EUI SCHOOL OF TRANSNATIONAL GOVERNANCE

How can AI Regulation be Effectively Enforced?

Comparing Compliance Mechanisms for AI Regulation with a Multiple-Criteria Decision Analysis

Lukas Tiberius Wiehler

Thesis submitted for assessment with a view to obtaining the degree of Master of Arts in Transnational Governance of the European University Institute

Florence, 20 May 2022

European University Institute
School of Transnational Governance

How can AI Regulation be Effectively Enforced?

Comparing Compliance Mechanisms for AI Regulation with a Multiple-Criteria Decision Analysis

Lukas Tiberius Wiehler

Thesis submitted for assessment with a view to obtaining
the degree of Master of Arts in Transnational Governance
of the European University Institute

Supervisor

Professor Andrea Renda, School of Transnational Governance

Student declaration to accompany the submission of written work
School of Transnational Governance

I Lukas Tiberius Wiehler certify that I am the author of the work 'How can AI Regulation be Effectively Enforced? Comparing Compliance Mechanisms for AI Regulation with a Multiple-Criteria Decision Analysis' I have presented for examination for the Master of Arts in Transnational Governance. at the European University Institute. I also certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, that it is the work of others.

I warrant that I have obtained all the permissions required for using any material from other copyrighted publications.

I certify that this work complies with the Code of Ethics in Academic Research issued by the European University Institute (IUE 332/2/10 (CA 297).

I declare that this work consists of 10361 words.

Signature and date:

Lukas Tiberius Wiehler, 20/05/2022, Firenze

ABSTRACT:
Newly emerging AI regulations need effective and innovative enforcement and compliance mechanisms to assure that fundamental and human rights are protected when using an AI system. This study compares four different compliance mechanisms namely 'Real-Time and Automated Conformity Assessment', 'Standardization and Certification', 'Algorithmic Impact Assessment' and 'Algorithmic Auditing' as well as three different assurers of compliance namely deployers, notified bodies and civil society organisations. With an MCDA, this research has shown that civil society-based compliance mechanisms are believed to be less effective, less feasible and more costly compared to all other compliance mechanisms. Second, external compliance mechanisms (by notified bodies) were rated to be more effective but also more difficult to implement compared to internal compliance mechanisms. Third, algorithmic auditing scored highest among all policy options. Fourth, despite its experimental nature, automated and real-time compliance mechanisms are not scored significantly lower than other compliance mechanisms.

1. Introduction

AI systems– on the market today and yet to be developed– may pose severe threats to human and fundamental rights**.** Legislators are thus challenged to design ambitious regulations controlling the risks of these new technologies. Around the world, policymakers have developed and are developing principles and guidelines constituting inter alia that high risk AI systems should be non-maleficence, beneficial, autonomous, just and explicable (Floridi & Cowls, 2019). However, aside from high-level principles for ethical AI, comprehensive legal frameworks should establish how such guidelines will be enforced (Mökander et al., 2021). In this sense, a solid AI regulation also establishes compliance and enforcement mechanisms that are effective, feasible and cost-efficient for government and deployers. As the regulation of AI is a relatively new field of scholarship only some attention has been paid to compliance mechanisms for AI systems(Mökander et al., 2021). Just as AI regulation generally, AI regulation's compliance mechanism remains thus a partially exploratory and uncertain policy field.

This research aims to shed new light on this policy domain by comparing the main proposals for compliance mechanisms with a Multi-Criteria Decision Analysis (MCDA) that is based on expert scoring of policy options. It will be investigated which compliance mechanism is potentially most or least effective, feasible and cost-efficient. MCDA is an ideal tool for decision making, especially when the decision environment is uncertain, and the assessment criteria are multiple and complex. The analysis promises to be insightful as to the best of the author's knowledge this method has never been used for AI regulation or cyber governance in general. With the MCDA three assurers of compliance (actors executing the compliance mechanisms) are compared to each other, namely the deployers, notified bodies and civil society organisations. Further, four different proposed policies for compliance mechanisms are ranked by the MCDA namely 'Real-Time and Automated Conformity Assessment', 'Standardisation and Certification', 'Algorithmic Impact Assessment' and 'Algorithmic Auditing'.

Before conducting the MCDA, AI systems will be defined and the risks and needs for rigorous regulation of these systems will be outlined. Subsequently, this paper will ponder upon the importance of solid enforcement and compliance mechanism. Third, compliance mechanisms for AI regulation as well assurers of enforcement along the AI lifecycle will be characterized. Next,

the three assurers and four compliance mechanisms selected for this MCDA will be thoroughly introduced and analysed. Fifth, this research will specify all steps of the MCDA. Finally, the results of the MCDA will be outlined and discussed.

2. Literature Review

2.1 AI Systems, their Risks & the Need for Regulation

2.1.1 AI Systems and their Promises

AI systems employed today and in future will and already do shape all aspects of human societies. The decisions, recommendations and judgments performed by those systems – previously in hands of human experts– will have a significant impact on humans and their environments (Renda, 2019; Spielkamp, 2019; Turchin & Denkenberger, 2020). By AI system this paper understands– in line with the originally proposed definition of the European Commission (European Commission, 2021) –broadly "a system that receives machine and/or human-based data and inputs, infers how to achieve a given set of human-defined objectives using learning, reasoning or modelling implemented with a variety of the techniques and approaches [e.g. machine learning, logic and knowledge-based approaches, etc] generates outputs in the form of content (generative AI systems), predictions, recommendations or decisions, which influence the environments and human beings." Innovation in the field of AI may be helpful in virtually all domains of human life. It is often associated with increased consistency, efficiency, and advanced answers to complex multi-variate problems (Taddeo & Floridi, 2018) that offers "major opportunities to improve our economic, societal and environmental wellbeing" (Walsh et al., 2019, p.5). According to Mökander and Axente (2021), these benefits are associated with the systems' relative autonomy, complexity, and scalability.

2.1.2 AI Systems and their Risks

Relative autonomy, complexity, and scalability, these attributes may also result in adverse negative effects on several fundamental and human rights (Access Now, 2018; European Union Agency for Fundamental Rights, 2020; Global Partners Digital, 2020). AI systems may facilitate and reproduce biases and discrimination (Mattu, 2016; QC & Dee, 2020; Review, 2019), erode human

agency and autonomy, negatively affect the freedom of expression (Solaiman et al., 2019), the right to free election and interfere with data protection as well as the right to privacy and family life. In governments, it may further impact good administration, access to a fair trial and justice (UC Berkeley, 2019). AI systems may hamper consumer protection, the right to freedom of assembly and association, as well as sustainability and protection against sustained impairment of the living standards of future generations. Finally, it may lead to online addiction (Alter, 2017), the detrition of gainful and humane working conditions (Frey & Osborne, 2017; Gray & Suri, 2019) and negatively affect vulnerable and marginalized groups such as migrants (CEPS et al., 2021). Concerning the above-mentioned risks, it needs to be emphasized, however, that not all AI systems will pose a significant risk, some simple systems (like toys or chatbots) may only pose low or no risk. Thus, this research is concerned with AI systems that pose a high risk to human and fundamental rights.

2.1.3 AI Systems and the Need for Regulation

Considering these potentially hazardous effects in today's information societies, threats to fundamental rights need to be addressed by regulations to ensure public trust in the systems as well as in public regulators (Ebers et al., 2021). Regulating AI systems poses unique challenges to regulators because the inputs, outputs and operations are often opaque to users and regulators (Rai, 2020) but also to deployers (AI systems as "black box"). Regulation of AI is further considered difficult and risky because of the autonomous, uncertain (Parker, 2012) and unforeseeable nature of AI systems which is arguably grounded in its inherent non-determinism (Scherer, 2015). Regulations are further complicated as AI -systems are data-intensive, continuously evolving, adjusting and self-adapting (Felderer & Ramler, 2021).

