P., Merritt, A., & Helmreich, R. (2003). Line operation safety audits: Definition and operating characteristics. *Proceedings of the 12th International Symposium on Aviation Psychology*, 663–668.

KPMG. (2020). *KPMG offers ethical AI Assurance using CIO Strategy Council standards*. https://home.kpmg/ca/en/home/media/press-releases/2020/11/kpmg-offers-ethical-ai-assurance-using-ciosc-standards.html

Knapp, M. C. (2021). *Contemporary auditing* (12th ed.). Cengage Learning.

Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, *5*(6), 679–685. https://doi.org/10.1038/s41562-021-01128-2

Kojima, T., Shane Gu, S., Reid Google Research, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. doi.org/10.48550/arxiv.2205.11916

Kolhar, M., Abu-Alhaj, M. M., & Abd El-Atty, S. M. (2017). Cloud data auditing techniques with a focus on privacy and security. *IEEE Security and Privacy*, *15*(1), 42–51. https://doi.org/10.1109/MSP.2017.16

Korbak, T., Elsahar, H., Kruszewski, G., & Dymetman, M. (2022). On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. *ArXiv*. https://arxiv.org/abs/2206.00761v2

Koshiyama, A., Kazim, E., & Treleaven, P. (2022). Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. *IEEE*, *55*(4), 40–50. https://doi.org/10.1109/MC.2021.3067225

Kostopoulos, L. (2021). *Decoupling human characteristics from algorithmic capabilities*. Medium. https://lkcyber.medium.com/decoupling-human-characteristics-from-algorithmic-capabilities-9c49b314b0a5

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020a). Defining AI in policy versus practice. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 72–78. https://doi.org/10.1145/3375627.3375835

Krafft, T. D., Zweig, K. A., & König, P. D. (2020b). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation and Governance*, *16*(1), 119–136. https://doi.org/10.1111/rego.12369

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., Van Esch, D., Ulzii-Orshikh, N., Tapo, A., … & Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, *10,* 50–72. https://doi.org/10.1162/TACL_A_00447/109285

Kritikos, M. (2019). Artificial Intelligence ante portas: Legal & ethical reflections. In *European Union*. https://www.europarl.europa.eu/at-your-service/files/be-heard/religious-and-non-confessional-dialogue/events/en-20190319-artificial-intelligence-ante-portas.pdf

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133). https://doi.org/10.1098/rsta.2018.0084

Kroll, J. A. (2021). Accountability in computer systems. In *The Oxford handbook of ethics of AI* (pp. 181–196). Oxford University Press.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(3), 663. https://doi.org/10.1002/ejoc.201200111

Kugelmass, H. (2016). "Sorry, I'm Not Accepting New Patients": An audit study of access to mental health care. *Journal of Health and Social Behavior*, *57*(2), 168–183. https://doi.org/10.1177/0022146516647098

Kumar, D., Mason, J., Bailey, M., Gage, P., Consolvo, K. S., Bursztein, E., Durumeric, Z., & Thomas, K. (2021). Designing toxic content classification for a diversity of perspectives. *Proceedings of the Seventeenth Symposium on Usable Privacy and Security*, 299–318. https://data.esrg.stanford.edu/study/toxicity-perspectives

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *31st Conference on Neural Information Processing Systems*. https://obamawhitehouse.-archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-

Kuusisto, A. (2001). *Safety management systems Audit tools and reliability of auditing at 12 o'clock noon* [Doctoral dssertation, Tampere University of Technology]. https://publications.vtt.fi/pdf/publications/2000/P428.pdf

LaBrie, R. C., & Steinke, G. H. (2019). Towards a framework for ethical audits of AI algorithms. *25th Americas Conference on Information Systems*, 1–5. https://dblp.org/rec/conf/amcis/LaBrieS19.html

Landers, R. N., & Behrend, T. S. (2022). Auditing the AI Auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*, *78*(1), 36–49. https://doi.org/10.1037/amp0000972

Langenkamp, M., Costa, A., & Cheung, C. (2020). Hiring fairly in the age of algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3723046

Langkafel, P. (2015). *Big Data in medical science and healthcare management*. Walter de Gruyter GmbH & Co KG.

Lapowsky, I. (2019). *How Cambridge Analytica sparked the great privacy awakening*. WIRED. www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society*, *7*(3), 437–451. https://doi.org/10.1017/als.2020.19

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *New Media & Society*, *9*(2), 1–16. https://doi.org/10.14763/2020.2.1469

Lauer, D. (2020). You cannot have AI ethics without ethics. *AI and Ethics*, *1*(1), 21–25. https://doi.org/10.1007/s43681-020-00013-4

Laux, J., Wachter, S., & Mittelstadt, B. (2021). Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA ☆. *Computer Law & Security Review*, *43*, 105613. https://doi.org/10.1016/j.clsr.2021.105613

Lawrence, G., Kearins, O., O'Sullivan, E., Tappenden, N., Wallis, M., & Walton, J. (2005). The West Midlands breast cancer screening status algorithm - Methodology and use as an audit tool. *Journal of Medical Screening*, *12*(4), 179–184. doi.org/10.1258/096914105775220705

Lea, H., Hutchinson, E., Meeson, A., Nampally, S., Dennis, G., Wallander, … & Khader, S. (2021). Can machine learning augment clinician adjudication of events in cardiovascular

trials? A case study of major adverse cardiovascular events (MACE) across CVRM trials. *European Heart Journal*, *42*. doi.org/10.1093/EURHEARTJ/EHAB724.3061

Lee, M. S. A., Floridi, L., & Singh, J. (2022). Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. In *The 2021 Yearbook of the Digital Ethics Lab* (pp. 157–182). Springer.

Lee, M., Floridi, L., & Denev, A. (2020). Innovating with confidence: Embedding governance and fairness in a financial services risk management framework. In *Ethics, governance, and policies in artificial intelligence* (pp. 353–371). Springer Publishing.

Lee, S. C. (2021). Auditing algorithms: A rational counterfactual Framework. *Journal of International Technology and Information Management*, *30*(2), 2021. https://doi.org/10.58729/1941-6679.1464

Lee, T.-H., & Azham, M. A. (2008). The evolution of auditing: An analysis of the historical development. *Journal of Modern Accounting and Auditing*, *4*(12), 1548–6583. https://www.researchgate.net/publication/339251518

Legg, C., & Hookway, C. (2020). Pragmatism. In *Stanford encyclopedia of philosophy*. PhilPapers.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications*, *157*, 17–24. doi.org/10.48550/arXiv.0706.3639

Leicht-Deobald, U., Busch, T., Schank, · Christoph, Weibel, A., … & Kasper, G. (2019). The Challenges of Algorithm-Based HR Decision-Making for Personal Integrity. *Journal of Business Ethics*, *160*, 377–392. https://doi.org/10.1007/s10551-019-04204-w

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy and Technology*, *31*(4), 611–627. doi.org/10.1007/s13347-017-0279-x

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*, 799–823. https://doi.org/10.1146/ANNUREV-PSYCH-010213-115043

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. Zenodo. https://doi.org/10.5281/zenodo.3240529

Levendusky, M. (2013). Partisan media exposure and attitudes toward the opposition. *Political Communication*, *30*(4), 565–581. doi.org/10.1080/10584609.2012.737435

Leveson, N. (2011). *Engineering a safer world: Systems thinking applied to safety*. MIT Press.

Li, L., Ma, R., Guo, Q., Xue, X., & Qiu, X. (2020). BERT-ATTACK: Adversarial attack against BERT using BERT. *EMNLP 2020 - 2020 Conference on Empirical Methods in*

Natural Language Processing, Proceedings of the Conference. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.500

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., … Koreeda, Y. (2022). Holistic evaluation of language models. In *Center for Research on Foundation Models*. https://arxiv.org/pdf/2211.09110.pdf

Lin, S., Openai, J. H., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*, 3214–3252. doi.org/10.18653/V1/2022.ACL-LONG.229

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 35–43. https://doi.org/10.1145/3233231

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine-learning scholarship. *Queue*, *17*(1), 1–15. https://doi.org/10.1145/3317287.3328534

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, *26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., & Oakden-Rayner, L. (2022). The medical algorithmic audit. *The Lancet Digital Health*, *4*(5), e384–e397. https://doi.org/10.1016/S2589-7500(22)00003-6

Loi, M., Ferrario, A., & Viganò, E. (2020). Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and Information Technology*, *Lipton 2018*. https://doi.org/10.1007/s10676-020-09564-w

Lomborg, S., Kaun, A., Sne, |, & Hansen, S. (2023). *Automated decision-making: Toward a people-centred approach*. https://doi.org/10.1111/soc4.13097

Luccioni, A., & Viviano, J. D. (2021). What's in the Box? An Analysis of undesirable content in the common Crawl corpus. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (2)*, 182–189. doi.org/10.18653/V1/2021.ACL-SHORT.24

Luckcuck, M., Farrell, M., Dennis, L. A., Dixon, C., & Fisher, M. (2019). A summary of formal specification and verification of autonomous robotic systems. *Integrated Formal Methods: 15th International Conference, IFM 2019, Bergen, Norway, December 2–6, 2019, Proceedings*, *11918*(5), 538–541. https://doi.org/10.1007/978-3-030-34968-4_33

Luhmann, N., & Barrett, R. (2018). *Organization and decision*. Cambridge University Press.

Lurie, E., & Mustafaraj, E. (2019). Opening up the black box: Auditing google's top stories algorithm. *32nd FLAIRS Conference 2019*, 376–381. https://aaai.org/ocs/index.php/FLAIRS/FLAIRS19/paper/view/18316/17433

MacCarthy, M., & Propp, K. (2021). *Machines learn that Brussels writes the rules: The EU's new AI regulation*. The Brookings Institution. www.brookings.edu/blog/techtank/2021-/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376445

Magee, B. (2016). *Ultimate questions*. Princeton University Press.

