AI Regulation Is (not) All You Need

Laura Lucaj, Patrick van der Smagt, Djalel Benbouzid

Machine Learning Research Lab, Volkswagen Group, Munich, Germany, laura.lucaj@argmax.ai

Preprint. Original is published at: FAccT '23:
Proc. 2023 ACM Conference on Fairness, Accountability, and Transparency, June 2023, Pages 1267-1279. DOI 10.1145/3593013.3594079

Abstract

The development of processes and tools for ethical, trustworthy, and legal AI is only beginning. At the same time, legal requirements are emerging in various jurisdictions, following a deluge of ethical guidelines. It is therefore key to explore the necessary practices that must be adopted to ensure the quality of AI systems, mitigate their potential risks and enable legal compliance. Ensuring that the potential negative impacts of AI on individuals, society, and the environment are mitigated will depend on many factors, including the capacity to properly regulate its deployment and to mandate necessary internal best practices along lifecycles. Regulatory frameworks must evolve from abstract requirements to providing concrete operational mandates that enable better oversight mechanisms in the way AI systems operate, how they are developed, and how they are deployed. In view of the above, this paper explores the necessary practices that can be adopted throughout a comprehensive lifecycle audit as a key practice to ensure the quality of AI systems and enable the development of compliance mechanisms. It also discusses novel governance tools that enable bridging the current operational gaps. Such gaps were identified by interviewing experts, analysing adaptable tools and methodologies from the software engineering domain, and by exploring the state of the art of auditing. The results present recommendations for novel tools and oversight mechanisms for governing AI systems.

1 Introduction

AI systems are rapidly transforming many areas of life with disruptive applications, such as the impressive advancement of generative AI models like ChatGPT, a text-generating chatbot, or the text-to-image generation provided by cascaded diffusion models such as DALL·E [1]. At the same time, the headlines unveiling the unintended consequences of the deployment of AI systems are constantly growing. Ethical challenges with the deployment of AI systems are not a new phenomenon and have been widely recognised in the literature [2, 3, 4]. For instance, it has been shown that Machine Learning (ML) systems can produce predictions that disproportionately affect vulnerable minorities, particularly in sensitive contexts such as criminal justice, hiring, and healthcare; and contribute to the re-enforcement of the structural bias already present in society [5, 6, 7, 8].

A recent example that highlights why some ML models can produce harm when deployed is provided by the analysis of the Stability AI releases of models comparable to DALL·E. Through reverse-engineering these open-source implementations, not only could the safety filters be bypassed but it was shown that these safety mechanisms were limited to a handful of criteria and overlooked a large part of the potential source of danger[9]. These findings display how the lack of documentation has prevented developers to access sufficient information for ex-ante mitigation measures, to properly understand and detect the safety risks [9]. Another example of deployment of controversial systems without properly assessing their potential impact, is the use of COMPAS, the algorithm has been shown to disproportionally predict high-risk scores of recidivism for African-American defendants, resulting in high false positive rates, whilst giving white defendants lower scores for similar cases, in turn resulting in high false negatives rates for the white population [10, 11]. Moreover, research has unveiled the inter-sectional accuracy disparities of commercial gender classifiers [6]. Such results demonstrate the alarming consequences of deploying face recognition systems in sensitive contexts such as law enforcement. In the domain of healthcare, an algorithm was less likely to refer African-American patients, who were equally sick, to programs to receive improved care due to complex medical needs [12, 13].

The issues mentioned above expose the absence of proper oversight mechanisms internally to test the systems before deployment, as well as external mechanisms, such as regulations, certification, and third-party auditing to enforce and mandate better practices on the companies developing such systems. Policy interventions seem therefore necessary, to enforce accountability mechanisms that could prevent the problematic consequences of the deployment of AI. To address such concerns, regulatory proposals have started to emerge around the world, such as the horizontal approach adopted by the European Union (EU) for the proposal for an AI regulation (AI Act) [14], the Bills proposed by the Canadian Government [15]. Moreover, the vertical approach adopted by the

Chinese government, where specific applications of AI are subject to regulatory requirements when deployed in certain contexts has been proposed, such as recommendation systems or generative AI models [16]. Whereas other governments around the world have been signaling their intent to endeavor into AI regulation, such as the Blueprint on an AI Bill of Rights, published by the White House [17]. Whilst the policy landscape around the world is rapidly evolving, the issues emerging with the deployment of AI are also increasing. This paper shows that the current governance landscape is inadequate to address all the challenges throughout the lifecycle of AI systems. Hence, regulation will not be enough. Ensuring that AI contributes to humans and the environment's well-being will depend on many factors such as the capacity to properly regulate its deployment, but also, by mandating necessary internal best practices along its lifecycle.

This paper aims, in particular, to identify the current operational gaps emerging from the horizontal approach adopted by the European Union. The proposed horizontal regulatory framework will target all AI systems across all sectors, which will be subject to the same set of risk-assessment criteria and legal requirements. Although the AI Act is not final yet, the current status of the regulation lays out four risk categories and defines a broad set of requirements for systems falling in each one [14]. The first category concerns systems that pose an "unacceptable risk", such as social scoring, whose deployment is forbidden [14]. The systems falling into the "high-risk" category are subject to a set of conformity assessment requirements [14]. Whereas the systems in the "limited risk" category, such as chatbots, have to adhere to a set of transparency requirements [14]. The fourth category concerns systems that pose no risk or minimal risk, which are subject to voluntary requirements [14]. The current draft, however, does not contain operational details on the necessary practices that could enable the conformity assessment. Hence, the burden of defining the specific compliance requirements falls onto Europe´s main standardisation bodies [18]. Moreover, the horizontal approach could lead to legal uncertainty, as many processes to provide oversight on the systems must be adopted in the early stages of the lifecycle.

Hence, this paper identifies a set of practices that can be leveraged to decrease the burden on policymakers, standardisation bodies, and businesses to enable compliance. On the one hand, businesses need operational guidance in the meantime, for the internal practices they must adopt, to enable the necessary conformity assessment practices, as waiting for the vertically oriented organisations to define the details of the compliance requirements might lead to significant compliance costs. As AI systems evolve rapidly, it is necessary to adopt oversight mechanisms in the early phases of their lifecycle to enable conducting comprehensive conformity assessment practices later on. On the other hand, policy-makers need recommendations on novel governance tools that can reduce the burden on them and the standardisation bodies in exploring the right methodologies and practices for determining the necessary requirements for legal compliance. Moreover, the regulatory landscape must move from abstract requirements to provide concrete operational mandates that enable better oversight mechanisms in the way AI systems operate, how they are developed and deployed, to disincentivise the accustomed "naive" approach that some AI companies follow by releasing models into the wild without proper adversarial testing and accountability for the impact produced by their systems. Although it is the purpose of legislation, to be formulated in a way that enables adaptable legal oversight on evolving technologies. Providing operational guidance can contribute to incentivise the adoption of the right measures internally and therefore foster the third-party oversight ecosystem. The motivation for this study is to understand the practices that must be adopted by developers to prevent such outcomes as well as the governance measures that have to be implemented to enforce the adoption of such practices. Hence, this research explores the necessary practices that can be adopted throughout a comprehensive lifecycle audit as well as the parallel novel governance tools that enable bridging the current operational gaps in the existing regulatory approach of the European Union.

A significant amount of research has been devoted to crafting ethical guidelines for practitioners, to guide their development of AI systems [19, 20, 21]. Nonetheless, the research on the practices that must be mandated by the regulation and required for compliance with standards is missing, leaving practitioners with significant legal uncertainty on what to adopt internally to comply with upcoming regulatory requirements. Such research is fundamental, as the ever-evolving nature of AI systems requires a set of oversight mechanisms that must be established and adopted throughout the lifecycle from the early design phases to the post-market monitoring. Hence, regulation will not be enough if the operational guidance on practices that public institutions must mandate or delegate to the private sector is missing. Investigating such gaps is fundamental, as the development of ethical AI systems is not merely a technical – ethical problem that requires only "tech ethics" solutions, but it concerns the complex dis-

tribution of power throughout the ecosystems [22]. There is a multitude of literature focusing on fairness [23, 24], transparency and explainability [25, 26], accountability [27, 7, 28], privacy and many other principles around the development of ethical AI [29, 30]. However, this paper will focus on the gaps restraining the translation of ethical principles, guidelines, and legal requirements into actionable practices. In order to tackle such a challenging topic, this research investigates the state-of-the-art research on practices that have been established in other industries, such as software engineering, aviation, and finance that could be leveraged for AI.