As most current laws, tools, and enforcement mechanisms were designed to merely oversee *human* decision-making and not *automated decisions,* they are fundamentally deemed insufficient to regulate AI systems (Mökander et al, 2021, Brundage et al., 2020). And while jurisdictions around the world are in the process to fill this regulation gap (Fanni et al., 2021), the AI systems evolve at a fast pace and may have outpaced the development of tools that review AI systems and assure their reliability and trustworthiness (Schulam & Saria, 2019).

A need to regulate AI systems may also be identified from the side of deployers, users and the affected public. A range of studies and surveys has identified a fundamental distrust amongst workers and the general public against AI systems (Dafoe & Oxford, 2019.; Edelman, 2019). Looking at the developers of AI systems, it can be argued that only binding legal guidelines, as well as external sanctions, will ensure rigid impact assessment. For example, in an interview-based study conducted by the EU Agency for Fundamental Rights (2020) most developers and deployers of AI systems stated that they are generally aware of fundamental rights issues, but admitted that since only data protection impact assessment was legally required (in the EU under GDPR; Lachaud, 2020), no impact assessment on fundamental rights was conducted. A meta-study further revealed that the mere existence of codes of ethics in software development is relatively meaningless if not enforced by the organization (Schell-Busey, 2022). Against this backdrop, lawmakers are challenged to design innovative, future-proof and effective regulations for AI systems.

2.1.4 Emerging Legal Frameworks for AI Systems:

Responding to the risks and challenges described above several international organisations are drafting or have agreed on high-level rules and ethical guidelines (Jobin et al., 2019) establishing standards for ethical, responsible and/or trustworthy AI systems (Gesley et al., 2019). Examples are the Recommendation of the Council on Artificial Intelligence (OECD, 2019), Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) and the Ethically Aligned Design guidelines (IEEE, 2019). According to Floridi & Cowls (2019), these guidelines – though different in detail – evolve around the same five principles: non-maleficence, autonomy, justice, explicability and beneficence. Similarly, states and supranational organisations have drafted guidelines for future AI regulation, for instance, the EU, Australia, Japan and Singapore. All of them commonly mention that AI should promote benefits, be human-centred (taking into account the needs and values of individuals and communities that interact with it), be fair, be explainable (understandable, transparent), secure, safe, reliable and accountable (Fanni et al., 2021).

While important to be defined, these principles will only be meaningful if translated into mid-level norms and subsequently detailed low-level requirements (Mittelstadt, 2019). Otherwise, they run the risk of either being semantically too narrow or too broad and flexible (Arvan, 2018). States

and supra- and international organisations are therefore drafting legislation that formalizes ethical guidelines into requirements. One of the first and most far-reaching examples in this regard is the EU's AI Act (from here on AIA). It is built on a risk-based approach and has translated the 'Ethical Guidelines on Trustworthy AI' (AI HLEG, 2019) into a list of essential requirements ('Title III, Chapter 2 AIA', *European Commission*, 2021) which are foreseen to be legally binding for high-risk AI systems. Other major economies are expected to publish drafts for comprehensive AI regulations soon, e.g. the UK is expected to publish the first draft of an AI regulation in the first half of 2022.

But even with settled high-level principles, guidelines as well as detailed norms and practical requirements in place, a central challenge to all newly emerging AI regulations is to translate them into practice. Hence, to assure the effectiveness of and compliance with given AI regulations, governments and public agencies need potent governance mechanisms to do so.

2.2 From Theory to Practice: Compliance, Enforcement, Assurance

2.2.1 An Introduction to Compliance Mechanisms

Governance mechanisms to enforce laws and control compliance with regulations are a central piece of every AI regulation. These governance mechanisms are in the following called *compliance mechanisms* (in line with the Ad hoc Committee on AI of the Council of Europe, CAHAI, 2020). Fundamentally, to be successful, every regulation needs to be linked to effective enforcement mechanisms, i.e., activities, structures, and controls wielded by various parties to influence and achieve normative ends (Baldwin et al., 1999). Compliance mechanisms for AI regulation may be defined as the process by which the expected function of an AI system as well as its conformity with laws, standards and regulations (national, international as well as privately set standards) is checked, controlled for and demonstrated to others (US GAO, 2021; CDEI, 2021). Compliance mechanisms thereby provide trustworthy information on the reliability, fairness, safety and performance as well as potential risks of the AI system (CAHAI, 2020). Practical mechanisms for compliance should help regulators, agencies, and users to monitor and understand if an AI system is adhering to a legal framework, provide transparency, and ensure accountability, expandability, and legibility (CAHAI, 2020). Further, compliance mechanisms should take a

holistic view, taking into account the computational model but also inputs, outputs, the operational process as well as the larger socio-technical context of the system (Mökander & Axente, 2021)

A concept closely related if not equivalent to compliance mechanisms is AI assurance (Kazim & Koshiyama, 2020; Leslie et al., 2022). This terminology is dominantly used in the UK context, where it is defined as "the systematic examination of the extent to which a software product is capable of satisfying stated and implied needs" (Felderer & Ramler, 2021, p.13). Correspondingly, the UK is expected to introduce an 'AI assurance ecosystem' alongside its awaited AI regulation (CDEI, 2021). Scholars like Brundage et al. (2020) further speak of *mechanisms for supporting verifiable claims,* a notion that may also be conceptually linked to compliance mechanisms. They argue that in contrast to mathematical claims about an AI system that may be verified with 100% certainty, general claims about AI systems (e.g., the system is fair) may be verified only to the degree that evidence could be collected. Here, Brundage et al (2020) think of socio-technical claims e.g., the non-biasness of a data set.

2.2.2 Effective Compliance Mechanisms

Compliance mechanisms may further be differentiated into the ex-ante and ex-post assessment of conformity. Ex-ante refers to the assessment of compliance with the legal framework before being deployed on the market. However, given the continuously evolving, adjusting and self-adapting nature of AI systems, the ex-post compliance mechanism, i.e., the market monitoring is of high significance. It may further be argued that the ex-post compliance mechanism should equally be able to evolve and adapt to the changing nature of AI systems. In this way, compliance mechanisms should be designed to assure legal conformity throughout the entire lifecycle of the AI system (Mökander et al., 2021).

Thoroughly implemented compliance mechanisms should render unethical behaviour impossible. Accordingly, the existence of well-intended but under-enforced ethical requirements has in the past arguably led to unethical behaviour (Floridi, 2019) like 'ethics- blue washing' (making unsubstantiated ethical claims about an AI system), ethics- lobbying (ethical standards that are used to justify the inexistence of necessary legally binding legislation) or ethics-shopping (pick and choose standards from different sources to suit pre-existing unethical behaviour). Therefore,

without compulsory compliance mechanisms, deployers may be incentivised to practice ethics washing (Gibney, 2020)

Compliance mechanisms remain a developing policy field. The translation of high-level guidelines and principles into practical guidance – the enforcement of AI regulation, the check for and the demonstration of compliance and function – still needs to be explored, developed and neatly designed. For example, Raji et al. (2020) hold that both governments and industries lack the tools to translate the general guidelines into verifiable criteria. Checking, implanting and monitoring ethical guidelines is also seen as a big challenge for businesses by PwC (PwC, 2019) For Mökander et al.(2021) thus while the question of 'what' - what are the ethical principles for AI systems – is more or less solved, the 'how' question – how should the principles be enforced and compliance be assured – remains to be answered.

2.2.2 Proposed Compliance Mechanism: A Literature Review

A wide variety of compliance mechanisms has been proposed by regulators, scholars and AI developers in the field. They focus on different aspects and components of the AI systems (e.g., training data, code, output, etc.) and may show different efficiency in different contexts (e.g. different regulatory cultures). This research has identified four main approaches to AI compliance that may be compared in the Multi-Criteria Decision Analysis, namely Algorithmic Auditing, Algorithmic Impact Assessment, Standards and Certification as well as Real-Time and Automated Conformity Assessment. These four compliance mechanisms will be defined and analysed in more detail below. Beyond these four main governance mechanisms, there is a range of proposed compliance mechanisms deserving mention. Due to the scope of this research, they could however not be considered in the MCDA. It may be differentiated between compliance mechanisms in the development phase, in the ex-ante pre-market phase and the ex-post market monitoring phase.