Maggetti, M., Gilardi, F., & Radaelli, C. (2015). Designing research in the social sciences. *Designing Research in the Social Sciences*, 21–41. doi.org/10.4135/9781473957664

Mahajan, V., Venugopal, V. K., Murugavel, M., & Mahajan, H. (2020). The algorithmic audit: working with vendors to validate radiology-AI algorithms—How we do it. *Academic Radiology*, *27*(1), 132–135. https://doi.org/10.1016/j.acra.2019.09.009

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law and Security Review*, *34*(4), 754–772. https://doi.org/10.1016/j.clsr.2018.05.017

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics*. doi.org/10.1007/s43681-022-00143-x

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial Intelligence. *ArXiv*. https://doi.org/10.48550/arXiv.2002.06177

Marda, V., & Narayan, S. (2021). On the importance of ethnographic methods in AI research. In *Nature Machine Intelligence* (Vol. 3, Issue 3, pp. 187–189). Nature Research. https://doi.org/10.1038/s42256-021-00323-0

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics*, *160*(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3

Maruyama, G., & Ryan, C. S. (2014). *Research methods in social relations*. John Wiley & Sons.

Matthews, J., Babaeianjelodar, M., Lorenz, S., …& Hughes, C. (2019). The right to confront your accusers: Opening the black box of forensic DNA software. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 321–327. doi.org/10.1145/3306618.3314279

Mau, S. (2019). *The metric society: On the quantification of the social*. John Wiley & Sons.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 622–628. https://doi.org/10.18653/V1/N19-1063

Mayson, S. G. (2019). Bias in, bias out. *Yale Law Journal*, *128*(8), 2218–2300. https://www.yalelawjournal.org/article/bias-in-bias-out

Mccarthy, J. (2007). *What is artificial intelligence?* Dimitris Diochnos. https://www.diochnos.com/about/McCarthyWhatisAI.pdf

McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., & Buyx, A. (2022). Embedded ethics: A proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics*, *23*(1), 1–10. https://doi.org/10.1186/s12910-022-00746-3

McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2018). Learning differentially private recurrent language models. *ICLR 2018 Conference Blind Submission*. https://openreview.net/pdf?id=BJ0hF1Z0b

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 729–733. doi.org/10.1145/3236024.3264833

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6). https://doi.org/10.1145/3457607

Merrer, E. Le, Pons, R., & Trédan, G. (2022). *Algorithmic audits of algorithms, and the law* (hal-03583919). http://arxiv.org/abs/2203.03711

Merriam-Webster. (2023). *Audit*. https://www.merriam-webster.com/dictionary/audit

Merton, R. K. (1948). The bearing of empirical research upon the development of social theory. *American Sociological Review*, *13*(5), 505. https://doi.org/10.2307/2087142

Merton, R. K. (1987). Three fragments from a sociologist's notebooks: Establishing the phenomenon, specified ignorance, and strategic research materials. *Review Literature And Arts Of The Americas*, *13*(1), 1–28. doi.org/10.1146/annurev.so.13.080187.000245

MetaAI. (2023). *System Cards, a new resource for understanding how AI systems work*. https://ai.facebook.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing algorithms. *Foundations and Trends in Human-Computer Interaction*, *14*(4), 272–344. https://doi.org/10.1561/1100000083

Metcalf, J., Anne Watkins, E., Singh, R., Clare Elish, M., & Moss, E. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–746. https://doi.org/10.1145/3442188.3445935

Meuwese, A. (2020). Regulating algorithmic decision-making one case at the time. *European Review of Digital Administration & Law*, 209-212. doi.org/10.4399/978882553896019

Microsoft. (2019). *Microsoft AI principles*. Communication. https://www.microsoft.com/en-us/ai/our-approach-to-ai

Microsoft. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI*. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

Mikians, J., Gyarmati, L., Erramilli, V., & Laoutaris, N. (2012). Detecting price and search discrimination on the Internet. *Hotnets*. www.researchgate.net/publication/232321801

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. SAGE Publications.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064. https://aclanthology.org/2022.emnlp-main.759

Minkkinen, M., Laine, J., & Mäntymäki, M. (2022). Continuous auditing of artificial intelligence: A conceptualization and assessment of tools and frameworks. *Digital Society*, *1*(3), 21. https://doi.org/10.1007/s44206-022-00022-2

Minkkinen, M., Zimmer, M. P., & Mäntymäki, M. (2021). Towards ecosystems for responsible AI: Expectations, agendas and networks in EU documents. *Proceedings of the 20th IFIP Conference on E-Business, e-Service and e-Society*, 220–232. doi.org/10.1007/978-3-030-85447-8_20

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

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mittelstadt, B. (2016). Auditing for transparency in content personalization systems. *International Journal of Communication*, *10*,4991–5002. www.researchgate.net/publication/309136069

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Moerel, E. M. L., & Storm, M. (2019). *Autonomous systems and the law* (N. Aggarwal, H. Eidenmüller, L. Enriques, J. Payne, & K. van Zwieten (Eds.)). CH Banks. https://doi.org/10.2139/ssrn.3356631

Mökander, J., & Axente, M. (2021). Ethics based auditing of automated decision making systems: Intervention points and policy implications. *AI & SOCIETY*, *0123456789*, 1–19. https://doi.org/10.1007/s00146-021-01286-x

Mökander, J., & Floridi, L. (2021). Ethics - based auditing to develop trustworthy AI. *Minds and Machines*, *0123456789*, 2–6. https://doi.org/10.1007/s11023-021-09557-8

Mökander, J., & Floridi, L. (2022a). From algorithmic accountability to digital governance. *Nature Machine Intelligence 2022*, 1–2. https://doi.org/10.1038/s42256-022-00504-5

Mökander, J., & Floridi, L. (2022b). Operationalising AI governance through ethics-based auditing: An industry case study. *AI and Ethics*, 1–18. https://doi.org/10.1007/s43681-022-00171-7

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022a). Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*, *32*(2), 241–268. https://doi.org/10.1007/s11023-021-09577-4

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Science and Engineering Ethics*, 1–30. https://doi.org/10.1007/s11948-021-00319-4 ORIGINAL

Mökander, J., Sheth, M., Gersbro-Sundler, M., Blomgren, P., & Floridi, L. (2022b). Challenges and best practices in corporate AI governance: Lessons from the biopharmaceutical industry. *Frontiers in Computer Science*, *4*, 106836. https://doi.org/10.3389/fcomp.2022.1068361

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023a). Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00289-2

Mökander, J., Sheth, M., Watson, D. S., & Floridi, L. (2023b). The Switch, the Ladder, and the Matrix: Models for Classifying AI Systems [Article]. *Minds and Machines*, *33*(1), 221–248. https://doi.org/10.1007/s11023-022-09620-y

Molnar, C. (2021). *Interpretable machine learning. A guide for making black box models explainable*. Lean Publishing.

Mondal, S., Das, S., & Vrana, V. G. (2023). How to bell the cat? A theoretical review of generative artificial intelligence towards digital disruption in all walks of life. *Technologies*, *11*(2), 44. https://doi.org/10.3390/TECHNOLOGIES11020044

Morgan, C. D. L., Krueger, R. A., & Morgan, E. D. L. (2016). Successful focus groups: Advancing the state of the art when to use focus groups and why. In *Advancing the State of the Art* (pp. 3–20). SAGE Publications Inc.

Morina, G., Oliinyk, V., Waton, J., Marusic, I., & Georgatzis, K. (2019). Auditing and achieving intersectional fairness in classification problems. *ArXiv*. https://doi.org/10.48550/arXiv.1911.01468

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mokander, J., & Floridi, L. (2021a). Ethics as a service: A pragmatic operationalisation of AI Ethics. *Minds and Machines*, *31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020a). From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141. https://doi.org/10.1007/s11948-019-00165-5

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., Floridi, L., & Morley, J. (2021b). Operationalising AI ethics: Barriers , enablers and next steps. *AI & SOCIETY*, *Villarreal 2020*. https://doi.org/10.1007/s00146-021-01308-8

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020b). The ethics of AI in health care: A mapping review. *Social Science and Medicine*, *260*(June). https://doi.org/10.1016/j.socscimed.2020.113172

Mueller, B. (2021). *How much will the artificial intelligence act cost Europe?* Information Technology and Innovation Foundation. https://itif.org/publications/2021/07/26/how-much-will-artificial-intelligence-act-cost-europe/

Muller, M., Chilton, L. B., Kantosalo, A., Maher, M. Lou, Martin, C. P., & Walsh, G. (2022). GenAICHI: Generative AI and HCI. *2022 CHI Conference on Human Factors in Computing Systems*, *110*. https://doi.org/10.1145/3491101.3503719

Munn, L. (2022). The uselessness of AI ethics. *Ai and Ethics*, 1–9. https://doi.org/10.1007/s43681-022-00209-w

Mustafaraj, E., Lurie, E., & Devine, C. (2020). The case for voter-centered audits of search engines during political elections. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 559–569. https://doi.org/10.1145/3351095.3372835

Myllyaho, L., Raatikainen, M., Männistö, T., Mikkonen, T., & Nurminen, J. K. (2021). Systematic literature review of validation methods for AI systems. *Journal of Systems and Software*, *181*, 111050. https://doi.org/10.1016/J.JSS.2021.111050

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 5356–5371.

Nadler, E., Arondekar, B., Aguilar, K. M., Zhou, J., Chang, J., Zhang, X., & Pawar, V. (2021). 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. *Journal of Cancer Research and Clinical Oncology*, *147*(3), 671–690. https://doi.org/10.1007/S00432-020-03414-4

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. *2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1953–1967. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.154

Narula, N., Vasquez, W., & Virza, M. (2018). zkLedger: Privacy-preserving auditing for distributed ledgers. *Proceedings of the 15th USENIX Symposium on Networked Systems*

Design and Implementation, 65–80. www.usenix.org/system/files/conference/nsdi18/nsdi18-narula.pdf

National Institute of Standard and Technology. (2022). *AI risk management framework: Second draft notes for reviewers: Call for comments and contributions*. https://arxiv.org/pdf/2302.08500

National Institute of Standard and Technology. (2022). *AI risk management framework*. https://www.nist.gov/itl/ai-risk-management-framework

National Institute of Standards and Technology. (2002). *Risk management guide for information technology systems recommendations of the National Institute of Standards and Technology*. www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative-/securityrule/nist800-30.pdf

National Institute of Standards and Technology. (2023). *Red Team (Glossary)*. https://csrc.nist.gov/glossary/term/red_team

Naudé, W., & Dimitri, N. (2020). The race for an artificial general intelligence: Implications for public policy. *AI and Society*, *35*(2), 367–379. https://doi.org/10.1007/S00146-019-00887-X/METRICS

Nejadgholi, I., & Kiritchenko, S. (2020). On cross-dataset generalization in automatic detection of online abuse. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 173–183. https://doi.org/10.18653/v1/P17

Neumark, D., Bank, R. J., & Van Nort, K. D. (1996). Sex Discrimination in Restaurant Hiring: An Audit Study. *The Quarterly Journal of Economics*, *111*(3), 915–941. https://doi.org/10.2307/2946676

Ng, A. (2021). *Can auditing eliminate bias from algorithms?* The Markup. https://themarkup-.org/the-breakdown/2021/02/23/can-auditing-eliminate-bias-from-algorithms

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, https://doi.org/10.18653/V1/2020.ACL-MAIN.441

Niemiec, E. (2022). Will the EU Medical Device Regulation help to improve the safety and performance of medical AI devices? *Digital Health*, 1–8. doi.org/10.1177/20552076221089079

Nilsson, M., & Weitz, N. (2019). Governing trade-offs and building coherence in policy-making for the 2030 agenda. *Politics and Governance*, *7*(4), 254–263. https://doi.org/10.17645/PAG.V7I4.2229

Noble, S. U. (2018). *Algorithms of oppression*. NYU Press.