2 Related work

We begin with an overview of the current landscape of practices and methodologies that can be leveraged to address the gaps between the abstract legal and ethical requirements and the methods and tools that AI practitioners must adopt to develop ethical and trustworthy systems. This encompasses traditional software methodologies as well as documentation practices for AI systems. We put a particular emphasis on AI algorithmic auditing, as it could be a promising future direction for translating the widely debated principles of legal and ethical AI into actionable measures. Auditing enables the evaluation of a system against a determined set of criteria and standards, that could enable the exchange of relevant information about its performance, and provide recommendations on how to improve it [31]. Therefore, auditing could be a key practice to enable the conformity assessment required by the current draft of the AI Act.

2.1 Software Engineering Practices

As ML systems are increasingly being deployed at large-scale, practices drawn from the software engineering field that ensure compliance with the law and ethical standards have also been investigated [32, 33]. Looking at the software engineering field can provide insights into the essential technical challenges that arise as organisations develop large-scale AI solutions, such as keeping the systems robust and secure, as well as the adoption of quality assurance practices [34, 33]. Similarly to the field of AI, software engineering requires necessary maintenance and evolution practices that carry enormous costs and implementation challenges [35, 36]. The practices, such as bug bounties, audit trails, incident analysis and red team exercises, conducted in software engineering, can provide concrete methods to enable AI developers to assess the performance of the system as well as enabling to gain trustworthiness [37, 38, 33, 39, 40, 41]. One of the biggest challenges is to address unknown concerns by analysing the domain of deployment and the limitations or risks that could be potentially exploited by actors with malicious intentions [33].

For instance, red team exercises are often deployed within the industry to address antitrust and cybersecurity concerns and mitigate the potential misuse of the software, as well as, its vulnerabilities [37, 42]. A red team consists of members outside the organisation that stage adversarial attacks on the system to attempt unveiling its vulnerabilities or resilience [33]. Audit trails are another fundamental practice that can clarify accountability, by recording who was responsible for what, and at which stage [32, 43]. Moreover, audit trails can record the history of the development of an AI system and provide an overview of the important processes that were taken to develop the system, and keep track of the failures registered in the pipeline [32].

Bias and safety bounties are an established practice in software engineering to analyse potential risks and vulnerabilities that might have been overlooked, in the development phase, but necessarily need to be addressed before deployment [37, 33, 44]. The mechanism consists of financially rewarding security experts, upon the exposure of the vulnerabilities of the system, to enable the company to address such risks, before encountering compliance costs [37, 33]. In conclusion, researchers can benefit from having access to databases, and recording AI incidents that can expose the potential relation, between the deployment of certain systems and the contexts, nonetheless, reputations concerns and competitive pressures play a significant role in disincentivising such practices and result in a classic "collective action problem scenario" [45]. However, such software engineering practices can only be helpful to address certain phases of the pipeline of the lifecycle model of an AI system. Whilst for software systems individual components can be carefully tested, reviewed, and monitored, it is very difficult to handle AI components as distinct modules[33]. AI systems operate in a complex entanglement of chips, software development tools, large amounts of training data, extensive code libraries, and many deployment cases where validation and verification must be adapted, and all these may change on a daily base [32, 34, 46].

2.2 Documentation Practices

Emerging guidelines have addressed documentation practices for certain features and particular steps along the pipeline of an ML model such as model cards and datasheets [47, 48]. Model cards are a

framework that provides a transparent method for reporting the performance characteristics of an ML model [47]. Such documentation provides the details of the benchmarks on which a model was trained, such as the different cultural, demographic, and phenotypic backgrounds and the respective intersectional groups (age, gender, Fitzpatrick skin type) [47]. Moreover, model cards are essential in disclosing the exact intended purpose and context of the deployment of a model [47]. As ML models are typically evaluated against fixed datasets, documentation practices such as datasheets can enable understanding the characteristics of that evaluation [48]. Datasheets are inspired by the standardised forms of information sharing and rigorous testing, established in the electronic hardware industry, to enable the overview of the components performances under different test conditions [48]. Hence, datasheets can provide details on the tests that have been conducted on the dataset, the recommended use of it, and the respective regulations governing its use [48]. By analysing such information, practitioners are guided in understanding whether the data is equally representative of the populations and therefore if its fit for purpose, and which scenarios should be avoided [48]. However, documenting large datasets for AI systems is not trivial, as they are regularly or continuously updated and therefore present challenges that are hard to tackle with the current state-of-the-art technology of code repositories [40].

Analogous to a supplier's declaration of conformity (SDoC) which provides information on how a product conforms to the technical standards or regulations enforced in the country it is deployed, Factsheets have been adopted, to present a comprehensive documentation framework [49, 50]. Such documentation aims to record the practices conducted in the development of an AI model as well as disclosing the exact intended purpose of an AI system, to increase the consumer's trust, as well as addressing the potential ethical concerns emerging in this phase [49, 50].

Another important documentation practice is the collection of incidents. For instance, the AIAAIC Repository, developed by Charlie Pownall is an independent open library that collects incidents and controversies emerging from the deployment of AI systems since 2012[1]. The incident database can provide practitioners the oversight of the potential issues they might be dealing with based on their domain. Such practices can become part of a lifecycle audit process by enabling the development of evidence and logging the necessary information for auditors.

[1] AIAAIC Repository: www.aiaaic.org

2.3 Auditing Methodologies

Algorithmic auditing of ML systems has gained considerable recognition as an opportunity to harness the potential of ML models, as well as detect and mitigate the problematic patterns and consequences of their deployment in sensitive decision-making contexts [51, 52, 53]. Auditing is not new and has significantly contributed to promoting accountability and consistency in other highly-regulated industries such as finance and air mobility [33, 32]. Advancing the research on the application of such methodologies to AI, in parallel to governance tools, can enable to hold the organisations that develop and deploy AI accountable, by addressing risks through the implementation of control and oversight mechanisms [54, 52, 55]. However, in spite of the increased awareness of the worthwhile endeavour of researching in the field of algorithmic auditing, audit practices remain under-standardised and poorly investigated [31]. Hence, clearly defining audits is necessary to create alignment between the relevant practices; it also enables audits to give a more comprehensive picture of a system and avoid the so-called "ethics washing" phenomenon, where companies can hide the controversial impact of their AI systems behind an auditor's compliance stamp [56]. Auditing is a process that can unveil whether an organisation's past or present behaviour is aligned with the relevant principles, standards, and regulations [33]. It is considered one of the major mechanisms for supporting verifiable claims and converging towards trustworthy machine learning systems, by providing a stage for third parties to verify claims of the practices conducted by organisations to develop AI [33]. Auditing can be conducted in various ways, it can be an internal process conducted by internal teams of the companies developing an AI model (first-party auditing), it can be conducted by contractors (second-party audits), or by an independent entity with no contractual agreement with the developing company with a consequent public disclosure of the results (third-party auditing), or even by an external body with sanctioning power provided by the government [33, 31].