In the development phase, compliance mechanisms are for example awareness-raising with software engineers (Floridi et al, 2018), diversifying the developer teams (Sánchez-Monedero et al., 2020), and a proactive design that integrates ethical guidelines from the very beginning, e.g reflective design, Values@Play, and Value-Sensitive Design (Aizenberg & van den Hoven, 2020; IEEE, 2019; Morley et al., 2020; van de Poel, 2020). However, this compliance mechanism has to

be viewed with caution as studies have identified difficulties to embed values into the design throughout the technology cycle (van de Poel, 2020). Further tools in the development stage are the review of potentially biased input data (AI Ethics Impact Group, 2020), employing only codes and decision models that have previously been verified (Dennis et al., 2016) or licensing developers allowed to code high-risk AI systems (Mittelstadt, 2019).

For pre-market ex-ante compliance mechanisms, an interesting tool is the regulatory sandbox (e.g. Leslie et al., 2022). Regulatory sandboxes may be used to test innovative systems or products for safety and compliance in a safe and delimited setting before they enter the market (Makarov & Davydova, 2020). The idea is the reduce the administrative burden, time-to-market and lower the cost for the organisation and in this way enable innovation without compromising on compliance with high-level principles (Ranchordas, 2021). Though the details are not fleshed out, regulatory sandboxes are also foreseen by the EU's AIA (EU Commission, 2021). For ex-post market monitoring, a potential compliance mechanism discussed is the so-called 'human in the loop protocol' that is supposed to oversight the AI systems, potentially intervene once harmful outcomes are detected or even might be held responsible (Jotterand & Bosco, 2020; Rahwan, 2018; R. Fanni et al., 2020).

Aside, from 'hard' compliance mechanisms, governments may also incentivise a cultural shift changing the way societies interact with new technologies. For example, by changing school curriculums to train digital literacy or by incentivising and supporting a fruitful culture for whistleblowing or investigative (data) journalism (European Parliament Think Tank, 2019). Lastly, high fines may further deter developers from (intentional) violating AI regulation (European Commission, 2021). Fines, however, are only functional in intersection with other compliance mechanisms as inconsistency with predefined principles must be identified. Compliance mechanisms are hence crucial to determine and fine potential misconduct.

2.2.3 Assurer of the System:

Legal frameworks aiming to regulate AI systems should also clearly define and empower actors so-called assurers of compliance to assess conformity and provide oversight (CAHAI, 2020). For effective enforcement, these assurers should be organisationally independent of the developers and

deployers of the system (Council of Europe, 2019). A range of actors may be thinkable to be assurers of compliance for instance expert committees, academics, sectoral regulators, auditing firms, dedicated corporate units for internal checks, civil society organisations or private sector auditors. Fundamentally, assurers should have sufficient expertise, competencies and resources. They may (but not necessarily need to) have intervening powers (CAHAI, 2020). With this research, three main potential assurers are compared namely deployers (internal compliance mechanism), notified bodies (external compliance mechanism) and civil society organisations (bottom-up compliance mechanism).

2.3 Policy Options for the Multi-Criteria Analysis

2.3.1 Possible Compliance Mechanisms

2.3.1.1 Algorithmic Auditing

AI Auditing is seen as one of the most promising tools to translate vague concepts, rules, and guidelines into a practically and effectively enforced legal framework (Bauer, 2017; Brundage et al., 2020; Kim, 2017; Morley et al., 2020). An Algorithmic Audit may be seen as a regulatory inspection used to assess whether an algorithmic system complies with "AI Regulation, data protection law, equalities legislation, or insurance industry requirements, for instance" (Lovelace & DataKind, 2020, p.12). Unlike impact assessment, algorithmic audits look at the technical entirety of the algorithmic system and its lifecycle (Raji & Buolamwini, 2019).

Auditors of the AI systems need significant access and statutory powers to audit the system. Therefore, audits can only be performed by actors who work in close collaboration with developers (e.g., auditing professionals) or by public agencies having regulatory authority. However, data gathering power and access should be secure and well delimited (Lovelace & DataKind, 2020). Aside from examining code (which is controversial and may offer only limited or slow information), potential auditing tools are e.g. "techniques from bias auditing, [...] mandating access to data about the algorithm's users, inspecting how the system is operating, speaking with developers or users, or looking at models underpinning an algorithmic system" (Lovelace & DataKind, 2020, p.4). Algorithmic auditing may be both employed ex-post and ex-ante.

Past literature has established that AI auditing may prove suitable for identifying a range of harmful effects of AI systems (e.g., discrimination, distortion, exploitation, misjudgement) with different types of auditing tools (Bandy, 2021). In this way, AI audits promise to operationalise, "assess and assure the legality, ethics, and safety of [AI systems]" (Mökander, Morley, et al., 2021). For instance, Bandy (2021) has reviewed studies on auditing and found examples where audits have revealed problematic behaviours, e.g., search algorithms capable of distortion and advertisement algorithms culpable of discrimination. Raji & Buolamwini (2019) examine a successful case of AI auditing, namely the auditing tool "Gender Shades" that checks for biases in gender and skin shades and subsequently documents how developers react to the findings, i.e., adjust and reduce the biases. Aside from these singular examples, however, there is no standardised approach on how to conduct audits. Also, given the wide range of different types of AI systems, algorithmic audits differ depending on sector and usage (Lovelace & DataKind, 2020).

Based on auditing experience in other sectors, two major lessons can be drawn. First, auditing differs from a code of conduct in that it is purpose-oriented and "aims to demonstrate adherence to a predefined baseline" (ICO, 2020). Second, operational independence between auditor and auditee is a precondition for a successful audit. This means that even if we speak of 'internal audit' an organisational separation between developers and auditors is presupposed (Floridi et al., 2022).

2.3.1.2 Algorithmic Impact Assessment:

In contrast to algorithmic auditing which requires significant technical insights into the AI systems, algorithmic impact assessment is less technical as it takes a broader social, environmental, legal or ethical perspective (Kazim & Koshiyama, 2020; Koshiyama & Engin, 2019). Algorithmic impact assessment may be differentiated between Algorithmic risk assessments which are conducted ex-ante and Algorithmic impact evaluations which are conducted ex-post (Lovelace & DataKind, 2020).

2.3.1.2.a. *Algorithmic Risk Assessments:*

Algorithmic risk assessments are deployed to identify and assess the potential risks of an AI system as well as anticipated impacts before the system is deployed. With algorithmic risk assessment, a

holistic perspective is aspired to look beyond the computational model and thus at societal, social, environmental, legal, and ethical impacts i.e. how users will interact with or be affected by it (McGregor et al., 2019). In this way, impact assessments e.g. look at the expected (societal) outcomes of the AI systems and propose means of mitigating risks and concerns (ECP, 2019). So far, impact and risk assessments are used mostly by the public sector e.g. for human rights, environmental regulation or data protection (Reisman et al., 2018). However, future users may also be creators, procurers or deployers. The EU's General Data Protection Regulation (Kaminski & Malgieri, 2019) includes a risk assessment that may partly serve as a blueprint for AI risk assessment (Kaminski & Malgieri, 2019).

Examples of already used frameworks for algorithmic risk assessment include the Canadian Government's algorithmic impact assessment (Canadian Government, 2020) for the public sector's use of AI systems. In this case, the risk assessment is implemented with a mandatory online questionnaire asking e.g. about the stakes of the decisions, type of technology, project motivation and the vulnerability of the users (Lovelace & DataKind, 2020). As an outcome, the assessment generates a risk level and subsequently establishes mandatory requirements (Treasury Board of Canada, 2021). Further examples include the AI Now Institute's algorithmic impact assessment, called the 'Human Rights, Democracy, and the Rule of Law Impact Assessment' (Leslie et al., 2022) or the 'Human Rights, Ethical and Social Impact Assessment'-HRESIA proposed by Mantelero (2018).