Nozza, D., Bianchi, F., & Hovy, D. (2021). HONEST: Measuring hurtful sentence completion in language models. *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2398–2406. https://doi.org/10.18653/V1/2021.NAACL-MAIN.191

O'Donoghue, C., Splittgerber, A., & O'Brien, S. (2021). *The proposed European regulation on artificial intelligence – A summary of the obligations, scope and effect*. Reed Smith. https://www.reedsmith.com/en/perspectives/2021/05/the-proposed-european-regulation-on-artificial-intelligence--a-summary-of

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Books.

O'Neill, O. (2021). *A philosopher looks at digital communication*. Cambridge University Press.

Oakden-Rayner, L., Gale, W., Bonham, T. A., Lungren, M. P., … & Palmer, L. J. (2022). Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: A diagnostic accuracy study. *The Lancet Digital Health, 4*(5), doi.org/10.1016/S2589-7500(22)00004-8

Olympic Region Clean Air Agency. (2020). *It's the age of the algorithm and we have arrived unprepared*. https://orcaarisk.com/

OpenAI. (2016). *Generative models*. https://openai.com/blog/generative-models/

OpenAI. (2022). *Best practices for deploying language models*. https://openai.com/blog/best-practices-for-deploying-language-models/

Organisation for Economic Co-operation and Development. (2015). Principles of Corporate Governance 2015. In *G20/OECD Principles of Corporate Governance 2015*. OECD Publishing. https://doi.org/10.1787/9789264236882-EN

Organisation for Economic Co-operation and Development. (2019). *Recommendation of the council on artificial intelligence*. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Organisation for Economic Co-operation and Development. (2021). *Government at a glance*. https://doi.org/10.1787/1c258f55-en

Organisation for Economic Co-operation and Development. (2022). *OECD's Framework for the Classification of AI Systems*. https://doi.org/10.1787/cb6d9eca-en.

Organisation for Economic Co-operation and Development. (2023). *OECD's live repository of AI strategies & policies*. https://oecd.ai/en/dashboards/overview

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *ArXiv*. https://arxiv.org/abs/2203.02155v1

Oxborough, C., Cameron, E., Rao, A., Birchall, A., Townsend, A., & Westermann, C. (2018). *Explainable AI*. PwC. https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf

Page, S. E. (2018). *The model thinker: What you need to know to make data work for you*. Basic Books, Inc.

Pager, D. (2007). The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *Annals of the American Academy of Political and Social Science*, *609*(1), 104–133. doi.org/10.1177/0002716206294796

Pammolli, F., Righetto, L., Abrignani, S., Pani, L., Pelicci, P. G., & Rabosio, E. (2020). The endless frontier? The recent increase of R&D productivity in pharmaceuticals. *Journal of Translational Medicine*, *18*(1), 1–14. https://doi.org/10.1186/s12967-020-02313-z

Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., & Pedreschi, D. (2021). FairLens: Auditing black-box clinical decision support systems. *Information Processing and Management*, *58*(5). https://doi.org/10.1016/j.ipm.2021.102657

Parikh, P. M., Shah, D. M., Parikh, K. P., Parikh, P. M., Shah, D. M., & Parikh, K. P. (2023). Judge Juan Manuel Padilla Garcia, ChatGPT, and a controversial medicolegal milestone. *Indian Journal of Medical Sciences*, *75*(1), 3–8. https://doi.org/10.25259/IJMS_31_2023

Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, *87*(3), 457–477. https://doi.org/10.1086/708691

Partnership on AI. (2020). *Researching diversity, equity, and inclusion in the field of AI*. https://partnershiponai.org/researching-diversity-equity-and-inclusion-in-the-field-of-ai/

Pasquale, F. (2016). The Black Box Society: The secret algorithms that control money and information. *Information, Communication & Society*, *19*(12), 1727–1728. https://doi.org/10.1080/1369118x.2016.1160142

Patacchini, E., Ragusa, G., & Zenou, Y. (2015). Unexplored dimensions of discrimination in Europe: homosexuality and physical appearance. *Journal of Population Economics*, *28*(4), 1045–1073. https://doi.org/10.1007/s00148-014-0533-9

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), 100336. https://doi.org/10.1016/J.PATTER.2021.100336

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S., & Turini, F. (2018). Open the black box data-driven explanation of black box decision systems. *Computer Science*, *1*(1), 1–15. http://arxiv.org/abs/1806.09936

Peirce, C. S. (1903). *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectured on pragmatism*. Suny Press.

Peña-López, I. (2021). *OECD framework for the classification of AI systems*. Organization for Economic Co-Operation and Development. https://doi.org/10.1787/cb6d9eca-en.

Pentland, A. (2019). *A perspective on legal algorithms*. MIT Computational Law Report. https://law.mit.edu/pub/aperspectiveonlegalalgorithms/release/3

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022a). Red teaming language models with language models. *ArXiv*. https://doi.org/10.48550/arxiv.2202.03286

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S.,., … Kaplan, J. (2022b). Discovering language model behaviors with model-written evaluations. *ArXiv*. https://arxiv.org/abs/2212.09251v1

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S., & Niebles, J. . (2019). *The AI index 2019 annual report*. https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf

Personal Data Protection Commission. (2020). *Model artificial intelligence governance framework* (2nd ed.). International Association of Privacy Professionals.

Personal Data Protection Commission. (2022). *Launch of AI Verify - An AI Governance Testing Framework and Toolkit*. www.pdpc.gov.sg/news-and-events/-announcements/-2022/05/launch-of-ai-verify---an-ai-governance-testing-framework-and-toolkit

Peter, F. (2010). Political Legitimacy. In *Stanford Encyclopedia of Philosophy*. Stanford Univerity Press. https://plato.stanford.edu/entries/legitimacy/

Peyrard, M., Ghotra, S., Josifoski, M., Agarwal, V., Patra, B., Carignan, D., Kıcıman, E., Tiwary, S., & West, R. (2021). Invariant language modeling. *ArXiv*. https://doi.org/10.48550/arxiv.2110.08413

Pierné, G. (2013). Hiring discrimination based on national origin and religious closeness: Results from a field experiment in the Paris area. *IZA Journal of Labor Economics*, *2*(1), 1–4. https://doi.org/10.1186/2193-8997-2-4

Platt, J. (1992). Cases of cases . of cases. In Ragin & Becker (Eds.), *What is a Case? Exploring the Foundations of Social Inquiry* (1–242). Cambridge University Press.

Pope, C., & Mays, N. (1995). Qualitative research: Reaching the parts other methods cannot reach: An introduction to qualitative methods in health and health services research. *BMJ*, *311*(6996), 42. https://doi.org/10.1136/bmj.311.6996.42

Pound, P., & Campbell, R. (2015). Exploring the feasibility of theory synthesis: A worked example in the field of health related risk-taking. *Social Science & Medicine*, *124*, 57–65. https://doi.org/10.1016/J.SOCSCIMED.2014.11.029

Pound, P., & Campbell, R. (2015). Exploring the feasibility of theory synthesis: A worked example in the field of health related risk-taking. *Social Science & Medicine*, *124*, 57–65. https://doi.org/10.1016/J.SOCSCIMED.2014.11.029

Powell, D. A., Erdozain, S., Dodd, C., Costa, R., Morley, K., & Chapman, B. J. (2013). Audits and inspections are never enough: A critique to enhance food safety. *Food Control*, *30*(2), 686–691. https://doi.org/10.1016/J.FOODCONT.2012.07.044

Power, M. (1997). *The audit society: Rituals of verification*. Oxford University Press.

Powers, T. M., & Ganascia, J.-G. (2021). The ethics of the ethics of AI. In Dubber, Pasquale, & Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Prasad, M. (2021). Pragmatism as problem solving. *Socius*, *7*. doi.org/10.1177/2378023121993991

PwC. (2020). *PwC ethical AI framework*. https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html

Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence*, *3*(2), 104–110. https://doi.org/10.1038/s42256-021-00298-y

Punch, K. (2014). *Introduction to social research : quantitative & qualitative approaches* (Third edit). SAGE Publications Ltd.

Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data cards: Purposeful and transparent dataset documentation for responsible AI. *ACM International Conference Proceeding Series*, 1776–1826. https://doi.org/10.1145/3531146.3533231

Rachovitsa, A., & Johann, N. (2022). The human rights implications of the use of AI in the digital welfare state: Lessons learned from the Dutch SyRI case. *Human Rights Law Review*, *22*(2). https://doi.org/10.1093/HRLR/NGAC010

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., … & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*. https://github.com/OpenAI/CLIP.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., … Irving, G. (2022). Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*. https://doi.org/10.48550/arXiv.2112.11446

Raghavan, M., Barocas, S., Kleinberg, J., Levy, K., & Levy, K. 2020. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. doi.org/10.1145/3351095.3372828

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Rainforest Alliance. (2023). *Our approach*. https://www.rainforest-alliance.org/approach-/?_ga=2.137191288.953905227.1658139559-1130250530.1658139559

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, *46*(1), 192–210. https://doi.org/10.5465/AMR.2018.0072

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *AIES 2019* -

Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 429–435. https://doi.org/10.1145/3306618.3314244

Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022). The Fallacy of AI Functionality. *ACM International Conference Proceeding Series*, 959–972. doi.org/10.1145/3531146.3533158

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. doi.org/10.1145/3351095.3372873

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider oversight: Designing a third party audit ecosystem for ai governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571. doi.org/10.1145/3514094.3534181

Ramanadhan, S., Revette, A. C., Lee, R. M., & Aveling, E. L. (2021). Pragmatic approaches to analyzing qualitative data for implementation science: An introduction. *Implementation Science Communications 2021 2:1*, *2*(1), 1–10. https://doi.org/10.1186/S43058-021-00174-1

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen OpenAI, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *ArXiv*. https://doi.org/10.48550/arxiv.2204.06125

Rao, A. S. (2020). *Democratization of AI. A double-edged sword*. Toward Data Science. https://towardsdatascience.com/democratization-of-ai-de155f0616b5

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, *4*, 1–13. doi.org/10.1038/s41746-021-00455-y

Rauh, M., Mellor, J., Uesato, J., Huang, P.-S., Welbl, J., Weidinger, L., … & Hendricks, L. A. (2022). Characteristics of harmful text: Towards rigorous benchmarking of language models. *ArXiv*. https://doi.org/10.48550/arxiv.2206.08325

Reddy, E., Cakici, B., & Ballestero, A. (2019). Beyond mystery: Putting algorithmic accountability in context. *Big Data and Society*, *6*(1), 1–7. https://doi.org/10.1177/2053951719826856

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute*, *April*, 22. https://ainowinstitute.org/aiareport2018.pdf

Renda, A. (2018). The trolley problem and self-driving cars – A CSI into the ethics of algorithms. *CEPS Policy Insight*. https://papers.ssrn.com/=3131522

Renda, A., Arroyo, J., Fanni, R., Laurer, M., Maridis, G., & Devenyi, V. (2021). *Study to support an impact assessment of regulatory requirements for artificial intelligence in*

Europe. European Commission. https://artificialintelligenceact.eu/wp-content/uploads/2022/06/AIA-COM-Impact-Assessment-3-21-April.pdf

Reynolds, L., Ai, M., Ai, K., & Mcdonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *The 2021 CHI Conference on Human Factors in Computing Systems*, 1–7. doi.org/10.1145/3411763.3451760

Rhea, A. K., Markey, K., D'Arinzo, L., Schellmann, H., Sloane, M., Squires, P., Arif Khan, F., & Stoyanovich, J. (2022). An external stability audit framework to test the validity of personality prediction in AI hiring. *Data Mining and Knowledge Discovery*, *36*(6), 2153–2193. https://doi.org/10.1007/S10618-022-00861-0/FIGURES/8

Richards, L. (2009). *Handling qualitative data: A practical guide* (2nd ed.). SAGE Publications.

Richardson, R. (2022). *Defining and Demystifying Automated Decision Systems*. https://digitalcommons.law.umaryland.edu/mlr/vol81/iss3/2

Rizk, J. G., Barr, C. E., Rizk, Y., & Lewin, J. C. (2021). The next frontier in vaccine safety and VAERS: Lessons from COVID-19 and ten recommendations for action. *Vaccine*, *39*(41), 6017. https://doi.org/10.1016/J.VACCINE.2021.08.006

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2020). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI and Society*, *0123456789*. https://doi.org/10.1007/s00146-020-00992-2

Robertson, A. (2022). *Clearview AI agrees to permanent ban on selling facial recognition to private companies*. The Verge. www.theverge.com/2022/5/9/23063952/clearview-ai-aclu-settlement-illinois-bipa-injunction-private-companies

Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing partisan audience bias within Google search. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–22. https://doi.org/10.1145/3274417

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, … Bengio, Y. (2019). Tackling climate change with machine learning. *ArXiv*. doi.org/10.48550/arXiv.1906.05433

Rorty, R. (2021). *Pragmatism as anti-authoritarianism*. Harvard University Press.

Rosenfeld, R. (2000). Two decdes of statistical language modeling where do we go form here? Where do we go from here? *Proceedings of the IEEE*, *88*(8), 1270–1275. https://doi.org/10.1109/5.880083

Roski, J., Maier, E. J., Vigilante, K., Kane, E. A., & Matheny, M. E. (2021). Enhancing trust in AI through industry self-governance. *Journal of the American Medical Informatics Association*, *28*(7), 1582–1590. https://doi.org/10.1093/JAMIA/OCAB065

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *Proceedings of*

the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 41–58. https://doi.org/10.18653/V1/2021.ACL-LONG.4

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudinger, R. (2019). *GitHub - rudinger/winogender-schemas: Data for evaluating gender bias in coreference resolution systems.* GitHub,. https://github.com/rudinger/winogender-schemas

Rudner, T. G. J., & Toner, H. (2021). Key concepts in AI safety: Robustness and adversarial examples. In *Cyber Security Evaluation Tool*. https://cset.georgetown.edu/research/key-concepts-in-ai-safety-robustness-and-adversarial-examples

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/S11263-015-0816-Y/FIGURES/16

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Russell, S. J., & Norvig, P. (2015). *Artificial intelligence: A modern approach* (3rd ed.). Pearson.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, *36*(4), 105–114. doi.org/10.48550/arXiv.1602.03506

S.3572 - Algorithmic Accountability Act of 2022. (2022). In *117th Congress (2021-2022)*. Office of U.S Senator Ron Wyden.

Sabato, S., Sarwate, A. D., & Srebro, N. (2013). Auditing: Active learning with outcome-dependent query costs. *Advances in Neural Information Processing Systems*, 1–24.

Saldaña, J. (2009). *The coding manual for qualitative researchers*. SAGE Publications.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *ArXiv*, *2018*. http://arxiv.org/abs/1811.05577

Salkind, N. J. (2010). *Encyclopedia of research design*. SAGE Publications. https://www.jstor.org/stable/44648549

Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). *AI Watch. Defining Artificial Intelligence: Towards an operational definition and taxonomy of artificial intelligence*. https://doi.org/10.2760/382730

Samuel, A. L. (1960). Some moral and technical consequences of automation-A refutation. *American Association for the Advancement of Science*, *132*(3429), 741–742. http://ww.jstor.org/stable/1705808

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to "solve" the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 458–468. https://doi.org/10.1145/3351095.3372849

Sandu, I., Wiersma, M., & Manichand, D. (2022). Time to audit your AI algorithms. *Maandblad Voor Accountancy En Bedrijfseconomie 96(7/8): 253-265*, doi.org/10.5117/MAB.96.90108

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms. *ICA 2014 Data and Discrimination Preconference*, 1–23. doi.org/10.1109/DEXA.2009.55

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., & Smith, N. A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5884–5906. httpsdoi.org/10.18653/V1/2022.NAACL-MAIN.431

Sapiezynski, P., Zeng, W., Robertson, R. E., Mislove, A., & Wilson, C. (2019). Quantifying the impact of user attention on fair group representation in ranked lists. *The Web Conference 2019*, 553–562. https://doi.org/10.1145/3308560.3317595

Sargeant, H. (2022). Algorithmic decision-making in financial services: Economic and normative outcomes in consumer credit. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00236-7

Savage, D. D., & Bales, R. (2017). Video games in job interviews: Using algorithms to minimize discrimination and unconscious bias. *ABA Journal of Labor & Employment Law*, *32*(2), 211–228. https://www.jstor.org/stable/44648549

Schaefer, T. (2016). Applied qualitative research design: A total quality framework approach. In *Public Opinion Quarterly* (Vol. 80, Issue 1, pp. 215–217). Oxford University Press.

Schat, E., van de Schoot, R., … & Mendrik, A. M. (2020). The data representativeness criterion: Predicting the performance of supervised classification based on data set similarity. *PLOS ONE*, *15*(8). doi.org/10.1371/JOURNAL.PONE.0237009

Scherer, M. (2016). Regulating artificial intelligence systems: Risks, challenges, competences, and strategies. *Harvard Journal of Law & Technology*, (2), 98. doi.org/10.1007/s00521-010-0388-2

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, *2*(1), 31–42. https://doi.org/10.1109/tts.2021.3052127

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2021). Explaining the principles to practices gap in AI. *IEEE Technology and Society Magazine*, 81–94. doi.org/10.1109/MTS.2021.3056286

Schneider, G. (2019). Mind and machine in drug design. *Nature Machine Intelligence*, *1*(3), 128–130. https://doi.org/10.1038/s42256-019-0030-7

Schneider, J., Abraham, R., & Meske, C. (2020). AI governance for businesses. *Information Systems Management*. https://doi.org/10.48550/arXiv.2011.10672

Schonander, C. (2019). *Enhancing trust in artificial intelligence: Audits and explanations can help*. CIO. https://www.cio.com/article/220496

Schöppl, N., Taddeo, M., & Floridi, L. (2022). *Ethics Auditing: Lessons from Business Ethics for Ethics Auditing of AI*. 209–227. https://doi.org/10.1007/978-3-031-09846-8_13

Schroeder, R. (2007). *Rethinking science, technology, and social change*. Stanford University Press. https://doi.org/10.1515/9781503626454

Schuett, J. (2021). Defining the scope of AI regulations. *ArXIv*. http://arxiv.org/abs/1909.01095

Schuett, J. (2022). Three lines of defense against risks from AI. *ArXiv*. https://doi.org/10.48550/arxiv.2212.08364

Schuett, J. (2023). Risk Management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, 1–19. https://doi.org/10.1017/ERR.2023.1

Schulam, P., & Saria, S. (2019). Can you trust this prediction? Auditing pointwise reliability after learning. *The 22nd International Conference on Artificial Intelligence and Statistics*, *89*, 1022–1031. https://proceedings.mlr.press/v89/schulam19a.html

Schumpeter, J. A. (1942). *Capitalism, socialism, and democracy*. Allen & Unwin.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. *NIST Special Publication*, 1270. https://doi.org/10.6028/NIST.SP.1270

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2). https://doi.org/10.1177/2053951717738104

Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, *35*.