Mechanisms similar to auditing are conducted in other industries to ensure the trustworthiness of the product by independent oversight, such as accounting firms that conduct external audits, insurance companies that compensate for failures or consumer advocacy groups that give a seal of approval to a product or a service [32]. For instance, in the financial sector, several methodologies have been globally established to enable third-party auditing, such as the International Financial Reporting Standards (IFRS) [57]. Moreover, safety-critical industries rely on ro-

bust practices to assess the performance of systems such as the aerospace and aviation industries, nuclear energy, and healthcare, heavily rely on standardised practices and internal audits to maintain a required level of quality [7, 58, 32]. In the aviation industry, practices such as flight data recorders (FDR), have contributed to making civil aviation safer by providing a clear overview of the design of the products that operate in safety-critical domains [32, 38]. Such practices are fundamental in understanding crashes, by recording the right actions that prevented an accident and therefore by providing essential information on best practices [32].

A major contribution towards defining what auditing means for ML systems was proposed by [7]. The authors proposed a comprehensive framework for internal auditing to enable proactive interventions within the organisation for impact assessment [7]. The research tackles the practices necessary along the pipeline of an ML model to record important design decisions and to identify the causal relationship between such decisions and the risks that might emerge and relate to ethical failures [7]. This process allows the developers to detect the potential unintended consequences before the deployment [7].

2.4 Toolkits

Auditing AI is often seen through the lens of a single ethical concern, such as discrimination and bias or privacy concerns. To enable translating ethical principles into practice, several toolkits have been designed by academic institutions or private organisations. Hence, a number of auditing tools and frameworks tackle specific areas, e.g. diversity, bias in datasets [59, 60, 61] or explainability [62, 63, 60] have been developed. Moreover, open-source software frameworks have been proposed, to enable the analysis of fairness metrics and manage trade-offs between fairness and optimal model performance [64, 65]. To address the gaps that are caused by a focus on technical solutions to provide algorithmic equity, tools such as the Algorithmic Equity Toolkit can promote awareness to users of the impact that automated decision-making systems can have on their communities [66]. Moreover, the Model Card Authoring Toolkit enables community members to understand, whether ML models operate in alignment with their collective values [67].

Nonetheless, the practices analysed above must be included in a continuous auditing procedure to avoid a "reductionist" understanding of auditing that must encompass all the necessary oversight mechanisms throughout the lifecycle of technologies.

3 Methodology

Working at the intersection of the fields of computer science, software engineering, political science, and jurisprudence required to adopt a research methodology that allows understanding the asymmetries present between the regulatory requirements and the actual real-world issues that developers face to enable conformity assessment throughout the lifecycle of their technology as well as the practices conducted on auditing that enable compliance with the upcoming AI Act. To address the research objectives of this study, a qualitative analysis has been selected by conducting semi-structured interviews, with experts working at the novel research intersection of algorithmic auditing and compliance with the upcoming regulation. Table 1 provides an overview of the interview candidates. Through these interviews, novel solutions emerged to bridge the existing gaps between the legal and technical expertise needed to address such research. We choose semi-structured interviews, to fulfill the research objective of this thesis, as they offer a good balance by allowing to ask open-ended questions that do not restrict the interviewee's opinions and answers on a topic [68].

The interviews were processed through a *framework analysis*, a methodology drawing from several methods of qualitative research to provide targeted answers about specific populations and issues, with the aim of applying its findings to policy and practice [69, 70, 71]. Framework analysis is an inherently comparative form of thematic analysis that enabled the organised structure of inductively and deductively derived themes for conducting a cross-sectional analysis with a combination of data description and abstraction [71]. It enables the identification, description, and interpretation of key patterns within the topic or phenomenon of interest [72, 73]. More details about the specific steps of the framework analysis conducted in this paper can be found in appendix A.2. Under this framework, we drew the relationships between the different parts of the data and drew explanatory conclusions clustered around key themes. These insights are commonly hard to retrieve as both the regulation and the ecosystem of algorithmic auditing are still under development. Many of the interviewed experts have not yet published research on this topic, others work within companies to understand how to translate legal requirements into compliance strategies, which is information that is not publicly available. Other experts, are currently developing new methodologies for algorithmic auditing within academia and have not yet tested them in practice. Therefore, by interviewing such a diverse set of experts we were able to retrieve interesting results to answer our re-

search questions and explore the practices that can be leveraged or must be developed to address the gaps at the intersection of algorithmic auditing and compliance with the AI Act.

4 Analysis: Key Findings

Following the methodology above, we conducted a qualitative analysis of the concerns raised by the interviewees and we identified four common key themes that emerged.

4.1 Key Theme 1: State of the Art best Practices

The first theme that emerged through the analysis concerns the practices that the relevant stakeholders are already adopting or exploring. The experts revealed the best methodologies that attempt at translating the ethical and legal requirements into actionable measures; and gave an overview of what is missing that could be adopted. An important practice that emerged is the "leveraging of the testing methodologies, already established in the software engineering domain, to ensure that errors are identified before deployment" (Candidate 5). Among the testing methodologies that could be adapted to AI are "red team testing methodologies from the field of penetration testing" (Candidate 3). Looking at the start-up world, however:

> There are a couple of providers, if I talk to them I don't have the feeling actually that they exactly know what to test or how to test for standardisation and that's exactly the reason right I mean that nobody will know because standardisation is not there yet. (Candidate 8)

The findings show the experts' efforts into understanding how to "achieve transparency and explainability, whilst preserving intellectual property for AI, by following the procedures established by patent law" (Candidate 1). Another best practice is "corporate social responsibility" and the necessity to adopt better oversight mechanisms for corporate responsibility as "it is problematic to translate them into a legal requirement" (Candidate 2). Moreover, accountability is a key practice "to define roles and obligations for all relevant stakeholders" (Candidate 5), and "further to comply with the upcoming regulation" (Candidate 6).

4.2 Key Theme 2: Operational Issues in the AI Act

Most interviewees agreed that one of the current biggest challenges around regulating AI is to operationalise the legal requirements of the AI Act. It must be noted that the interviews were conducted before December 6th, 2022, when the Council of Europe published an updated, compromised version of the AI Act. Such reformulation should deliver sufficient information to distinguish software systems from AI systems[2]. Nonetheless, many of the recommendations delineated by the experts are still relevant, such as "solving the enforcement issues and making the intersection with other legislation clearer", the "risk of becoming a check-list legislation", and the "lack of technical expertise in the EU jurisdictions". Five candidates expressed that practitioners are experiencing doubts due to the current grey zones emerging from the unclear language around risk-classification of the AI Act (Candidates 3, 4, 5, 6, 8). Another important issue is the "intersection with other legislations such as consumer law, I think that when you are integrating AI into a variety of other products or services, you will be confronted with double obligations" (Candidate 1). It is increasingly fundamental to understand "who is my opposing part in this regulation right? I mean who's the authority that regulates me? Who is the authority I should give the information to?" (Candidate 8). Throughout the interviews, the lack of technical expertise in public institutions has been a clear recurring theme. The interviewees expressed how this "uneven distribution of talent among the relevant stakeholders is creating the current operational gaps one finds in the proposal" (Candidate 2). One of the main identified causes related to the lack of such expertise in the European jurisdiction is the competitive salaries offered by big tech companies (Candidates 2, 4, 5). However, one interviewee mentioned that "you can reach out to the commission and you do find people who are open to listen to you. So the question is, do they then in turn have an impact on the organisation?" (Candidate 9). Hence, it is rather an issue regarding making sure that technical expertise does not have a marginal impact on the inclusion of operational guidance in the legislation. One solution to address the lack of technical expertise could be to " introduce a body of advisory body, not only consisting of officials, like the European data protection board. It's comprised of just representatives of the authorities, but there's nobody from corporations in there." (Candidate 9). Another related theme that emerged through the in-

[2] For more information, read the European Council's Compromised text, online at: `https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf`

Table 1: *Interview Candidates*

Candidate ID	Role	Organisation	Background
1	Researcher	Academia	Law
2	Researcher	Academia	Law
3	Researcher	Academia	Computer Science
4	Team Lead	SME	Engineering
5	Tech and Regulatory Affair Counsel	Startup	Law
6	Manager	Multinational	Computational Linguistics
7	Researcher	Academia	Information Assurance
8	Founder/Board Member	Startup	Law
9	Chief Privacy Officer	Multinational	Law
10	Data Strategy Principal	Multinational	Engineering

terviews is "the need to address technical issues with appropriate technical language to guide practitioners" (Candidate 6).