2.3.1.2.b *Algorithmic Impact Evaluation:*

Algorithmic impact evaluation monitors the AI system and its impact on a population after its deployment (ex-post). In this way, algorithmic impact evaluation parallels conventional policy or economic impact assessments that rate processes when operative (Kazim & Koshiyama, 2021). In some jurisdictions, these evaluations are part of the regular implementation process of a policy, e.g. for the EU's DG Internal Market and Services (Fitzpatrick, 2012). Like algorithmic risk assessment, the impact assessment takes a contextual, social and cultural perspective and draws amongst others from assessment frameworks in Science and Technology Studies (Konrad et al., 2017).

A variety of practical tools are available to assess impact including checklists (e.g. on data protection; EU Comission, 2016), lists of questions (e.g. the Assessment List for Trustworthy AI for self-assessment; AI HLEG, 2020) or self-assessment tool like the 'Human rights compliance assessment quick check' (Danish Institute for Human Rights, 2016). Impact evaluations can be conducted by independent researchers, governments or public agencies, though may need some access to system information, such as details of people subject to the system. Examples of a framework for algorithmic risk assessment include the 'Human Rights, Democracy, and Rule of Law Assurance Case' (HUDERAC) proposed by the Alan Turning Institute (Leslie et al., 2022) or the Stanford's 'Impact evaluation of a predictive risk modelling tool for Allegheny County's Child Welfare Office'(Goldhaber-Fiebert & Prince, 2019)

As impact evaluations are conventionally not directly conducted by deployers, they may raise questions about accountability because developers are not necessarily obliged to enact the recommendations. An example of this problem may be Facebook's failure to react to a human rights impact evolution in Myanmar (Latonero & Agarwal, 2021; Warofka, 2018). Looking generally at the scholarship produced on algorithmic impact assessment (risk assessment and impact evolution), it may be said that so far, there is no clear consensus on best practices and generally on how algorithmic risk assessment and algorithmic impact evaluation are best designed.

2.3.1.3 Real-Time and Automated Compliance Mechanisms

AI Systems and their underlying algorithms are pervasive and their risks quickly evolving. As AI systems interact with the real world, their behaviour may be quickly changing and evolving (Parker, 2012). A regulatory reaction to this new form of agile and uncertain technology may be real-time and continuous conformity and compliance checks. With automated and continuous auditing, drifts in functionality may be identified right when they occur and can be addressed timely and on the spot (Leslie et al., 2022). They are thought to be useful mainly for ex-post compliance mechanisms.

For real-time and automated compliance mechanisms potential technologies could be regtech and subtech solutions. These technologies were initially developed in the fintech sector but are now believed to reform regulatory processes generally with "low costs, effective and efficient

compliance, accurate information and real-time data, flexibility, easy reporting, security and analytics" (Johansson et al., 2019, p. 72). The technology is believed to supplement manual reporting and compliance checks in multiple regulatory environments (Arner, 2017; Deloitte, 2021). As the core data, processes and governance of compliance with different regulations are always similar (Nicoletti, 2018) duplication of work may thus be avoided (Hill, 2018).

Given the above-mentioned benefits of regtech and subtech for example Butler & O'Brien (2019) see great potential for the regulation of AI systems, machine learning and natural language processes. According to Renda (2018), the technology may be suitable to serve as a compliance mechanism for "digital labour, robotic process automation, machine learning, cognitive learning, big and smart data analysis, biometric technology, and natural language processing" (p. 80). Aside from compliance mechanisms, regtech and subtech may also increase cyber security, reduce operational risk, combat fraud, and other crimes and may be used to warn on a range of other issues.

2.3.1.4 Standardization and Certification

Standards are the technical and detailed settlements on specifications of products amongst stakeholders that often codify norms and legal requirements (Lewis et al., 2021). Private standardisation bodies negotiating and coordinating the standards are often made up of industry as well as government representatives, consumer organisations, trade unions and environmental organisations (Nativi & De Nigris, 2021). Under the EU's New Legislative Framework for instance pre-market controls and conformity assessments to testify product's performance and safety are carried out against a list of essential requirements by the manufacturers. With this self-assessment products may be marked with the 'CE' branding and enjoy the freedom of movement in the EU (European Parliament and the European Council, 2012). They are thus considered to be useful mainly for ex-ante conformity assessment.

The rationale of NLF is that "the manufacturer, having detailed knowledge of the design and production process, is best placed to carry out the complete conformity assessment procedure. Conformity assessment should therefore remain the obligation of the manufacturer" (European Parliament and of the Council,2008). In the EU, not all products fall under the NLF, for instance,

pharmaceuticals are assessed by the (public) European medical agency before granting market approval. In this way, defining the scope and application of standardization is naturally political (Cihon, 2019)– e.g. which products fall under NFL and which are under the assessment of public authorities. Also, standard-setting is a matter of international competition and power politics as major economies are aiming to push their homegrown standards to become harmonized international standards (Almeida et al., 2021; Cantero Gamito, 2021)

Standards may promote the rapid transfer of technologies from research to implementation. By defining requirements for products, services or processes, they ensure interoperability and quality (Ebers et al., 2021). Norms and standards can thus potentially make a significant contribution to explainability and security as well ass support acceptance and trust of AI systems (Din e.V. and DKE, 2020). Examples of relevant industry frameworks for AI systems are the 'Principles for Accountable Algorithms and a Social Impact Statement for Algorithms' (Diakopoulos, 2017) or the emerging standards by the IEE on algorithm's transparency, privacy or bias, the 'IEEE P7000: Model Process for Addressing Ethical Concerns During System Design' ('IEEE 7000™ Projects, 2021)or 'IEEE P7003: Algorithmic Bias Considerations' (Koene et al., 2018).

Private standardisation bodies are, however, not free of criticism. According to Veale & Zuiderveen Borgesius (2021) the rule-making bodies such as CEN or CENELEC function under private law. With the NLF they are argued to "serve the European consumer ill" (McGee & Weatherill, 1990, p.69). This is because relatively under-funded consumer organisations must catch up with industry representatives in opaque harmonisation processes. Often enough, well-equipped industry associations dictate the standards in their interest rather than the consumers' (Veale and Borgesius, 2021). In this light, leaving the "real rulemaking" (Veale and Borgesius, 2021, p.105) to European standardisation organisations, especially for potentially harmful and complex AI systems seems to be questionable. Oversight with democratic legitimation or by civil society actors may be particularly valuable for AI technologies that have severe consequences for human and fundamental rights.

2.3.2 Assurers of Compliance

2.3.2.1 Deployers (Internal Compliance Mechanisms)

With internal control of AI systems, the deployer is the main actor to ensure compliance with given guidelines and regulations. Internal compliance mechanisms for example outlined by the EU's AIA ('conformity assessment based on internal control', European Commission, 2021) imply properly documented internal checks to guarantee the conformity of the AI System with all requirements of the AI legislation. It further may include establishing a robust quality and risk management system, detailed technical documentation concerning internal governance processes, and safeguards against adversarial or negligent behaviour, e.g., through declarations of conformity (Mökander, Axente, et al., 2021). Harmonized standards defined by standardization bodies will foreseeably be the basis of this process (Ebers et al., 2021). Internal compliance mechanisms benefit from excellent access to the AI models and internal processes (e.g., to intermediate models or training data which are usually protected by trade secrets). Further, in case of failure to communicate proactively and transparently as well as in case of communication of incomplete or misleading information, authorities may impose fines (Mökander, Axente, et al., 2021).

An example of internal compliance mechanisms is proposed by Raji et al (2020). The framework for internal compliance mechanisms is called 'end - to end framework for internal algorithmic auditing' and was developed by researchers from Google. They hold that internal assurers are not less interested "to evaluate how well the product candidate, once in real-world operation, will fit the expected system behaviour encoded in standards" (p.3). According to Raji et al (2020), it is beneficial that assurers of the AI systems are also employees of the deploying company. This is because performance gaps may directly be solved with the product team whereas external assurers of compliance may have to go through a complex communication protocol before the information reaches the deployer. In this regard, it is argued that internal control could more likely result in beneficial organisational change in the deploying company. Internal compliance mechanisms are thus argued to result in informed model design decisions (Shah, 2018).