Senft, S., & Gallegos, F. (2009). *Information technology control and audit* (3rd ed.). CRC Press.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute trends across three eras of machine learning. *Proceedings of the International Joint Conference on Neural Networks*, 1–8. doi.org/10.1109/IJCNN55064.2022.9891914

Shabong, Y., & Aripaka, P. (2020). *Ex Cambridge Analytica boss banned over "unethical services": UK agency*. Reuters. https://www.reuters.com/article/britain-cambridge-analytica-idINKCN26F32E

Sharma, G. D., Yadav, A., & Chopra, R. (2020). Artificial intelligence and effective governance: A review, critique and research agenda. *Sustainable Futures*, *2*(December 2019), 100004. https://doi.org/10.1016/j.sftr.2019.100004

Sharma, S., Henderson, J., & Ghosh, J. (2019). *CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models*. https://doi.org/10.48550/arXiv.1905.07857

Shelby, R., Google, J., Henne, K., Rismani, S., Moon, A., Rostamzadeh, N., Nicholas Paul, N., Yilla, M., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2022). Sociotechnical harms: Scoping a taxonomy for harm reduction. doi.org/10.48550/arxiv.2210.05791

Shen, H., Devos, A., Eslami, M., & Holstein, K. (2021). Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–29. https://doi.org/10.1145/3479577

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *2019 Conference on Empirical Methods in Natural Language Processing*, 3407–3412. https://doi.org/10.18653/v1/D19-1339

Shevlane, T. (2022). Structured Access. In J. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. Young, & B. Zhang (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.

Silva, M., Santos De Oliveira, L., Andreou, A., Vaz De Melo, P. O., Goga, O., & Benevenuto, F. (2020). Facebook ads monitor: An independent auditing system for political ads on Facebook. *Proceedings of The Web Conference 2020*, 224–234. https://doi.org/10.1145/3366423.3380109

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., … & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Simig, D., Wang, T., Dankers, V., Henderson, P., Batsuren, K., Hupkes, D., & Diab, M. (2022). Text Characterization Toolkit (TCT). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 72–87. www.aclanthology.org/2022.aacl-demo.9

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial Bias in Pulse Oximetry Measurement. *The New England Journal of Medicine*, *383*(25), 2477–2478. https://doi.org/10.1056/NEJMc2029240

Slee, T. (2020). The incompatible incentives of private-sector AI. In *The Oxford Handbook of Ethics of AI* (pp. 106–123). Oxford University Press.

Sloane, M. (2021). *The algorithmic auditing trap*. OneZero. https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d

Sloane, M., Solano-Kamaiko, I. R., Yuan, J., Dasgupta, A., & Stoyanovich, J. (2023). Introducing contextual transparency for automated decision systems. *Nature Machine Intelligence 2023 5:3*, *5*(3), 187–195. https://doi.org/10.1038/s42256-023-00623-7

Smart Dubai. (2019). *AI ethics principles & guidelines*. https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf

Smieliauskas, W. J., & Bewley, K. (2010). *Auditing: An international approach* (5th ed.). McGraw-Hill Ryerson Higher Education.

Smith, B. C. (2019). *The promise of artificial intelligence: Reckoning and judgment*. MIT Press.

Smith, E. (2014). Research design. In H. Reis & C. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 27–48). Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.006

Smith, M., & Miller, S. (2022). The ethical application of biometric facial recognition technology. *AI and Society*, *37*(1), 167–175. https://doi.org/10.1007/S00146-021-01199-9/METRICS

Smith-Goodson, P. (2022). *NVIDIA's new H100 GPU smashes artificial intelligence benchmarking records*. Forbes. www.forbes.com/sites/moorinsights/2022/09/14/nvidias-new-h100-gpu-smashes-artificial-intelligence-benchmarking-records/?sh=5e8dca9ce728

Smuha, N. A. (2021). From a "race to AI" to a "race to AI regulation": Regulatory competition for artificial intelligence. *Law, Innovation And Technology*, *13*(1), 57–84. https://doi.org/10.1080/17579961.2021.1898300

Sobieszek, A., & Price, T. (2022). Playing games with AIs: The limits of GPT-3 and similar large language models. *Minds and Machines*, *32*(2), 341–364. doi.org/10.1007/s11023-022-09602-0

Sohoni, N. S., Dunnmon, J. A., Angus, G., Gu, A., & Ré, C. (2020). No subclass left behind: fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, *33*, 19339–19352. doi.org/10.48550/arXiv.2011.12945

Sokol, K., Santos-Rodriguez, R., & Flach, P. (2022). FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency. *Software Impacts*, *14*, 100406. https://doi.org/10.1016/j.simpa.2022.100406

Song, C., & Shmatikov, V. (2019). Auditing data provenance in text-generation models. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 196–206. https://doi.org/10.1145/3292500.3330885

Sookhak, M., Akhunzada, A., Gani, A., Khurram Khan, M., & Anuar, N. B. (2014). Towards dynamic remote data auditing in computational clouds. *Scientific World Journal*, *2014*. https://doi.org/10.1155/2014/269357

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Bilal Zafar, M. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. doi.org/10.1145/3219819.3220046

Springer, A., & Whittaker, S. (2019). Making Transparency Clear. In *Algorithmic Transparency for Emerging Technologies Workshop*. https://ceur-ws.org/Vol-2327/IUI19WS-IUIATEC-5.pdf

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., … Wu, Z. (2022). *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*. https://research.tilburguniversity.edu/en/publications/-beyond-the-imitation-game-quantifying-and-extrapolating-the-capab

Stake, R. E. (1995). *The art of case study research*. SAGE Publications. https://doi.org/10.48550/arXiv.2206.04615

Steed, R., Panda, S., Kobren, A., & Wick, M. (2022). Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 3524–3542. https://doi.org/10.18653/V1/2022.ACL-LONG.247

Steghöfer, J. P., Knauss, E., Horkoff, J., & Wohlrab, R. (2019). Challenges of scaled agile for safety-critical systems. *Product-Focused Software Process Improvement: 20th International Conference*, 350–366. https://doi.org/10.1007/978-3-030-35333-9_26

Steinberg, P. F. (2015). Can we generalize from case studies? *Global Environmental Politics*, *15*(3), 152–175. https://doi.org/10.1162/GLEP_a_00316

Stephan, U., Patterson, M., Kelly, C., & Mair, J. (2016). Organizations Driving Positive Social Change: A Review and an Integrative Framework of Change Processes. *Journal of Management*, *42*(5), 1250–1281. doi.org/10.1177/0149206316633268/ASSET/-IMAGES/LARGE/10.1177_01492063-16633268-FIG2.JPEG

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize from human feedback. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2009.01325v3

Stodt, J., & Reich, C. (2021). Machine learning development audit framework: Assessment and inspection of risk and quality of data, model and development process. *International Journal of Computer and Information Engineering*, *15*(3), 187–193. https://opus.hs-furtwangen.de/frontdoor/index/index/docId/7795

Stoel, D., Havelka, D., & Merhout, J. W. (2012). An analysis of attributes that impact information technology audit quality: A study of IT and financial audit practitioners. *International Journal of Accounting Information Systems*, *13*(1), 60–79. https://doi.org/10.1016/j.accinf.2011.11.001

Stratton, S. J. (2019). Literature reviews: Methods and applications. *Prehospital and Disaster Medicine*, *34*(4), 347–349. https://doi.org/10.1017/S1049023X19004588

Strenge, B., & Schack, T. (2020). AWOSE - A process model for incorporating ethical analyses in agile systems engineering. *Science and Engineering Ethics*, *26*(2), 851–870. https://doi.org/10.1007/s11948-019-00133-z

Susi, M. (2019). Balancing fundamental rights on the internet - The proportionality paradigm and private online capabilities. In M. La Torre, L. Niglia, & M. Susi (Eds.), *The Quest for Rights. Ideal and Normative Dimensions* (pp. 179–193). Edward Elgar Publishing.

Susskind, R., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts* (1st ed.). Oxford University Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Bradford Books.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Won, H., Aakanksha, C., Quoc, C., Le, V., Chi, E. H., Zhou, D. & Wei, J. (2022). Challenging BIG-bench tasks and whether chain-of-thought can solve them. *ArXIv*. https://arxiv.org/abs/2210.09261v1

Swedberg, R. (2014). *The Art of social theory* (Course Boo). Princeton University Press.

Sweeney, L. (2013). Discrimination in online Ad delivery. *Communications of the ACM*, *56*(5), 44–54. https://doi.org/10.1145/2447976.2447990

Taddeo, M. (2016). Data philanthropy and the design of the infraethics for information societies. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083). https://doi.org/10.1098/rsta.2016.0113

Taddeo, M. (2016). On the risks of relying on analogies to understand cyber conflicts. *Minds and Machines*, *26*(4), 317–321. https://doi.org/10.1007/s11023-016-9408-z

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, *40*(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Taeihagh, A., Ramesh, M., & Howlett, M. (2021). Assessing the regulatory challenges of emerging disruptive technologies. *Regulation and Governance*, *15*(4), 1009–1019. https://doi.org/10.1111/rego.12392

Tam, G. K. L., Kothari, V., & Chen, M. (2017). An analysis of machine- and human-analytics in classification. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 71–80. https://doi.org/10.1109/TVCG.2016.2598829

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *ArXiv*. https://doi.org/10.48550/arXiv.2102.02503 Focus to learn more

Tasioulas, J. (2018). First steps towards an ethics of robots and artificial intelligence. *SSRN Electronic Journal*, *7*(1), 61–95. https://doi.org/10.2139/ssrn.3172840

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

The Institute of Internal Auditors. (2009). *IIA position paper: The role of internal auditing in enterprise-wide risk management* (CAE Bulletin). www.theiia.org/globalassets/-documents/-resources/the-role-of-internal-auditing-in-enterprise-wide-risk-management-january-2009/pp-the-role-of-internal-auditing-in-enterprise-risk-management.pdf

The Institute of Internal Auditors. (2022). *About Internal Audit*. www.theiia.org/en/about-us/about-internal-audit/

Thomas, G. (2010). Doing case study: Abduction not induction, phronesis not theory. *Qualitative Inquiry*, *16*(7), 575–582. https://doi.org/10.1177/1077800410372601

Thomson, R., Plumridge, L., & Holland, J. (2003). Longitudinal qualitative research: A developing methodology. *International Journal of Social Research Methodology: Theory and Practice*, *6*(3), 185–187. https://doi.org/10.1080/1364557032000091789

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., … & Le, Q. (2022). *LaMDA: Language models for dialog applications*. Google. https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html?hl=fr&m=1