4.3 Key Theme 3: Leveraging Technically Feasible Practices

The third theme identified the necessary internal practices that companies must adopt and further develop in order to audit their systems as well as to comply with the AI Act. Auditing emerged through most interviews as a key established methodology that "will be at the core of compliance with the AI Act" (Candidate 5) and that can address the current issues of understanding the impact of the systems through the development phase. Leveraging established documentation practices such as "model cards" and "datasheets" was a recurring topic throughout the interviews (Candidates 4, 5, 7). The establishment of these practices would "enable better oversight mechanisms on the market as well as create a kind of economic incentive for market actors to improve the quality of documentation of existing practices in order to comply with the AI Act" (Candidate 5). Another important topic that emerged through two interviews is the "necessity of establishing assurance practices" (Candidate 3), which are "methodologies to improve oversight mechanisms by verifying the validity of the claims made" (Candidates 3, 7). Assurance practices provide a framework to conduct "controls to ensure a certain technology does not produce wrong results" (Candidate 7). For instance, claiming that a system is transparent needs to be validated by an auditor based on evidence "that there is a documentation for non-expert users." (Candidate 3). Moreover, developing methodologies to assess data quality emerged throughout three interviews. An interviewee suggested that a "solution could be to apply to data the same regulations and requirements that are enacted in the supply chain to disincentivise the development of practices that explore how the data was sourced, similarly to what happens in the trade of raw materials" (Candidate 2). Data quality practices for instance should not merely cover the statistical properties of the training sets but also "guarantee that no infringements to human rights were conducted in the sourcing phase" (Candidate 2). An important gap that the interviews identified is the lack of cooperation between the relevant stakeholders in AI systems; there is a lack of interdisciplinarity, both in research and jurisdiction (Candidates 1, 2, 4, 5, 6, 9, 10). The development and deployment of AI is "a multi-stakeholder subject, so one needs somebody who knows about these legal issues and how to translate them" into requirements for an AI product, this person should also know how legal departments work" (Candidate 6). Another interview candidate explained the necessity to: "I guess it would be good to reshape the organisation that every team has then one legal guy as an interface expert. In order to have the linkage in each team to all of these different regulations." (Candidate 10) Finally, one interview highlighted the "necessity of applying tools implemented at a system level" (Candidate 6). In general software systems, part of the quality assessment practices is implemented at the system level, documentation is mostly generated automatically from code, tests are executed by the Continuous Integration system, and versioning is an integral part of the daily work, to give a few examples.

4.4 Key Theme 4: Governance Tools and Novel Policy-Making Solutions

The fourth theme explored the relevance and the experts' knowledge on the new regulatory instruments, to support the effectiveness of the AI Act. Practices such as policy-prototyping, regulatory sandboxes, standards, and certification were analysed to unveil the practices that policymakers must explore to address such gaps and provide better operational guidance through legislation. One of the main gover-

nance tools needed is harmonised standards, which could help to solve the grey zones in understanding the regulatory requirements and the current need to go through a third-party assessment to access the market. Certification emerged throughout four interviews as fundamental in improving the ecosystem (Candidates 1, 3, 4, 5). Certifications "enable companies to cooperate with experts, that can guide them to internally adopt the measures that will enable compliance with standards and norms" (Candidate 4).

Nonetheless, the field would only benefit from the development of accredited certification:

> It depends on what kind of certification you're talking about. I can give you a certificate. It has no worth at all, but I can certify what I want. But, if we talk about accredited certification, there's a clear procedure for that. That's one body in each country that is allowed to provide accredited certification. It's also allowed to certify other companies that can perform accredited certification. But, accredited certification means there are very well-specified rules, things to look into, processes, and so on. We do not have any of these. (Candidate 3)

Policy prototyping emerged through three interviews as a novel policy-making tool that can address the operational gaps of the AI Act and help policymakers adapt legal requirements, given the fast-evolving pace of AI systems (Candidates 1, 2, 5). It enables experimentation with a potential rule to understand how it can be applied and what methodologies must be implemented for compliance (Candidates 1, 2, 5). One point that was raised by one expert is "the importance of the whistle-blower acts and to enable better oversight mechanisms by incentivising companies to adopt compliance methodologies" (Candidate 2). Moreover, regulatory sandboxes have emerged through four interviews as an interesting policy-making tool to experiment with regulation and understand how to operationalise the legal requirements (Candidates 1, 2, 4, 5).

> The regulator or the authority creates some kind of a very practical interface, where both the market and public authorities, are able to engage in a very dynamic, or even testing approach way. So basically what we are doing with this is to optimise communication channels between the market and the regulator and therefore create, in the long run, a more dynamic flow, when it comes to regulatory adaptation. (Candidate 5)

Digital Hubs, an environment where collaboration can take place to support the market's compliance challenges with the AI Act, emerged through one interview as a solution to enable forming public-private partnerships that could overcome all the shortcomings, of the lack of resources of public institutions. "I would suggest more something like digital hubs, where you can have them in, I mean, some of them are already forming as public-private partnerships in a way" (Candidate 2).

5 Discussion

A significant amount of current research has been devoted to analysing the shortcomings of the AI Act. However, the field necessitates more investigation into the practices that can be leveraged to comply with such legal requirements and that eventually have to be mandated by legislators for proof of compliance. We identify the research on algorithmic auditing as a worthwhile endeavour to understand which practices must be conducted internally to produce oversight mechanisms along the lifecycle of an AI system. Hence, the analysis enabled gaining an in-depth overview of the necessary practices that must be developed along the lifecycle of AI systems, as well as the novel governance tools that can bridge the current operational gap of the AI Act proposal.

The research unveiled the necessity of exploring the intersection of algorithmic auditing and AI regulation, as the field currently lacks operational practices that translate the legal requirements into actionable measures. The main finding of the research was that both the regulation and auditing methodologies must focus on the set of practices that are conducted throughout the lifecycle of an AI system and determine its impact, instead of focusing on the outcomes. Algorithmic auditing is still a nascent research field, therefore the AI Act mentions that audits cannot be mandated yet. Nonetheless, many of the practices that the experts have explored during the interviews overlap with auditing practices such as testing methodologies, assurance practices, and documentation throughout the lifecycle.

In spite of the importance of taking the necessary measures to govern AI systems and the indispensability of delivering a well-drafted regulation, such top-down governance measures will not be enough to address the current challenges in the ethical AI ecosystem. Regulation is important but only when it contains operational guidance, upon which standards can be drafted to support and guide the compliance methodologies for practitioners. Moreover, standards must be supported by the presence of effective third-party oversight mechanisms through exter-

nal independent auditors and accredited certification bodies. Hence, this paper investigates what practices can be leveraged to fill the current gaps in the field and advance research on the growing necessity to develop effective auditing mechanisms that enable the translation of legal requirements into actionable measures.

Four major themes have been identified by our research, overall to explain the necessary practices that must be leveraged to address the gap at the intersection of AI Auditing and regulation.

5.1 Lack of Holistic Methodologies

Many experts agreed that it is not necessary to redevelop certification methods but rather to explore the practices that have been established in other domains such as software engineering or healthcare. One of the most important practices that have been discussed is testing. Testing has to become a fundamental part of developing AI systems, just as it is a high percentage of developing software. The mindset of AI practitioners has to shift from producing models that "work" to understanding how they should work, in which contexts they can be deployed, and what is the safe extent of their deployment. Adversarial testing must be conducted before such systems enter the market. This necessity becomes clear with the current issues emerging with large language models such as Meta's Galactica, a tool for aiding scientific writing, or Chat-GPT, a chatbot. In spite of the hype generated by such models, several failures have been registered, such as the inability to distinguish truth from false claims when aiding scientific writing [74]. Such incidents are a consequence of the lack of effective oversight mechanisms, both to be internally developed by the companies to assess the potential impact of such systems throughout the lifecycle as well as externally by enabling accountability for the decisions made to develop such systems. However, through the interviews, it emerged that in spite of the awareness of the importance of testing, the tools and methodologies are missing, meaning that even if new providers are emerging on the market, it is not clear if they know what they are doing, because of the lack of standards and operational requirements in the legislation that can serve as a benchmark of what needs to be done to prove compliance.