On the other side, internal compliance mechanisms raise concerns about the incentives for accurately self-reporting. The main interest for deployers according to Da Silva (personal

communication, May 2022) is to put the AI system on the market first or as quickly as possible and to keep them on the market. The strive for rapid deployment and profit may lie in contradiction with a resource- and time-intensive conformity assessment that may necessitate a costly redesign of the system. This conflict of interest is also acknowledged by researchers from Google (Raji et al., 2020) as internal compliance controls are "never isolated from the practices and people conducting [it]" (p.7). According to Da Silva (personal communication, May 2022), the risk of unreliable internal compliance mechanisms may be particularly given for small and medium deployers that do not have the same public scrutiny as multinational cooperation.

2.3.2.2 Notified Bodies (External Compliance Mechanisms)

The notified bodies system is a framework designed by the EU and used for ex-ante conformity assessment (European Comission, 2020). As it is included in the EU's AIA act (*EU Commission*, 2021) it serves as a blueprint for this paper to conceptualize and outline notified bodies.

Notified bodies are external institutions appointed and certified by governments or public agencies that conduct external conformity assessments (in the AIA national governments select so-called notifying bodies that in turn certify notified bodies). The main task of notified bodies is to assess deployers' internal quality management system by closely evaluating the technical documentation (not necessarily the source code). By checking and judging these documents, a notified body determines if a deployer and their system comply with the relevant requirements (Mökander, Axente, et al., 2021). In a positive case, a system can be certified e.g., in the EU with the CE certification.

Naturally, the deployers should be required to cooperate as much as needed to make the relationship work and to enable a due exercise of the compliance checks by the notified body– most importantly access to all resources and documents. Further, deployers should be required to swiftly react to and improve on any major malfunctioning based on the judgment and communication of the notified body. Possible notified bodies are big auditing firms such as the TÜV group or big consultancy firms like PwC (Mökander et al., 2021).

2.3.2.3 Civil Society Organisations (Bottom-Up Compliance Mechanism)

Alternative assurers of compliance such as civil society organisations have been suggested to replace or complement public bodies, deployers or notified bodies (Raji, 2021 as cited in Miller, 2021; Renda, 2019). Deployers may be required to comply with or voluntarily sign up to regulatory schemes which are privately developed by civil society actors. Civil society organisations either define their own rules and criteria for certification or use pre-given requirements (e.g. from the AI regulation). A further possibility is that civil society is involved in the academic endeavour to establish ethical principles (Mittelstadt, 2019). The civil society groups may be mandated by the national authorities, gain regulator-facilitated data access and may address sector-specific AI systems (e.g. labour unions could address AI systems in platform work). Possible actors for civil society enforcement of ethical guidelines are e.g. labour unions, consumer protection organisations or fundamental rights organisations. (Renda, 2019).

In its simplest form, impact assessment reports may be disclosed to affected communities (or their civil society representations) in simple language, giving citizens the possibility to be informed and organize informed opposition against new AI systems (Akkus, 2018). This may also help to prevent public backlash. A more advanced example of an existing voluntary self-regulation scheme from another sector that may serve as a blueprint is ISEAL (International Social and Environmental Accreditation and Labelling Alliance) which is mandated and funded by a range of civil society organisations. ISEAL functions as a meta-regulatory scheme, it defines rules and criteria and assists in implementing and controlling them (Fransen, 2015).

For example, Raji (2021 as cited by Miller, 2021) proposes a more radical civil society-based compliance mechanism. She argues that three policy interventions could make mandatory and rigorous third-party conformity assessment a reality: “a national incident reporting system to prioritize audits; an independent audit oversight board to certify auditors, set audit standards, and oversee the audit process; and mandated, regulator-facilitated data access for certified third-party auditors” (Raji, 2021 as cited by Miller, 2021). She argues that conventionally conformity assessments would be provided by auditing or consultancy firms interested to stay contractors of the providers, therefore meeting first and foremost the providers’ or users’ demands (e.g., a police

department) and not of the communities affected. An external and civil society driven compliance mechanism on the other hand could represent the interests and needs of impacted communities and may go beyond biases and look at "ecological, safety, or privacy impacts as well as a system's failure to live up to appropriate standards for transparency, explainability, and accountability." (Raji, 2021 as cited by Miller 2021).

A community certification scheme for AI systems as e.g. also suggested by Ghernaouti et al (2021 as cited in Duberry et al., 2021) may push for technology that benefits communities and make "visible and question the social and political conditions for innovation." (p. 16) Arguably, regulations like the EU's AIA already foresee an accreditation scheme for notified bodies (as described above) which could be extended to qualified civil society actors. Still, it remains highly questionable if civil society organisations have the resources and means to become assurers of AI systems given their limited technical expertise (Da Silva, personal communication, May 2022).

3. Methodology:

3.1 The MCDA Approach

The Multiple Decision Criteria Analyses (MCDA) is a group of methods that aim to rank different management or policy options based on a set of relevant assessment criteria and the evaluation against the criteria (Belton & Stewart, 2002). The MCDA approach is among others designed for complex decisions, a high degree of uncertainty, unknown impacts and limited knowledge. An example may be the intersection of technical and social dynamics of complex systems (Rogeberg et al., 2018; Stewart & Durbach, 2016)

In the last twenty years, the method found wide application with a growing acceptance among scholars in various fields (Cegan et al., 2017; Kurth et al., 2017), for instance, to make decisions in humanitarian aid (Curran et al., 2014), to manage environmental risk (Yatsalo et al., 2011), to manage nanotechnology (Bates et al., 2015, 2016; Linkov & Moberg, 2012) or to search for suitable sites for infrastructure, transportation and land use in general (Hamilton et al., 2016; M. J. Hill et al., 2005; Yatsalo et al., 2011). So far, however, in the field of cyber politics and specifically the regulation of AI, this method to the best of the author's knowledge has not been

applied yet. This is surprising given that both the potential impact of AI systems and the impact of different approaches to AI regulation are thought to be complex, multi-faceted, uncertain or entirely unknown (Nordström, 2021). This makes it an interesting case to apply an MCDA approach especially as MCDA finds strength in comparing already used and more experimental policy options (e.g. Rogeberg et al., 2018). This research may thus explicate experts' opinions on different AI regulation's compliance mechanisms with a methodology new and compelling to the field.

Generally, for MCDA both data and expert judgment may be used to score the policy options against the assessment criteria. Expert elicitation provides especially useful when data is missing or inadequate (Pesce et al., 2018). In the case of AI regulation's compliance mechanism, to the best of the author's knowledge, little data is available. Thus, this MCDA study relies on an expert scoring of policy options to investigate and evaluate the *expected* impact and effect of the proposed policy options.

3.2 The Steps of this Multi-Criteria Decision Analysis

1. Decide on a List of Stakeholders and Experts
2. Decide on a List of Policy Options
3. Select Assessment Criteria
4. Participants score the Policy Options against the Criteria
5. Participants weigh the Criteria
6. Aggregate the Weights and Scores

3.2.1 Decide on a List of Stakeholders and Experts

First, after delimiting the topic, relevant experts and stakeholders were selected and contacted. Stakeholders for AI regulation include developers, deployers, civil servants, public agencies (e.g. the data protection agency), civil society organisations/ NGOs and scholars (CEPS et al., 2021). To find relevant candidates for the stakeholders, the public consultation section of the EU's proposed AIA was consulted (European Parliament and European Council, 2021). Stakeholders from all domains outlined above were contacted to get balanced perspectives. 18 stakeholders agreed to participate and 11 finally completed the questionnaire, they are listed in Table 1. If not stated otherwise, they consented to have their name, organisation and position published in this

research. Given the small number of participants – which is conventional for MCDA approaches –scoring of policy options reflects a small subset of experts and stakeholders. Regarding MCDA literature, it did not become clear how balanced representativity is reached in the selection of stakeholders (e.g. Poustie et al., 2015). This should be seen as a limitation of the stakeholder selection and potentially the MCDA method in general.