Thudi, A., Jia, H., Shumailov, I., & Papernot, N. (2021). On the necessity of auditable algorithmic definitions for machine unlearning. *31st USENIX Security Symposium*, 4007–4022. https://doi.org/10.48550/arXiv.2110.11891

Tolan, S. (2019). Fair and unbiased algorithmic decision making: Current state and future challenges. In *JRC Working Papers on Digital Economy* (2018-10). https://doi.org/10.48550/arxiv.1901.04730

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Tran, T. A., & Daim, T. (2008). A taxonomic review of methods and tools applied in technology assessment. *Technological Forecasting and Social Change*, *75*(9), 1396–1405. https://doi.org/10.1016/J.TECHFORE.2008.04.004

Truby, J., Brown, R. D., Ibrahim, I. A., & Parellada, O. C. (2022). A sandbox approach to regulating high-risk artificial intelligence applications. *European Journal of Risk Regulation*, *13*(2), 270–294. https://doi.org/10.1017/ERR.2021.52

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: Key problems and solutions. *AI & SOCIETY 2021 37:1*, *37*(1), 215–230. https://doi.org/10.1007/S00146-021-01154-8

Turley, G. A., Williams, M. A., & Tennant, C. (2007). Final vehicle product audit methodologies within the automotive industry. *International Journal of Productivity and Quality Management*, *2*(1), 1–22. https://doi.org/10.1504/IJPQM.2007.011465

Turley, S., & Cooper, M. (2005). *Auditing in the United Kingdom: A study of development in the audit methodologies of large accounting firms*. Prentice Hall.

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, *16*(3), 252–260. https://doi.org/10.1108/JICES-06-2018-0056

Tutt, A. (2017). An FDA for algorithms. *Administrative Law Review*, *69*(1), 83–123. https://doi.org/10.2139/ssrn.2747994

Ugwudike, P. (2021). AI audits for assessing design logics and building ethical systems: The case of predictive policing algorithms. *AI and Ethics*, *2*(1), 199–208. https://doi.org/10.1007/s43681-021-00117-5

Ulloa, R., Makhortykh, M., & Urman, A. (2019). Algorithm auditing at a large-scale: Insights from search engine audits. *Computer Science and Engineering*, *5*(7), 21–36.

Vakkuri, V., Kemell, K. K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2019). Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *Electronic Journal of Business Ethics and Organization Studies*, *27*(1), 4–15.

van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security*, *23*(4), 323–340. https://doi.org/10.1177/13882627211031257

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

Van Leeuwen, J. (2014). On Floridi's method of levels of abstraction. *Minds and Machines*, *24*(1), 5–17. https://doi.org/10.1007/s11023-013-9321-7

van Merwijk, C. (2022). *An AI defense-offense symmetry thesis*. LessWrong. www.-lesswrong.com/posts/dPe87urYGQPA4gDEp/an-ai-defense-offense-symmetry-thesis

Vanschoren, J. (2018). Meta-learning: A survey. *IEEE Computer Society*, 1–29. http://arxiv.org/abs/1810.03548

Vasetenkov, A. (2021). *AstraZeneca's knowledge graph: Drug discovery is a lot about connections*. Eckher Insights. https://www.eckher.com/c/21h530pr6z

Veale, M., & Borgesius, F. Z. (2022). Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*. https://doi.org/10.9785/cri-2021-220402

Vecchione, B., Levy, K., & Barocas, S. (2021). Algorithmic auditing and social justice: Lessons from the history of audit studies. *ACM International Conference Proceeding Series*, 1–9. https://doi.org/10.1145/3465416.3483294

Verband Der Elektrotechnik. (2022). *VCIO based description of systems for AI trustworthiness characterisation: (en)*. www.vde.com/resource/blob/-2177870/a24b13db01773747e6b7bba4ce20ea60/vde-spec-90012-v1-0--en--data.pdf

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *International Conference on Software Engineering*, 1–7. https://doi.org/10.1145/3194770.3194776

Verne, A., & Mir, S. (2019). *Interpretability of Machine Learning: What are the challenges in the era of automated decision-making processes?*

Villalobos, P., Sevilla, J., Besiroglu, T., Heim, L., Ho, A., & Hobbhahn, M. (2022). Machine learning model sizes and the parameter gap. *ArXiv*. http://arxiv.org/abs/2207.02852

Vincent, N., Johnson, I., Sheehan, P., & Hecht, B. (2019). Measuring the importance of user-generated content to search engines. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 505–516. https://arxiv.org/abs/1906.08576v1

Vinten, G. (1994). Participant observation: A model for organizational investigation? *Journal of Managerial Psychology*, *9*(2), 30–38. https://doi.org/10.1108/02683949410059299

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., … & Nerini, F. F. (2019). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*. https://doi.org/10.1038/s41467-019-14108-y

Vlok, N. (2003). *Technology auditing as a means of ensuring business continuity in a manufacturing organisation*. https://core.ac.uk/download/pdf/145048364.pdf

Voas, J., & Miller, K. (2016). Software certification services: Encouraging trust and reasonable expectations. *IEEE Computer Society*, 39–44. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1717342

Vogl, T. M., Seidelin, C., Ganesh, B., & Bright, J. (2020). Smart technology and the emergence of algorithmic bureaucracy: Artificial Intelligence in UK local authorities. *Public Administration Review*, *80*(6), 946–961. https://doi.org/10.1111/puar.13286

Vokinger, K. N., & Gasser, U. (2021). Regulating AI in medicine in the United States and Europe. In *Nature Machine Intelligence* (Vol. 3, Issue 9, pp. 738–739). Nature Research. https://doi.org/10.1038/s42256-021-00386-z

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, *31*(2), 841–888. https://doi.org/10.2139/ssrn.3063289

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Why fairness cannot be automated: Bridging the gap between EU Non-Discrimination Law and AI. *SSRN Electronic Journal*, *January*. https://doi.org/10.2139/ssrn.3547922

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *SSRN Electronic Journal*, *123*, 1–51. https://doi.org/10.2139/ssrn.3792772

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language

understanding systems. *Advances in Neural Information Processing Systems*, *32*. doi.org/10.5555/3454287.3454581

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, *ProceedingS*, 353–355. doi.org/10.18653/V1/W18-5446

Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., … & Li, B. (2021a). Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *Advances in Neural Information Processing Systems*. doi.org/10.48550/arxiv.2111.02840

Wang, J., & Yan, Y. (2012). The interview question. In *The SAGE Handbook of Interview Research: The Complexity of the Craft* (pp. 231–242). SAGE Publications Inc. https://doi.org/10.4135/9781452218403.n16

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Wang, S., Tu, Z., Tan, Z., Wang, W., Sun, M., & Liu, Y. (2021b). Language models are good translators. *ArXiv*. https://doi.org/10.48550/arxiv.2106.13627

Wang, Y., Wang, W., Joty, S., & Hoi, S. C. H. (2021c). CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8696–8708. https://doi.org/10.18653/v1/2021.emnlp-main.685

Wang, Z. J., Choi, D., Xu, S., & Yang, D. (2021d). Putting humans in the natural language processing loop: A survey. *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, 47–52.

Watson, D. (2021). *Explaining black box algorithms: Epistemological challenges and machine learning solutions* [Doctoral dissertation, University of Oxford]. https://ora.ox.ac.uk/objects/uuid:ba743054-3eaf-41fc-98e8-841255ee24ad

Watson, D. S., & Floridi, L. (2020). The explanation game: A formal framework for interpretable machine learning. In *Synthese*. Springer Netherlands.

Webb, E. J. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Rand McNally. https://doi.org/10.48550/arXiv.2112.04359

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., & Petrov, S. (2020). Measuring and reducing gendered correlations in pre-trained models. *ArXiv*. https://doi.org/10.48550/arxiv.2010.06032

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., … & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. doi.org/10.48550/arXiv.2206.07682

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., … Gabriel, I. (2021). Ethical and social risks of harm from language models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

Weiss, I. R. (1980). Auditability of software: A survey of techniques and costs. *MIS Quarterly: Management Information Systems*, *4*(4), 39–50. doi.org/10.2307/248959

Weiss, K., Khoshgoftaar, T. M., & Wang, D. D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 1–40. https://doi.org/10.1186/s40537-016-0043-6

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., & Huang, P. Sen. (2021). Challenges in detoxifying language models. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, 2447–2469. https://doi.org/10.48550/arxiv.2109.07445

Weller, A. (2017). Challenges for transparency. In *2017 ICML Workshop on Human Interpretability in Machine Learning*. https://openreview.net/forum?id=SJR9L5MQ-

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI Now report 2018*. AI Now Institute. https://ec.europa.eu/futurium/en/system/files/ged/ai_now_2018_report.pdf

Whittlestone, J., & Clarke, S. (2022). AI challenges for society and ethics. In *The Oxford Handbook of AI Governance*. Oxford University Press. doi.org/10.1093/oxfordhb/9780197579329.013.3

Whittlestone, J., Alexandrova, A., Nyrup, R., & Cave, S. (2019a). The role and limits of principles in AI ethics: Towards a focus on tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. https://doi.org/10.1145/3306618.3314289

Whittlestone, J., Nyrup, R., Alexandrova, A., & Dihal, K. (2019b). *Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research*. Nuffled Foundation.

Wiener, N. (1954). *The human use of human beings: Cybernetics and society* (2nd ed.). Doubleday Anchor Books.

Wiggins, C., Martin, E., & Jones, M. L. (2023). *How data happened: A history from the age of reason to the age of algorithms* (1st ed.). W. W. Norton & Company.

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677. https://doi.org/10.1145/3442188.3445928

Wilson, C., Marchetti, F., Di Carlo, M., Riccardi, A., & Minisci, E. (2020). Classifying intelligence in machines: A taxonomy of intelligent control. *Robotics*, *9*(3), 1–19. https://doi.org/10.3390/ROBOTICS9030064

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *EASE '14*. https://doi.org/10.1145/2601248.2601268

Wolf, F. M. (2011). Meta-analysis and synthesizing research. In *Meta-Analysis*. Little Green Books.

Woodside, A. G. (2016). Participant observation research in organizational behavior. In *Case study research* (pp. 331–352). Emerald Group Publishing Limited.