The risks associated with the deployment of ML models have been widely analysed and range from robustness issues due to private data leaks to providing and spreading misinformation, or the manipulation of users leading them to overestimate its capabilities and use it unsafely [75]. Other industries do not deploy products without a set of strict controls and trials, such as the medical field. To enter the market, systems have to comply with very strict requirements. This mindset is not there for AI and this must change. Testing is then connected to the other practices that have been discussed, such as transparency and explainability, to disclose the companies' decisions that determine the safe extent of the deployment of AI systems. Moreover, the adoption of testing and transparency can enable accountability for users subject to systems and incentivise corporate social responsibility. In spite of the relevance of such practices, a common theme that emerged through the interviews is the low adoption of tools and methodologies to operationalise such principles within the company's processes. In spite of the importance of such practices, comprehensive methodologies for auditing seem to be still under-investigated, which contributes to confirming the hypothesis analysed in this paper regarding the presence of gaps between the practices that companies must adopt internally and the practices and tools that enable compliance.

5.2 Necessity to Mandate Actionable Practices

The second main theme that emerged was the necessity of operational guidance in the legal requirements of the AI Act. One of the main identified reasons for the lack of actionable measures, as in the current draft of the legislative proposal, is the lack of technical expertise in the European Jurisdiction. Although many technical experts were involved in the drafting of the AI Act, their influence on the document seems marginal. This is manifested in the presence of repeated problematic terminology that most interviewees pointed out, such as "error-free" and "complete" data, and other examples where the legal requirements do not reflect technically feasible practices. The current compromised version of the AI Act solved some of the problematic terminology in response to the wide feedback received from practitioners, nonetheless, the lack of technical expertise in the public sector produces issues that go beyond problematic terminology and effects.

In order to have well-drafted laws and standards, the presence of expertise in translating technical measures and processes into legal requirements and standards is fundamental. Regulatory sandboxes, policy prototyping, and digital hubs enable a novel cooperative environment between public and private institutions and hence must be explored. These environments enable the development and collaboration of the expertise that is necessary to address the novel fields of algorithmic auditing and compliance with AI regulation. Moreover, AI systems are not deployed in a vacuum and are often utilised in several highly-regulated industries such as healthcare and

the automotive as well as critical infrastructure. This leads to the need for clarifying the potential overlap between the legal requirements of the AI Act and other domain-specific regulations.

5.3 Infancy of AI Auditing Field

The third theme that was investigated is the necessity of developing technically feasible practices to enable bridging the gap between the legal requirements and the actionable measures that companies must adopt internally to develop ethical, trustworthy and compliant systems. All interviewees agreed on the importance of moving from a theoretical ethical debate to understanding the actionable measures that enable the development of ethical systems that comply with regulations. Various practices have been considered fundamental, such as "auditing", "assurance mechanisms", "data quality". These practices will be essential for translating legal requirements, into actionable processes.

Nonetheless, further exploration and development is necessary, as there are many gaps in the research around these domains. The law requires technically feasible actions, especially when comparing the AI Act to the enforcement struggles that practitioners experienced with GDPR, where many policies were written and companies struggled to adopt them or translate them into actionable practices. Such legal uncertainty left the authorities with little capacity to verify them. Moreover, the AI Act must find an equilibrium between mandating technically feasible measures and not setting lower requirements than the practices that are actually feasible. However, in spite of the identified relevance of auditing, the results show the novelty and under-investigation that still characterises the field due to the lack of guidance into a comprehensive audit methodology. The practices suggested by the experts such as assurance mechanisms, documentation, testing, and data quality methodologies are necessary to enable effective audit mechanisms. Furthermore, such practices can bridge the gap between the legal requirements and the actionable measures necessary to prove compliance. Established documentation practices such as datasheets and model cards have been presented as a fundamental tool to bridge the gaps between the need for transparency and accountability and the necessity of developing internal oversight mechanisms for companies to log fundamental information about the design processes. Therefore, the AI Act could significantly contribute to the development of such auditing mechanisms, due to the necessity to adhere to the conformity assessment requirements. Moreover, to improve the auditing ecosystem, novel governance solutions are required such as accredited certification for auditors that can certify the conformity assessment practices, as well as standards that define the compliance requirements.

5.4 Novel Governance Tools

The last theme that was studied was the necessity of exploring novel policy-making tools to enable the development of effective oversight mechanisms that are hard to achieve with traditional top-down regulatory approaches. We argue that the lack of technical expertise in public institutions needs to be balanced. This can be done in collaborative environments for co-regulation, where policies can be tested and evaluated. Such settings have been explored through regulatory sandboxing and policy prototyping, as well as by innovation hubs. These processes and environments can further contribute to bridging the operational gap between the legal requirements and the best practices conducted by practitioners along the lifecycle of the systems they develop.

Other important tools that have been emerging are standards and certifications. However, standards should be integrated into a certification scheme mandated by legislators that would create a balance between the regulatory oversight mechanisms and the increasingly heated competition by ensuring mechanisms that incentivise the development of reliable, ethical, and trustworthy systems. One of the most important aspects that emerged is the importance of European legislators mandating actionable measures through legally binding obligations to create the necessary quality assurance practices that can be specified in the harmonised standards and guide the process for accredited certifications. Delegating regulatory powers to private bodies is problematic due to the lack of democratic oversight mechanisms by European legislative institutions, the scarce involvement of all relevant stakeholders such as consumers, and the current immunity of standards to judicial control [76].

Hence, the European Union must delineate a clear path for co-regulation, by subjecting the discussion of fundamental ethical and legal decisions as well as the relevant technical solutions to legislative procedures and debates that are cooperatively shaped by all relevant stakeholders, such as civil society, the industry, academia, among many. One of the solutions is to invest in significant in-house training of the technical expertise within the governments to enable proper independence instead of heavily relying on private partners [77]. Moreover, the risk of low availability of tech talent in public institutions and the potential issues of accessing sandboxing and digital hubs for startups and SMEs could lead to the risk of research and development monopoly from bigger tech

corporations, which would further deepen the gaps currently present on the market. There is a need for an environment where the private sector and public institutions cooperate in developing the best practices for the development and deployment of ethical AI, conducting research on auditing, and developing and testing software and documentation tools throughout the lifecycle. Furthermore, such cooperation has to bring the interests of SMEs to the table, as they have less access to such resources.

6 Limitations

There are significant constraints to the representativeness and accuracy of the findings. First of all, conducting qualitative research through interviews might lead to a sample bias that is not accurately representing the whole population. Moreover, another issue that emerged is the lack of statistical representativeness of the sample for the whole population, due to the lack of expertise, working at the intersection of legal and technical issues around AI. So sample size is, as always, an issue. Our sample size can well be not representative of a full population, for which reason we decided to focus on key messages rather than discuss statistics. So we compensated by conducting a qualitative analysis based on the framework approach, which of course may lead to the bias of the inductive elements of the researchers' interpretation. Another limitation is caused by the necessity of conducting most of the interviews online rather than in person. For a successful data collection through qualitative interviews, an essential feature is the ability of the interviewer to create a comfortable environment for the interviewee to express their concerns and opinions (Warren, 2002). However, interviewing online can potentially lead to the presence of observer bias, by preventing some interviewees from showing their real concerns, but rather saying what they thought the interviewer was expecting them to say.

7 Policy Recommendation

Based on the analysis conducted in this paper, this section outlines three main policy recommendations to bridge the current gaps between the lack of operational guidance in the AI Act draft, and the practices that must be adopted along the AI lifecycle.