Table 1: Participants of the MCDA

Name	Institution	Stakeholder Group
Anonymous Participant (at the request of the participant)	Federation of German Consumer Organisations	Consumer Organisation
Anonymous Participant (at the request of the participant)	Ada Lovelace Institute for AI	Scholar
David Nosák	Centre for Democracy and Technology	Civil Society Organisation
Jakob Mökander	Oxford Internet Institute, University of Oxford	Scholar, PhD Student
Jeannette Gorzala	AI Austria, European AI Forum	Policy Advisor
Jochen Friedrich	ETSI Board at ETSI	Standardisation Body
Julien Chasserieau	Digital Europe	Industry Representation
Koen Cobbaert	Philips Electronics N.V.	Industry Representation
Michele Oliva	Unipol Policy Department	User of AI System (Insurance Company)
Paul MacDonnell	Global Digital Foundation	Think Tank
Shea Brown	University of Iowa	Scholar

Note: Only participants that filled out the whole MCDA questionnaire are listed

3.2.2 Decide on a List of Policy Options

Based on a detailed literature review (see above), the policy options for this MCDA study were selected. Due to the limited scope of this research, the study only included four compliance mechanisms as well as the three assurers of the systems. The assurers of compliance are deployers (Internal compliance mechanisms), notified bodies (external compliance mechanism) and civil

society organisations (bottom-up compliance mechanisms). Further, four different proposed compliance mechanisms are compared with the MCDA namely Algorithmic Auditing, Algorithmic Impact Assessment, Standardisation as well as Real-Time and Automated Conformity Assessment. The policy options are outlined in detail in the literature review above.

Before the MCDA questionnaire was published, the list of policy options was peer-reviewed by fellow students of the cluster "Digitalization, Technology and Media" of the School of Transnational Governance. Subsequently, it was sent to and discussed with the supervisor of this thesis, Professor Andrea Renda who suggested changes and amendments. Here, it must be noted that in large scale MCDA studies, stakeholders are also invited to discuss the policy options, e.g., in a workshop. However, due to the small scale of this study, every step was conducted online and asynchronous as a lengthy meeting could not be expected from the participants. Thus, also the selection of policy options could not be discussed in a group session. This is a limitation since claims and judgements, as well as different perspectives of stakeholders, could not be exchanged ruling out the possibility of an 'internal peer review'.

To compensate for the lack of open consultation with the stakeholders, however, the policy options were shared with the stakeholders in a google doc before the MCDA. Feedback, questions, comments, and amendments could be shared via the comment and 'track changes' functions. Unfortunately, some stakeholders joined late and were thus not part of this first optional review of the policy options. Interestingly, none of the stakeholders commented and suggested any changes in the doc. This may be for two reasons: first, it may be that stakeholders were in overall agreement with the policy options. Second, it may indicate that engagement level and motivation to actively participate were relatively low which in turn may reduce the robustness of this research's results.

3.2.3 Select Assessment Criteria

To assign different scores to the different policy options, assessment criteria, as well as sub-criteria, had to be selected for the MCDA. The criteria aim to measure the performance of alternative policy options against a predefined goal, reduce uncertainty and increase the understanding of the analysed system (Linkow & Moberg, 2011). The criteria are supposed to be clear and succinct, meeting the consistent, comprehensible, non-redundant and exhaustive

requirements (Adiat et al., 2012). To arrive at a meaningful selection of assessment criteria, a range of relevant literature was reviewed. As a first step, a preliminary list of decision criteria was established. Just as with the policy options, this list was peer-reviewed by fellow students, send to the supervisor and shared with the stakeholders via google docs. Also in this case the participants did not comment on the google doc. This process was concluded with 17 sub-criteria with 3 cluster headings (main criteria). The cluster headings serve to reduce cognitive complexity when the criteria are weighted against each other. The criteria are listed in Table 2.

Table 2: Main Criteria and Sub-Criteria for the MCDA

Main Criteria	Sub-Criteria	
Effectiveness		Capability to ensure Legibility/Explicabity/Transparency
		Capability to ensure Accountability of AI Systems
		Capability to ensure Traceability of AI Systems' Decisions
		Ability to check for Robustness and Accuracy
		Ensure a Holistic View of the AI System
		Ensure AI compliance Mechanism are permitting Innovation of AI Systems
		Capability to discover Unknown Risks
		Creating Incentives for Developers to act Responsible
Feasibility of Application		Technical Feasibility
		Organizational Feasibility of Assurer of Compliance
		In-house Expertise at the Assurer of Compliance
		Dialectic Cooperation between Providers and Assurers of Compliance
Cost-Efficiency	Cost-Efficiency for Governments/ Public Agencies	One-off Costs
		Financial Burden
		Administrative Burden
	Cost-Efficiency for Deployers	One-off Costs
		Financial Burden
		Administrative Burden

Source: Established based on the literature review above.

3.2.4 Participants score the Policy Options against the Criteria

For the stakeholders to smoothly and easily score the policy options, an online questionnaire was created with the tool freeonlinesurvey.com. The stakeholders were asked to elicit a score for each policy option against each sub-criterion to express the subjective value they attribute to it. For the criteria 'effectiveness' and 'feasibility' the score ranges from 0 to 10, with 0 indicating 'very poor' performance and 10 indicating 'very good' performance. Correspondingly, for the category 'cost-efficiency for governments/ public agencies' and 'cost-efficiency for deployers' the scores range from 0 and 10, with 0 indicating 'minimal' and 10 indicating 'maximal' cost-efficiency. The participants were given explanations of the policy options (equivalent to the analysis in the literature review) before scoring.

On the first page of the online questionnaire, participants were further reminded that they should score the policy options based on their professional perspective, their cumulated knowledge but also their intuition, best estimation and educated guess. As the definition and explanations of the policy options given to the participants were still quite broad and as possible impacts of different enforcement mechanisms for AI regulation are still uncertain, it was communicated to the stakeholder that a degree of uncertainty was anticipated also in their answers.

3.2.5 Participants Weigh the Criteria

The participants were next asked to give a relative weight for each assessment criterion accounting for the relative importance they assign to this criterion. This was equally done on a scale from 0 to 10, with 0 indicating minimum importance and 10 indicating maximum importance. The weights of all participants were averaged to establish a default weight for this MCDA model (as done by Cegan et al., 2017). Subsequently, the results were normalised to add up to 100 (thus equalling percentage weights). The result of the weighting is shown in Table 3 in the analysis section.

3.2.6 Aggregate the Weights and Scores

In this step, all scores and weights were aggregated to reach a single comparable value for each policy option – the utility score. Given the small but conventional number of participants and the individual (not collective) scoring decision, the scores of all participants were first consolidated

using the mean. In this way, one value for all policy options again all 17 sub-criteria was established (corresponding to the methodology of Poustie et al., 2015).

Next, to establish the utility score for each policy option, the weighted sum method was applied, the simplest and most widely used approach for MCDA studies (Poustie et al, 2015, Hajkowicz & Higgins, 2008; Howard, 1991). With the weighted sum method, the consolidated scores are multiplied by their assigned relative weight and then summed together to gain the overall utility for each policy option. The methodology was chosen given the unclear relationship between the criteria's performances that would have made more complex multi-objective evolution algorithm-based approaches unjustified. The consolidated scores as well as the overall utility scores are listed in Tables 4 and 5 in the 'Analysis' section of this paper.