Wu, X., Liang, Z., & Wang, J. (2020). FedMed: A federated learning framework for language modeling. *Sensors*, *20*(14), 4048. https://doi.org/10.3390/S20144048

Xu, X., Chen, X., Liu, C., Rohrbach, A., Darrell, T., & Song, D. (2018). Fooling vision and language models despite localization and attention mechanism. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4951–4961. https://doi.org/10.1109/CVPR.2018.00520

y Arcas, B. A. (2022). Do large language models understand us. *Daedalus*, *151*(2), 183–197. https://doi.org/10.1162/daed_a_01909

Yanisky-Ravid, S., & Hallisey, S. K. (2019). Equality and privacy by design: A new model of artificial data transparency via auditing, certification, and safe harbor regimes. *Fordham Urban Law Journal*, *46*(2), 428–486. https://ir.lawnet.fordham.edu/ulj/vol46/iss2/5

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information Communication and Society*, *20*(1), 118–136. https://doi.org/10.1080/1369118X.2016.1186713

Yin, R. K. (1994). *Case study research: Design and methods* (2nd ed.). SAGE Publications.

Zang, S. (2022). *Chronicles of OPT development*. GitHub. https://github.com/facebookresearch/metaseq/tree/main/projects/OPT/chronicles

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science Technology and Human Values*, *41*(1), 118–132. https://doi.org/10.1177/0162243915605575

Zarsky, T. (2017). Incompatible: The GDPR in the age of Big Data. *Seton Hall Law Review*, *47*(4), 2. https://papers.ssrn.com/=3022646

Zerbino, P., Aloini, D., Dulmin, R., & Mininno, V. (2018). Process-mining-enabled audit of information systems: Methodology and an application. *Expert Systems with Applications*, *110*, 80–92. https://doi.org/10.1016/j.eswa.2018.05.030

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, 661–683. https://doi.org/10.1007/s13347-018-0330-6

Zhang, M., & Ré, C. (2022). Contrastive adapters for foundation model group robustness. *ICML 2022 Workshop on Spurious Correlations*. https://doi.org/10.48550/arxiv.2207.07180

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., … & Zettlemoyer, L. (2022). OPT: Open pre-trained transformer language models. *ArXiv*. http://arxiv.org/abs/2205.01068

Zhang, Z., Wang, S., & Meng, G. (2023). A Review on Pre-processing Methods for Fairness in Machine Learning. *Lecture Notes on Data Engineering and Communications Technologies*, *153*, 1185–1191. doi.org/10.1007/978-3-031-20738-9_128/FIGURES/2

Zicari, R. V., Brodersen, J., Brusseau, J., Dudder, B., Eichhorn, T., Ivanov, T., … & Westerlund, M. (2021). Z-Inspection®: A process to assess trustworthy AI. *IEEE Transactions on Technology and Society*, *2*(2), 83–97. doi.org/10.1109/tts.2021.3066209

Ziegler, D. M., Nix, S., Chan, L., Bauman, T., Schmidt-Nielsen, P., Lin, T., … & Thomas, N. (2022). Adversarial training for high-stakes reliability. *Advances in Neural Information Processing Systems*, *35*, 9274–9286. https://doi.org/10.48550/arXiv.2205.01663

Zinda, N. (2021). Ethics auditing framework for trustworthy AI: Lessons from the IT audit literature. In *Digital Ethics Lab Yearbook*. Springer.

Zittrain, J. L. (2014). The generative internet. *Connections: The Quarterly Journal*, *13*(4), 75–118. https://doi.org/10.11610/CONNECTIONS.13.4.05

APPENDIX 1 – DISCLAIMER FROM JESSICA MORLEY

Oxford, 10 March 2022

Subject: Collaboration with Jakob Mökander

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Jakob was the sole first author of the following article:

> Mökander, J., Morley, J., Taddeo, M. & Floridi, L. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Sci Eng Ethics* 27, 44 (2021). doi.org/10.1007/s11948-021-00319-4

Jakob conceived of the piece, conducted the literature review, and drafted the manuscript. I helped execute the analysis and helped revise the manuscript before publication. In summary, the work is substantially his own, and I have no objections to him submitting a lightly revised version of the article as a chapter of his doctoral thesis. Please do not hesitate to reach out if you have any questions.

Sincerely,

Jessica Morley

Doctoral Candidate
Oxford Internet Institute
University of Oxford
41 St. Giles', Oxford OX1 3LW

Email: jessica.morley@phc.ox.ac.uk

APPENDIX 2 – DISCLAIMER FROM MARIAROSARIA TADDEO

Oxford, 11 March 2022

Subject: Collaboration with Jakob Mökander

Dear Oxford Graduate Studies Committee/Examiners,

I, Mariarosaria Taddeo, was one of the co-authors of the article:

> Mökander, J., Morley, J., Taddeo, M. & Floridi, L. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Sci Eng Ethics* 27, 44 (2021). doi.org/10.1007/s11948-021-00319-4

My main contributions to the paper were providing contextual background and revising the final draft of the manuscript. Jakob had the idea for the study, provided its central framing, and was its lead author. I am happy for Jakob to include this paper as part of his DPhil thesis.

Sincerely,

Mariarosaria Taddeo

Associate Professor and Senior Research Fellow

ICSS DPhil Programme Director

Oxford Internet Institute

University of Oxford

41 St. Giles', Oxford OX1 3LW

Email: mariarosaria.taddeo@oii.ox.ac.uk

APPENDIX 3 – DISCLAIMER FROM LUCIANO FLORIDI

Oxford, 24 April 2023

Subject: Collaboration with Jakob Mökander

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Jakob was the sole first author of the following five articles:

> Mökander, J., Morley, J., Taddeo, M. & Floridi, L. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Sci Eng Ethics* 27, 44 (2021). doi.org/10.1007/s11948-021-00319-4
>
> Mökander, J., Floridi, L. Operationalising AI governance through ethics-based auditing: an industry case study. *AI Ethics* (2022). doi.org/10.1007/s43681-022-00171-7
>
> Mökander, J., Axente, M., Casolari, F. & Floridi, L. Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Act. *Minds & Machines* 32, 241–268 (2022). doi.org/10.1007/s11023-021-09577-4
>
> Mökander, J., Sheth, M., Watson, D. & Floridi, L. The Switch, the Ladder, and the Matrix: Models for Classifying AI Systems. *Minds & Machines* 33, 221–248 (2023). doi.org/10.1007/s11023-022-09620-y
>
> Mökander, J., Schuett, J., Kirk, H.., & Floridi, L. (2023) Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*. doi.org/10.1007/s43681-023-00289-2

As his supervisor, I gave input to Jakob's ideas, gave feedback to the presentation of the arguments, and edited the manuscripts. I am happy for Jakob to include each of the above articles as chapters in his thesis and present the material as his own ideas.

Sincerely,

Luciano Floridi

Professor of Philosophy and Ethics of Information

Oxford Internet Institute

University of Oxford

Email: luciano.floridi@oii.ox.ac.uk

APPENDIX 4 – DISCLAIMER FROM MARIA AXENTE

Oxford, 13 March 2022

Subject: Collaboration with Jakob Mökander

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Jakob Mökander was the lead author of our article:

> Mökander, J., Axente, M., Casolari, F. & Floridi, L. Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Act. *Minds & Machines* 32, 241–268 (2022). doi.org/10.1007/s11023-021-09577-4

Jakob came up with the idea for the article, coordinated the research, gathered, and analysed the data, and produced the first draft of the manuscript. As an industry practitioner, I helped validate the findings and contributed with real life examples. I also proof-read the final manuscript before publication. I am happy for Jakob to include the article as chapters in his DPhil thesis.

Sincerely,

Maria Axente

Responsible AI and AI for Good Lead, PwC
Advisory Board member of the UK Government All-Party Parliamentary Group on AI

Email: maria.axente@pwc.com

APPENDIX 5 – DISCLAIMER FROM FEDERICO CASOLARI

Oxford, 15 March 2022

Subject: Collaboration with Jakob Mökander

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Jakob Mökander was the lead author of our article:

> Mökander, J., Axente, M., Casolari, F. & Floridi, L. Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Act. *Minds & Machines* 32, 241–268 (2022). doi.org/10.1007/s11023-021-09577-4

As one of the co-authors, my main contribution was to validate the soundness of the arguments from a legal point of view. I also proof-read, revised, and added additional references to the final manuscripts before publication. That said, Jakob initiated and coordinated the research project, collected, and analysed the data, and wrote the first draft of the article. I am happy for Jakob to include a lightly revised version of our articles as a chapter in his DPhil thesis.

Sincerely,

Federico Casolari

Associate Professor of European Union Law
Deputy Head, Department of Legal Studies
Alma Mater Studiorum - Università di Bologna

Email: federico.casolari@unibo.it

APPENDIX 6 – DISCLAIMER FROM MARGI SHETH

Oxford, 13 April 2023

Subject: Collaboration with Jakob Mökander

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Jakob Mökander was the primary author of the following article:

> Mökander, J., Sheth, M., Watson, D., & Floridi, L. *et al.* The Switch, the Ladder, and the Matrix: Models for Classifying AI Systems. *Minds & Machines* 33, 221–248 (2023). doi.org/10.1007/s11023-022-09620-y

As a co-author and industry practitioner, I helped conceptualise the paper and validate the findings. I also proof-read the final manuscript before publication. However, Jakob came up with the idea for the article, coordinated the research, gathered, and analysed the data, and produced the first draft of the manuscript. I have no objection to Jakob including a lightly edited version of the article as a chapter in his DPhil thesis.

Sincerely,

Margi Sheth

Director Data Policy
R&D Data Office
AstraZeneca plc

Email: margi.sheth@astraZeneca.com

APPENDIX 7 – DISCLAIMER FROM DAVID WATSON

London, 14 April 2023

Subject: Collaboration with Jakob Mökander

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Jakob Mökander was the sole first author of the following article:

> Mökander, J., Sheth, M., Watson, D., & Floridi, L. The Switch, the Ladder, and the Matrix: Models for Classifying AI Systems. *Minds & Machines* 33, 221–248 (2023). doi.org/10.1007/s11023-022-09620-y

As the sole first author of the article, Jakob formulated the research question, gathered, and analysed the data, produced the first draft of the manuscript, and coordinated the project. As one of the co-authors, my main contribution was to I validate the findings. Given my background in ML and statistics, I focused primarily on the technical aspects of the article. I also proof-read the final manuscript before it was submitted for publication. That said, Jakob assumed most of the workload and I have no objection to him includes a lightly edited version of the article as a chapter in his DPhil thesis.