7.1 Mandating Audits through novel AI Lifecycle models

The definition of auditing must change. Currently, only the outcomes of an AI system are subjected to auditing, however, to properly address the problematic outcomes, the processes throughout the lifecycle of an AI system must be audited. This is fundamental because fixing the issues before deployment will be costly, and so will understanding what generated the unintended outcome and re-collecting and labeling the data. The practices conducted in algorithmic auditing that are currently delineated in the literature often do not consider auditing a continuous process but mainly as a practice that is conducted either for certification or at the end of the development of an AI system to check its performance via internal auditing or by hiring external auditors. Conducting audits in a continuous collaboration can increase the quality of the system and, at the same time, enable compliance with the upcoming regulations for AI. To further address the gaps identified in this paper, further research must investigate the practices and the software tools that should be adopted throughout the lifecycle of AI systems, leading to the definition of novel lifecycle models for AI systems, wherein quality-assessments are more system and of which auditing is an integral part.

7.2 Including technical expertise in public institutions

Solving such a gap is essential to improve the governance ecosystem of AI, as well as, addressing the over-reliance of policymakers on private institutions for understanding the measures that must be mandated by the regulation. Therefore, to prevent the battle for talent, such collaboration must be carefully designed, to decentivise the delegation of the drafting of policies to private organisations aiming at their own agenda. Hence, governments and public institutions must support independent research to prevent the monopoly on such topics of research from big tech companies. Moreover, public institutions and the European jurisdiction must invest in building technical expertise internally, as well as, finding new ways to attract tech talent and compete with big tech salaries.

7.3 Developing the Infrastructure and Tools for Accredited Certification and harmonised Standards

Another topic that must be addressed is the necessity for accredited certification and well-drafted standards that provide operational guidance on the internal practices to audit and effectively assess the impact of AI systems. Unlike other highly-regulated industries, accreditation is still missing for AI and many loopholes must be solved to understand, which bodies will be responsible for such processes. Nonethe-

less, to develop accreditation mechanisms the field must advance, and more training and expertise have to be built to address the challenges of developing software and auditing mechanisms for complex data-driven models. This is an issue affecting also the standardisation of AI. Hence, due to the lack of operational guidance in the AI Act, harmonised standards can guide practitioners understand the requirements for compliance and the practices that must be adopted internally.

8 Conclusion

In this paper, the danger of over-reliance on regulation and top-down governance mechanisms for AI is demonstrated by analysing the current gaps that are present in translating such legal requirements into actionable measures.

The study identified a consensus on the need to operationalise the legal requirement of the AI Act and further develop an internal mechanism to understand how systems operate and record all the fundamental practices.

Hence, regulating AI will not be enough unless complementary tools are present throughout the lifecycle such as testing, documentation and assurance practices, auditing, and cross-disciplinary training. At the same time, the research unveiled novel policy-making methods that can further bridge the operational gaps in the field such as the development of regulatory sandboxes, policy-prototyping exercises, AI harmonised standards, and accredited certification. This research provides an overview of the current gaps between auditing and governance measures of AI through a series of interviews with experts working ad researching at this intersection. If auditing is to evolve into becoming a key mechanism for, not only developing ethical and trustworthy development of AI but also enabling legal compliance with the upcoming regulation, it is important to understand the operational gaps present at its intersection with the conformity assessment requirements of the upcoming legislation.

In order to address the aforementioned gaps, this research outlines recommendations for policymakers and practitioners that will enable further research in the AI auditing and regulation ecosystems. These include: 1) mandating audits through the entire lifecycle, 2) including technical expertise in public institutions, and 3) developing the infrastructure and tools for accredited certification and harmonised standards.

Acknowledgements

We would like to thank all the researchers and experts that made this research possible by sharing their invaluable insights at this complex intersection. We would like to express our sincere gratitude to Prof. Dr. Urs Gasser for his valuable guidance throughout the research process. We sincerely appreciated the reviewers' suggestions and feedback, which significantly helped us to improve the quality of the paper. Our sincere appreciation goes to Dr. Alexandros Paraschos' feedback and help in finalising the paper.

References

[1] Aditya Ramesh et al. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* (2022).

[2] Danton S Char, Nigam H Shah, and David Magnus. "Implementing machine learning in health care—addressing ethical challenges". In: *The New England journal of medicine* 378.11 (2018), p. 981.

[3] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. "Recommender systems and their ethical challenges". In: *Ai & Society* 35.4 (2020), pp. 957–967.

[4] Benjamin Lange and Theodore M Lechterman. "Combating disinformation with AI: Epistemic and ethical challenges". In: *2021 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE. 2021, pp. 1–5.

[5] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[6] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[7] Inioluwa Deborah Raji et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 33–44.

[8] Alexander Amini et al. "Uncovering and mitigating algorithmic bias through learned latent structure". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 289–295.

[9] Javier Rando et al. "Red-Teaming the Stable Diffusion Safety Filter". In: *arXiv preprint arXiv:2210.04610* (2022).

[10] Justin B Biddle. "On predicting recidivism: Epistemic risk, tradeoffs, and values in machine learning". In: *Canadian Journal of Philosophy* 52.3 (2022), pp. 321–341.

[11] Carina Lewandowski. "Machine (Un) learning: An Investigation of Racial Bias in Predictive Recidivism Algorithms as a Product of Real-World, Structural Discrimination". In: (2021).

[12] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[13] Heidi Ledford. "Millions of black people affected by racial bias in health-care algorithms". In: *Nature* 574.7780 (2019), pp. 608–610.

[14] European Commission. "Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts". In: *EUR-Lex-52021PC0206* (2021).

[15] House of Commons of Canada. "Bill C-27: An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts". In: *1st Session, 44th Parliament, 70-71 Elizabeth II, 2021-2022* (2022).

[16] Matt O'Shaughnessy and Matt Sheehan. "Lessons From the World's Two Experiments in AI Governance". In: (2023).

[17] White House Office of Science and Technology Policy. "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People." In: *https://www. whitehouse.gov/ostp/ai-bill-of-rights/* (2022).

[18] Johann Laux, Sandra Wachter, and Brent Mittelstadt. "Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act". In: *Available at SSRN* (2023).

[19] Anna Jobin, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1.9 (2019), pp. 389–399.

[20] Thilo Hagendorff. "The ethics of AI ethics: An evaluation of guidelines". In: *Minds and Machines* 30.1 (2020), pp. 99–120.

[21] Luciano Floridi. "Establishing the rules for building trustworthy AI". In: *Nature Machine Intelligence* 1.6 (2019), pp. 261–262.

[22] Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

[23] Jean-Marie John-Mathews, Dominique Cardon, and Christine Balagué. "From reality to world. A critical perspective on AI fairness". In: *Journal of Business Ethics* 178.4 (2022), pp. 945–959.

[24] Lionel P Robert Jr et al. "Introduction to the special issue on AI fairness, trust, and ethics". In: *AIS Transactions on Human-Computer Interaction* 12.4 (2020), pp. 172–178.

[25] Nagadivya Balasubramaniam et al. "Transparency and explainability of AI systems: ethical guidelines in practice". In: *Requirements Engineering: Foundation for Software Quality: 28th International Working Conference, REFSQ 2022, Birmingham, UK, March 21–24, 2022, Proceedings*. Springer. 2022, pp. 3–18.

[26] João Figueiredo Nobre Brito Cortese et al. "Should explainability be a fifth ethical principle in AI ethics?" In: *AI and Ethics* (2022), pp. 1–12.

[27] Finale Doshi-Velez et al. "Accountability of AI under the law: The role of explanation". In: *arXiv preprint arXiv:1711.01134* (2017).

[28] Junaid Qadir, Mohammad Qamar Islam, and Ala Al-Fuqaha. "Toward accountable human-centered AI: rationale and promising directions". In: *Journal of Information, Communication and Ethics in Society* (2022).