4. Analysis

4.1 Results of the Relative Weighting of Assessment Criteria

The results of the weighted criteria are displayed in Table 3 and will be outlined and analysed in this section. Displayed are the mean weights given to the assessment criteria by all 11 participants normalized to add up to 100%. Overall, the sub-criterion accounting for the 'effectiveness' of compliance mechanisms was believed to be most relevant, with a percentual relevance of between 6,1 % and 7,1 %. The second most relevant sub-criteria were the 'feasibility of the application' with values from 5,3 % to 7,0%. The costs-efficiencies were weighted to be of relatively low importance. Here, surprisingly, the costs-efficiency for deployers was rated to be a little more relevant (with values from 4,5 % to 5,1 %) than the cost-efficiency for governments and public agencies (values from 3,8 % to 4,9 %).

Looking closer at the sub-criteria, the 'capability to ensure accountability of the AI System' seems to be most relevant for the stakeholders with 7,6 %. On the other hand, the 'one-off costs for governments and public agencies' was rated to be the least important criteria with 3,8%. Interestingly, the consulted experts held that it is more important for compliance mechanisms to ensure transparency (7.1%), accountability (7,6%) and traceability (7,0%) than to leave sufficient leeway to innovate (6,1%).

Table 3: Relative Weights of Sub-Criteria as elicited by the MCDA Stakeholders

Criteria	Sub- Criteria	Weight in %
Effectiveness	Capability to ensure Legibility/Explicabity/Transparency	7,1
	Capability to ensure Accountability of AI Systems	7,6
	Capability to ensure Traceability of AI Systems' Decisions	7,0
	Ability to check for Robustness and Accuracy	6,9
	Ensure a Holistic View of the AI System	6,1
	Ensure AI compliance Mechanism are permitting Innovation of AI Systems	6,1
	Capability to discover Unknown Risks	6,8
Feasibility of Application	Creating Incentives for Developers to act Responsible	6,9
	Technical Feasibility	7,0
	Organizational Feasibility of Assurer of Compliance	5,5
	In-house Expertise at the Assurer of Compliance	5,3
Cost-Efficiency for Governments and Public Agencies	Dialectic Cooperation between Providers and Assurers of Compliance	3,8
	One-off Costs	4,6
	Financial Burden	4,9
Cost-Efficiency for Deployers	Administrative Burden	4,5
	One-off Costs	5,1
	Financial Burden	4,8

4.2 Results of the MCDA for the proposed four Compliance Mechanisms

The aggregated scores for the four compliance mechanisms are listed in Table 4 and further graphically displayed in Figure 1 with a stacked bar chart. The sums of all averaged criteria multiplied by the criteria weights are displayed (see Table 4). As there is a different total number of sub-criteria per each main criteria, intra-policy option comparison (comparing the values of the main criteria within one policy option) is not possible.

Table 4: Overall Utility Scores for the four Compliance Mechanisms

Main Criteria	Algorithmic Auditing	Algorithmic Impact Assessment	Standardization and Certification	Real-Time and Automated Compliance Mechanisms
Effectiveness	3,31	2,75	2,79	2,98
Feasibility of Application	1,55	1,50	1,66	1,35
Cost-Efficiency for Governments	0,68	0,50	0,60	0,65
Cost-Efficiency for Deployers	0,85	0,74	0,83	0,80
Utility Scores	6,39	5,49	5,89	5,79

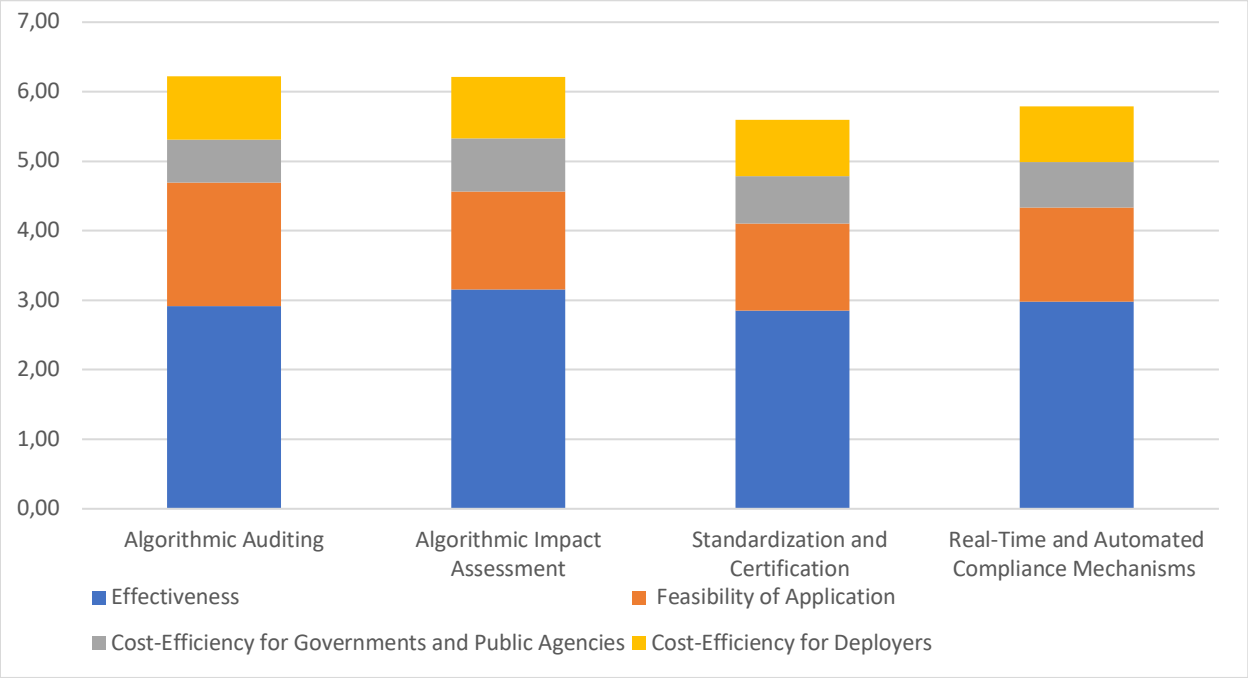
Source: Own Calculations

To the surprise of the author, the overall aggregated scores for AI enforcement mechanisms do not vary greatly, all ranging between 5,39 and 6,39. This is remarkable given that policy options like real-time and automated conformity assessment are at a very exploratory stage while standards and certification are widely practised enforcement mechanisms.

Looking at the results, algorithmic auditing was elicited with the highest overall score of 6,39. Equally, algorithmic auditing was believed to be most effective with a value of 3,31, most cost-efficient for governments as well as most cost-efficient for deployers. In contrast, algorithmic impact assessment received the overall lowest weighted ratings with 5,49. This is because, stakeholders have equally elicited the lowest effectiveness score for algorithmic impact assessments and viewed it to be the least cost-efficient, both for governments and deployers. Standardisation and Certification are believed to be most feasible (1,66) and second-most cost-efficient for deployers (0,83), however, scoring second-lowest on effectiveness (2,79).

Most interestingly, real-time/automated conformity assessment was valued at similar levels compared to the other compliance mechanisms with a value of 5,79 (compared to 5,89 for standardisation and 5,49 for algorithmic impact assessment), although being arguably most experimental. Expectedly, real-time and automated conformity assessment was ranked least feasible (1,35) as it is without best practice. On the other hand, real-time conformity assessment was elicited to be more effective (2,98) than standardisation (2,79) and algorithmic impact assessment (2,75).

Fig. 1: Overall Utility for the four Enforcement Mechanisms



4.3 Results of the MCDA for the three Assurers of Compliance

In Table 5, the ranking results (summed weights and scores) of three main assurers of compliance are listed, in Figure 2 they are visualized using a stacked bar chart. The differences in scoring for the assurers were equally marginal. However, some slight differences may be observed. Overall, internal and external compliance mechanisms have the same aggregated score (6,2). A civil society driven compliance mechanisms were rated overall significantly lower (5,5) and were assessed to be least effective (2,8), feasible (1,2) and cost-efficient (0,81). Further, while the experts consider notified bodies to be slightly more effective than internal compliance mechanisms, compliance mechanisms by notified bodies are also believed to be less feasible.