Sincerely,

David Watson

Lecturer in Artificial Intelligence
Department of Informatics
King's College London

Email: david.s.watson11@gmail.com

APPENDIX 8 – DISCLAIMER FROM JONAS SCHUETT

Oxford, 23 April 2023

Subject: Collaboration with Jakob Mökander

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Jakob Mökander was the sole first author of the following article:

> Mökander, J., Schuett, J., Kirk, H.., & Floridi, L. (2023) Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*. doi.org/10.1007/s43681-023-00289-2

As the sole first author of the article, Jakob formulated the research question, gathered, and analysed the data, produced the first draft of the manuscript, and coordinated the project. As one of the co-authors, my main contribution was to help conceptualise the study, provide feedback on the research methodology and article outline, and validate the soundness of the analysis. I also provided relevant sources to the bibliography and proofread the final manuscript before it was submitted for publication. However, Jakob assumed most of the workload, and I approve that he includes a lightly edited version of the article as a chapter in his DPhil thesis.

Sincerely,

Jonas Schuett

Research Fellow
Centre for the Governance of AI
Trajan House, Oxford

Email: jonas.schuett@governance.ai

APPENDIX 9 – DISCLAIMER FROM HANNAH ROSE KIRK

Oxford, 24 April 2023

Subject: Collaboration with Jakob Mökander

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Jakob Mökander was the sole first author of the following article:

> Mökander, J., Schuett, J., Kirk, H.., & Floridi, L. (2023) Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*. doi.org/10.1007/s43681-023-00289-2

As the sole first author of the article, Jakob formulated the research question, gathered, and analysed the data, produced the first draft of the manuscript, and coordinated the project. As one of the co-authors, my main contribution was to validate the soundness of the analysis and add technical depth to our description of model audits and dataset audits. I also provided relevant sources to the bibliography and proofread the final manuscript before it was submitted for publication. That said, Jakob assumed most of the workload and I approve that he includes a lightly edited version of the article as a chapter in his DPhil thesis.

Sincerely,

Hannah Rose Kirk

DPhil Candidate in Social Data Science
Oxford Internet Institute
University of Oxford

Email: hannah.kirk@keble.ox.ac.uk

APPENDIX 10 – QUESTION SHEET | SYSTEMATISED REVIEW

The following questions guided the critical examination of articles included in the *systematised literature review* (Grant & Booth, 2009) conducted in Chapter 3.

Meta info

1. Who are the authors of the article?
2. How was the research funded?
3. What is the purpose of the article?
4. Do conflicts of interests exist?

Context

5. How is EBA of ADMS defined?
6. What ethical challenges posed by ADMS do the author claim that EBA addresses?
7. How does EBA help address those challenges, according to the article?

Core contribution

8. What is the main argument presented in the article?
9. What is the subject of the audit: a person, an organisation, or a technical system?
10. Which tools and procedures are recommended to be included in the audit?
11. Who should conduct the audit?
12. According to which criteria are ADMS being evaluated during the audit?
13. Which parts of the auditing process are, or can reasonably be, automated?
14. How does the EBA procedure relate to existing ADMS governance structures?
15. What mechanisms incentivise the implementation of EBA of ADMS?
16. How does the EBA procedure ensure that decision-makers in organisations that design or deploy ADMS can be held accountable in case of irregularities?
17. Under which conditions is the outlined EBA procedure a feasible and effective mechanism for supporting the development of trustworthy ADMS?

Case study (if applicable)

18. What area of application (e.g., use case or sector) is examined?
19. What technology (e.g., symbolic vs sub-symbolic ADMS) is underpinning the examined application?
20. How probable, sensitive, and impactful are potential system failures for the ADMS?
21. Which ethics principles are covered by the EBA procedure?
22. What values and ethics principles have been embedded in the EBA procedure?

Limitations and gaps

23. What limitations of EBA as an ADMS governance mechanism do the authors discuss?
24. Which tensions between and within different ethics principles are highlighted in the article?
25. Which tradeoffs are highlighted between different ethics principles, technical system properties and organisational incentives?
26. How can these tensions and tradeoffs be managed, according to the article?
27. What challenges associated with developing EBA tools and are identified?
28. How can the methodological limitations and practical implementation challenges discussed in the article be managed?
29. What avenues for further research are suggested in the article?

APPENDIX 11 – QUESTION SHEET | INTERVIEWS

The following questions guided the semi-structured interviews I conducted with managers, software developers, and internal auditors within AstraZeneca, as reported on in Chapter 4.

Following best practices for *semi-structured interviews* (Edwards & Holland, 2013), I did not follow a strict manuscript but sought to have open dialogues with the interviewees. Hence, the below questions only indicate the information I sought to extract from the interview module as a whole: the focus of individual interviews varied depending on the participants' (professional) job description and (personal) interests.

<u>Meta info</u>

1. What is your job description/role within the organisation? (i.e., What are your primary responsibilities? What are your core tasks?)
2. What are the major initiatives you are executing or planning to execute to achieve your goals and address your issues?
3. How does your daily work relate to the design and deployment of AI systems?
4. How (if at all) have you and your team been involved in drafting the AstraZeneca AI ethics principles and the internal AI governance framework?

<u>Context</u>

5. How do (or would) you (and your team) define AI systems?
6. Are you (or your team) developing or using AI systems to support your objectives?
7. If yes, what are the primary benefits you hope to achieve by using or developing AI systems?
8. What AI technologies are underpinning the application? I.e., predictive/diagnostic, symbolic/connectionist, fully automated/decision support etc.
9. How probable, sensitive, and impactful are potential failures for the AI system?
10. What do you consider to be the most significant strengths of AI systems?
11. What do you consider to be the biggest ethical risks posed by AI systems, both from an organisational and societal perspective?
12. How do you see your department using AI in the next two years? What potential risks/governance factors do you think are relevant?

AstraZeneca AI Ethics principles and governance

13. How are you (and your team) currently managing ethical risks when using or developing AI systems?
14. What are the existing AI and data governance processes, policies, methods, initiatives, and tools that can be used to operationalise Data and AI Ethics and AI governance? And how effective are they?
15. How do you think Data and AI Ethics applies to your usage of AI?
16. What Data and AI Ethics principles do you consider important?
17. What is the value that such principles would bring to you (and to AstraZeneca)?
18. Who is accountable for decision-making regarding the design or usage of AI systems within your team?
19. How do you ensure that the design and use of AI systems respect AstraZeneca's risk and compliance policies?
20. Is there a process for measuring data and AI system quality – biases, accuracy, balance, etc.? If yes, who is responsible for this?

Ethics-based auditing

21. Do you believe that AI Governance in general, and ethics-based auditing in particular, will benefit you and your team? If so, how?
22. What is your role – and what are your responsibilities – within the emerging/recently implemented internal ethics-based auditing procedure?
23. What tools (e.g., software) and methods (e.g., assessment lists) are you using as part of the internal ethics-based auditing procedure?
24. Who/what is subject to the ethics-based audit: a person, an organisational unit, a software system or a technical component?
25. How does the process of ethics-based auditing of AI systems relate to existing structures of accountability and oversight?
26. What technical and practical constraints have you faced during the implementation of ethics-based auditing of AI systems?
27. How are you (and your team) managing these constraints associated with ethics-based auditing in practice?
28. What mechanisms incentivise the implementation of ethics-based auditing of AI systems for you and your team?
29. Is there a mechanism whereby the ethics-based auditing procedure is linked to personal accountability?

Suggestions for improvements

30. How stringent and enforceable should ethics-based auditing of AI systems be, in your opinion?
31. Which ethics principles do you think should be covered by ethics-based auditing procedures?
32. How do you think the AI systems you (and your team) are using should be evaluated and assessed? I.e., according to which methodology or metrics?
33. What is your recommendation to ensure ethics-based auditing is implemented successfully within AstraZeneca?

APPENDIX 12 – CODE HIERARCHY | THEMATIC ANALYSIS

As described in Section 4.5, I used *thematic analysis* (Braun & Clarke, 2006) to code the qualitative data collected as part of my longitudinal industry case study in Chapter 4. Table 5 summarises the 76 initial codes, 26 sub-themes, and 9 themes that resulted from this analysis.

Table 5. Code hierarchy generated through qualitative data analysis.

Initial codes	**Sub-themes**	**Themes**
Bias	Prevent harms	Balancing interests
Factual errors		
Privacy breaches		
Drug discovery	Reap benefits	
Innovation		
Operations		
Saving lives		
R&D	Risk appetite	
Red tape		
Agenda setting	Strategy	Catalysing change
Process improvement		
Regulatory preparedness		
AI design	Technology	
Digitalisation		
Implementation		
New use cases		
AZ principles	Values	
Responsible behaviour		
Audit fee	Financial costs	Costs
Infrastructure		
Core business	Opportunity costs	
Focus		
Audit meetings	Time investment	
Preparations		
Access	Challenges	External collaborations
Training data		
Visibility		
External software	Procurement	
Long-term contracts		
Negotiations		
Academic collaborations	R&D partnerships	
Joint ventures		
Startup collaboration		
Best practices	Decentralised organisation	Harmonising standards

Business areas		
Unclear mandate		
Regional differences		
Accountability	Enterprise level	
Compliance		
Resources		
Global policies		
Limited attention	Obstacles	Internal communication
Not relevant		
Agree on principles	Purpose	
Anchor decisions		
Employee consultation		
Onboard people		
Compliance documents	Roll-out	
Workshops		
Data governance	Already governed	Material scope
IT governance		
CSR		
Analytics	Define AI	
Data		
Project inventory		
Nothing new	Objections	
Standard practice		
Baseline	Key Performance Indicators	Measuring progress
Fairness		
Metrics		
Interviews	Methods	
Output logs		
Surveys		
Visualisation		
Accuracy	Quality management	
Control		
Safety		
Access	Data management	Verifying claims
Data storage		
Federated learning		
Assumptions	Documentation	
Completeness		
Process chart		
Role description		
Frequency	Model testing	
Tools		

APPENDIX 13 – NVIVO MIND MAP

Figure 17. Mind map of the themes and subthemes that emerged from my qualitative data analysis in Chapter 4.