[29] Ayodeji Oseni et al. "Security and privacy for artificial intelligence: Opportunities and challenges". In: *arXiv preprint arXiv:2102.04661* (2021).

[30] Karl Manheim and Lyric Kaplan. "Artificial intelligence: Risks to privacy and democracy". In: *Yale JL & Tech.* 21 (2019), p. 106.

[31] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. "Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1571–1583.

[32] Ben Shneiderman. "Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10.4 (2020), pp. 1–31.

[33] Miles Brundage et al. "Toward trustworthy AI development: mechanisms for supporting verifiable claims". In: *arXiv preprint arXiv:2004.07213* (2020).

[34] Saleema Amershi et al. "Software engineering for machine learning: A case study". In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE. 2019, pp. 291–300.

[35] Keith H Bennett and Václav T Rajlich. "Software maintenance and evolution: a roadmap". In: *Proceedings of the Conference on the Future of Software Engineering*. 2000, pp. 73–87.

[36] Chuanqi Tao, Jerry Gao, and Tiexin Wang. "Testing and quality validation for ai software–perspectives, issues, and practices". In: *IEEE Access* 7 (2019), pp. 120164–120175.

[37] Shahar Avin et al. "Filling gaps in trustworthy development of AI". In: *Science* 374.6573 (2021), pp. 1327–1329.

[38] Gregory Falco et al. "Governing AI safety through independent audits". In: *Nature Machine Intelligence* 3.7 (2021), pp. 566–571.

[39] Deep Ganguli et al. "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned". In: *arXiv preprint arXiv:2209.07858* (2022).

[40] Ben Shneiderman. *Human-Centered AI*. Oxford University Press, 2022.

[41] Nikiforos Pittaras and Sean McGregor. "A taxonomic system for failure cause analysis of open source AI incidents". In: *arXiv preprint arXiv:2211.07280* (2022).

[42] Lorenzo Russo et al. "Cybersecurity exercises: wargaming and red teaming". In: *Next Generation CERTs* 54 (2019), p. 44.

[43] Avi Rushinek and Sara F Rushinek. "Accounting software evaluation: hardware, audit trails, backup, error recovery and security". In: *Managerial Auditing Journal* 10.9 (1995), pp. 29–37.

[44] Aaron Yi Ding, Gianluca Limon De Jesus, and Marijn Janssen. "Ethical hacking for boosting IoT vulnerability management: A first look into bug bounty programs and responsible disclosure". In: *Proceedings of the Eighth International Conference on Telecommunications and Remote Sensing*. 2019, pp. 49–55.

[45] Amanda Askell, Miles Brundage, and Gillian Hadfield. "The role of cooperation in responsible AI development". In: *arXiv preprint arXiv:1907.04534* (2019).

[46] Shin-Shin Hua and Haydn Belfield. "AI & Antitrust: Reconciling Tensions between Competition Law and Cooperative AI Development". In: *Yale JL & Tech.* 23 (2020), p. 415.

[47] Margaret Mitchell et al. "Model cards for model reporting". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.

[48] Timnit Gebru et al. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[49] Matthew Arnold et al. "FactSheets: Increasing trust in AI services through supplier's declarations of conformity". In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 6–1.

[50] John Richards et al. "A methodology for creating AI FactSheets". In: *arXiv preprint arXiv:2006.13796* (2020).

[51] Inioluwa Deborah Raji and Joy Buolamwini. "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 429–435.

[52] Adriano Koshiyama et al. "Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms". In: *SSRN Electronic Journal* (2021).

[53] Jenna Wiens, W Nicholson Price, and Michael W Sjoding. "Diagnosing bias in data-driven algorithms for healthcare". In: *Nature medicine* 26.1 (2020), pp. 25–26.

[54] Matti Minkkinen, Joakim Laine, and Matti Mäntymäki. "Continuous auditing of Artificial Intelligence: A Conceptualization and Assessment of Tools and Frameworks". In: *Digital Society* 1.3 (2022), p. 21.

[55] Jakob Mökander et al. "Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation". In: *Minds and Machines* 32.2 (2022), pp. 241–268.

[56] Ellen P Goodman and Julia Trehu. "AI Audit Washing and Accountability". In: *Available at SSRN 4227350* (2022).

[57] Financial Reporting Council. *Auditors I Audit and Assurance I Standards and Guidance for Auditors I Financial Reporting Council (2020)*. 2021.

[58] Ioannis N Kessides. "Powering Africa's sustainable development: The potential role of nuclear energy". In: *Energy Policy* 74 (2014), S57–S70.

[59] Pedro Saleiro et al. "Aequitas: A bias and fairness audit toolkit". In: *arXiv preprint arXiv:1811.05577* (2018).

[60] Vijay Arya et al. "AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models." In: *J. Mach. Learn. Res.* 21.130 (2020), pp. 1–6.

[61] Rachel KE Bellamy et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias". In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1.

[62] Vijay Arya et al. "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques". In: *arXiv preprint arXiv:1909.03012* (2019).

[63] Harsha Nori et al. "Interpretml: A unified framework for machine learning interpretability". In: *arXiv preprint arXiv:1909.09223* (2019).

[64] Alexander D'Amour et al. "Fairness is not static: deeper understanding of long term fairness via simulation studies". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 525–534.

[65] Sarah Bird et al. "Fairlearn: A toolkit for assessing and improving fairness in AI". In: *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[66] Michael Katell et al. "Toward situated interventions for algorithmic equity: lessons from the field". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 45–55.

[67] Hong Shen et al. "The model card authoring toolkit: Toward community-centered, deliberation-driven AI design". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 440–451.

[68] Jaber F Gubrium and James A Holstein. "From the individual interview to the interview society". In: *Handbook of interview research: Context and method* (2002), pp. 3–32.

[69] Jane Ritchie and Liz Spencer. "Qualitative data analysis for applied policy research". In: *Analyzing qualitative data*. Routledge, 2002, pp. 187–208.

[70] Jane Ritchie et al. *Qualitative research practice: A guide for social science students and researchers*. sage, 2013.

[71] Judy Goldsmith and Emanuelle Burton. "Why teaching ethics to AI practitioners is important". In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[72] Nicola K Gale et al. "Using the framework method for the analysis of qualitative data in multi-disciplinary health research". In: *BMC medical research methodology* 13.1 (2013), pp. 1–8.

[73] Laurie J Goldsmith. "Using Framework Analysis in Applied Qualitative Research." In: *Qualitative Report* 26.6 (2021).

[74] Will Douglas Heaven. *Why Meta's latest large language model survived only three days online*. 2022.

[75] Laura Weidinger et al. "Ethical and social risks of harm from language models". In: *arXiv preprint arXiv:2112.04359* (2021).

[76] Martin Ebers. "Standardizing AI-The Case of the European Commission's Proposal for an Artificial Intelligence Act". In: *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics* (2021).

[77] Ramak Molavi Vasse'i. "The Ethical Guidelines for Trustworthy AI–A Procrastination of Effective Law Enforcement". In: *Computer Law Review International* 20.5 (2019), pp. 129–136.

[78] Carol AB Warren. "Qualitative interviewing". In: *Handbook of interview research: Context and method* 839101 (2002), pp. 103–116.

[79] Pranee Liamputtong. *Handbook of research methods in health social sciences*. Springer, 2019.

[80] Alan Bryman, Robert G Burgess, et al. *Analyzing qualitative data*. Vol. 11. Routledge London, 1994.

A Methodology

A.1 Data Collection

In this paper, the method of theoretical sampling has been selected. Theoretical sampling differs from other approaches, because it requires the researcher to look for potential interviewees who are required to possess specific characteristics of interest for the research [78]. To fulfil the objectives of this research, interview candidates needed to be experts working or researching in either academia, public institutions or private companies at the intersection of internal practices for compliance, auditing and regulation of AI. This requirement is particularly important, to enable drawing the patterns and identifying the gaps among the respondents. In particular, due to the novelty of the research at this intersection of AI regulation and auditing, the "snowball" technique was applied, to find more respondents through the social connections of the interview candidates. Setting the sample size to 7 individuals was a choice driven by the principle of "saturation" [79], for which the size of a sample can be viewed as satisfactory, once the facts emerging from the interviews become repetitive [79]. Furthermore, in order to safeguard the privacy of the interview candidates, all respondents were invited to sign a consent form before the interview, which informed them about being recorded, the purpose of the interview, and the use and transcription of data.

A.2 Framework Analysis Steps

The qualitative analysis follows the framework methodology, which consists of five steps: 1) data familiarisation; (2) thematic framework identification; (3) indexing; (4) charting; (5) mapping and interpretation.

Step 1: Data Familiarisation After conducting the interviews, the first step necessary is to identify the major themes in the data, and the answers to the research questions recurring through the data [69, 70, 73]. During this phase, the authors have to familiarise with the transcripts of the interviews conducted, to become aware of the recurring themes and to note them.

Step 2: Framework Identification Once the authors process the selected material, the key issues and recurring themes are identified, which will then enable to examine and reference the data [69]. The second phase consists of identifying the concepts that would provide a structure, for the analysis and its interpretation and builds on a combination of a-priori and emergent concepts [69, 73]. The concepts and recurring themes have to be ordered, in a way that enables addressing the focus of the study, by dividing the framework, into key themes (KT) and elaborating the sub-themes (ST) [73]. The identification of the framework, is not a mechanical process, but it requires making judgements about the content of the data, the relevance of certain themes, and making sure that the research question is addressed[69]. The framework is based on an iterative process and is initially tested on, both the interview transcripts and the literature review, then later, it is modified in the analysis phase, to enable transitioning from straightforward descriptions to conceptual abstractions [69, 70].

Step 3: Indexing After identifying the framework the next step is to link it systematically to all the data [70]. This stage is facilitated by the use of the software MAXQDA, in order to process the data more efficiently. The indexing step is fundamental to review the framework and understand whether it applies to the whole data as well amending the definitions of all key themes and sub-components identified. In this phase, the key themes remained the same, but the authors re-named them to provide more clarity to the reader. For instance, Key Theme 1 changed from "Best practices" to "State of the art and best practices". The authors identified new sub-themes, through the processing of the data, such as ST 1.2 "Corporate Social Responsibility", ST1.3 "Testing" and other sub-themes have been deleted, due to the lack of data explaining and analysing such topic, such as ST 1.1 "Fairness". In the processing phase of Key Theme 2, new sub-themes emerged such as ST 2.1 "Risk classification", ST 2.3 "Risk of becoming check-list regulation," and ST 2.5 "Disadvantages for SMEs". In Key Theme3. new topics emerges such as, ST 3.3 "Assurance practices", ST 3.5 "Cross-disciplinary training and collaboration", ST 3.6 "Clarifying the terminology in the AI Act," and ST 3.7 "Tooling and systemic practices". In conclusions, for Key theme 4, the ST 4.4 "Whistle-blower protection" emerged and ST 4.6 "Digital Hubs".

Step 4: Charting The next step, for the framework analysis, consists of ordering and abstracting the indexed data to systematically proceed with its analysis [71]. In this phase the authors create several charts, summarising the study data [71]. The authors organised the data in a matrix form, using the rows to delineate the key themes and sub-themes and the columns to represent each interviewee response to the theme. Because of the large volume of the data, as the recordings consisted of 703 minutes, the data had to be processed and inserted into the chart according to the themes and sub-themes identifies in the indexing phase. This phase was also supported

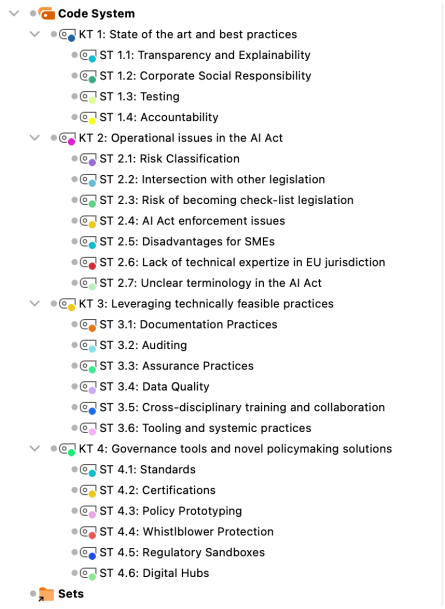

Figure 1: *Own Representation of the Index*

by the use of the tool MAXQDA, which generated a "quote matrix", containing all the data for the charting. Nonetheless, the authors had to clean the data and make it more readable, by breaking it down into sub-charts, one for each key theme. The summaries are represented in two-dimensional ((number of interviews=10)*(number of sub-themes)) matrices. To every cell, a code is assigned, such as (A 11, B 11), representing what a specific experts said about a sub-theme.

Step 5: Mapping and Interpretation The final step of the framework analysis consists in interpreting the key lessons and themes of the earlier phases [69, 73].

Emergent themes and topics, patterns and associations are usually noted already, in the indexing and charting phase of the analysis [80]. Therefore, this final phase enables to systematically address the key objective of the qualitative analysis [80]. The analysis proceeds with defining key concepts and providing explanations through the in-depth analysis and interpretation of the data. There are multiple ways to display the results of the mapping and interpretation phase, such as explaining attitudes and experiences, creating typologies or identifying the key patterns and concepts of a particular phenomena [69, 73, 72]. The authors choose to identify and describe the key themes that emerged with their respective sub-themes.

B Interview guideline

Hereby attached is the excerpt of the interview guideline. The purpose of the guideline is to guide the author and the experts on the topics covered and investigated. The interview questions are not shared with the interviewees before the interview.

Section 1: Exploring the interviewee role in the field of AI Ethics

1.1 Please describe your role and work within your organisation or institution.
1.2 How are you addressing the field of AI auditing and/or law compliance within the domain of your organisation or institution?
1.3 What are the current issues in the field of AI ethics on which you are focusing and which practices are you exploring within your role?

Section 2: Understanding the interviewees work in actionable measures for ethical AI

2.1 Which areas of the research of the field of AI ethics do you focus on and what are the most promising future directions?
2.2 What do you think the AI ethics debate is missing and needs to address urgently?
2.3 How are companies attempting to assess the impact of the AI systems they develop and deploy?
2.4 Within your domain, are you aware if companies are adopting or developing internal auditing practices?
2.5 If not, why? Is auditing even considered by the management teams and what are the barriers that prevent such work?
2.6 If you are not aware of any methods and tools, which areas of the research of the field of AI do you focus on and what are the most promising future directions?
2.7 Let me give you some example practices, do you know about them (data-sheets, model cards, factsheets) do you have any experience with them? Why do you think one is more helpful than the other?
- If not, why? Do you have any plans to adopt them in the future? Is there any discussion on adopting any in the future? If yes, which one?
- If yes, how do you document the findings? Who has access to this? Can I also get access?

Section 3: Understanding the interviewees opinion on the best practices in the field of auditing in light of the compliance with the upcoming regulation

The European Commission is working on the AI Act, which is expected to be enforced in a couple of years

3.1 Does your organisation or institution follow the latest policy development in the EU or globally?

3.2 If yes, how are you working in understanding how to operationalize the conformity assessment requirements?
If not, why? Do you believe that your company will not be affected?

3.3 How do you see the AI Act contributing to the field of AI ethics?

3.4 Many believe that the AI Act is hard to implement, that operational guidance is lacking, what do you think?

3.5 What can be done to make the AI Act more operational? Which tools or methods do you think are needed to help understand which actions can enable compliance with the law ?

3.6 If you are not aware of any practice that seems promising, what do you think it is necessary to adopt in light of the regulatory requirements?

3.7 What are the shortcomings of the AI Act in your opinion?

3.8 Within your role, how aware are companies about the upcoming AI regulation?

3.9 How is your role contributing to address such gap?

3.10 What expertise is missing in the field and how do you think that can be addressed?