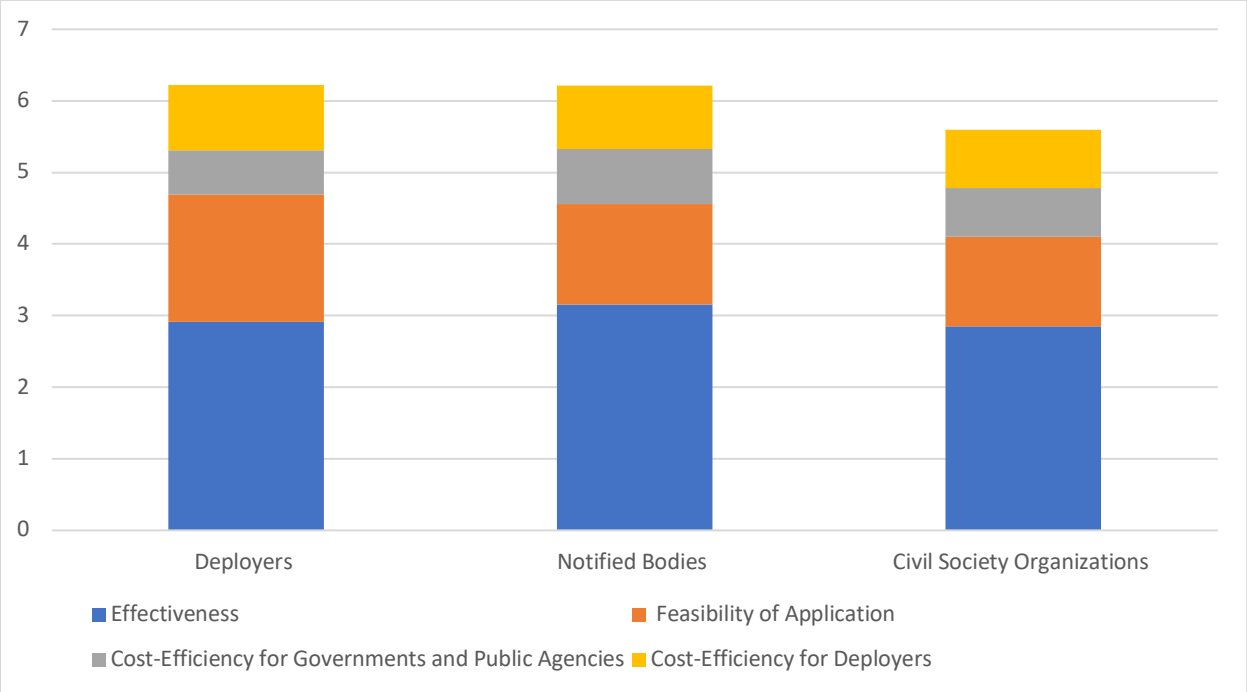
Table 5: Overall Utility Scores for the three Assurers of Compliance

Main Criteria	Deployers (Internal Compliance Mechanisms)	Notified Bodies (External Compliance Mechanisms)	Civil Society Organizations (Bottom-Up Compliance Mechanism)
Effectiveness	2,91	3,15	2,85
Feasibility of Application	1,78	1,41	1,25
Cost-Efficiency for Governments	0,62	0,77	0,68
Cost-Efficiency for Deployers	0,91	0,88	0,81
Utility Scores	6,22	6,21	5,59

Source: Own Calculations

Looking at cost-effectiveness, the results reveal surprisingly that internal compliance mechanisms are believed to be more cost-effective for the deployers themselves compared to the compliance checks by the other assurers. This result, however, may hint at a possible confusion stemming from the research design. Feedback from a participant confirmed that some participants were not sure if maximum cost-efficiency meant lowest or highest costs. Therefore, the values for cost-efficiency should be interpreted with caution.

Fig. 2: Overall Utility for the three Assurers of Compliance



4.4 Discussion

Looking at the overall results, it was acknowledged that the differentiation in scoring was not nearly as clear as anticipated. Both the assurers of compliance and the enforcement mechanisms were scored at surprisingly similar utility despite being arguably very different in their effectiveness, feasibility and cost-effectiveness (as outlined in the literature review). The goal of this MCDA approach was to shed light on differences between AI regulations' enforcement mechanisms by pooling experts' knowledge and judgments. Ideally, the expert participants would

have found a clear preference amongst the policy options or discarded a policy option for not being cost-efficient, feasible or effective.

While striking results were not yielded, experts of this study have confirmed general notions, judgements and ideas about AI regulation –thus potentially solidifying the canon of knowledge on compliance mechanisms. First, the idea of a civil society-based compliance mechanism was generally viewed with scepticism both in terms of feasibility, cost-efficiency and general effectiveness. This may correspond to research that frames this policy option as a slightly more radical, idealistic and less feasible idea (Raji, 2021). Second, compliance mechanisms based on external control was rated to be potentially more effective but also to be less feasible compared to compliance mechanism based on internal control. This corresponds with the notion that internal compliance mechanisms may be more accurate (Raji et al, 2020) due to direct access to the AI models. On the other hand, it also corresponds to the critique that internal compliance mechanisms may come with a conflict of interest (Da Silva, personal communication, May 2022). Third, algorithmic auditing was preferred over algorithmic impact assessment and was believed to be more effective and feasible than impact assessment. Fourth, despite its experimental nature, automated and real-time compliance mechanisms were perceived to be similarly feasible compared to the other policy proposals.

The relatively small differences in scoring may have several reasons. First, potentially the questionnaire was too extensive and long. Therefore, the participants may have not had the patience and perseverance to go through the whole questionnaire with the required accuracy. In support of this hypothesis, some stakeholders who initially agreed to devote 15 min for the MCDA quit the questionnaire mid-way (and were thus excluded from the results). Adding to this notion, the level of engagement among contacted experts was quite low. Second, some aspects of the research design as described above were perceived to be misleading. Third, it might be hypothesized that participants were cautious to elicit ‘extreme’ values given the general uncertainty surrounding potential enforcement mechanisms. In this sense, a coping mechanism to handle uncertainty could have been to settle for ‘safe and not opinionated’ mid- values. As mentioned above, the last limitation is the small– but conventional – number of participants. From reviewing MCDA literature, it was not clear to the author how a balanced group of stakeholders

could be selected. Although, stakeholders were contacted and appointed (as much as possible given the response rate) to balance between stakeholder groups and to avoid biases in any direction, a distortion due to the combination of experts cannot be ruled out.

5. Conclusion

This research has used a Multi-Criteria Decision Analysis to capture expert opinions on the expected effectiveness, feasibility and cost-efficiency of different policy options for compliance and enforcement of AI guidelines. Four different compliance mechanisms and three different assurers of compliance were compared.

First, the research has briefly defined AI systems and their potential to advance human societies. It has further outlined the risks that AI systems pose to human and fundamental rights and that constitute the need for innovative, effective and future-proof AI regulation. Subsequently, this research has listed different national and international approaches to formulating principles and guidelines for ethical, trustworthy and responsible AI. It was emphasized that principles and guidelines need to be decoded into norms and essential requirements that only become meaningful –as in every regulatory framework –if properly implemented into practice with adequate enforcement and compliance mechanisms. Based on these considerations, a definition of compliance mechanisms was established namely the process by which the expected function of an AI system, as well as its conformity with laws, standards and regulations, is checked, controlled for and demonstrated to others. A range of different compliance mechanisms at all stages of the AI lifecycle was briefly presented.

As reliable quantitative data for AI regulation's compliance mechanism is lacking, a Multi-Criteria Decision Analysis was utilized to assess and compare different policy options. MCDAs structure decision processes with multiple assessment criteria, general uncertainty and lack of knowledge for example by integrating expert elicitation. Based on a literature review, four different compliance mechanisms namely algorithmic auditing, algorithmic impact assessment, real-time and automated conformity assessment as well as standardisation and certification were compared against each other. Further, three different assurers (actors carrying out compliance

mechanisms) were compared: the deployers themselves (internal conformity assessment), notified bodies (external compliance mechanisms) and finally civil society organisations (bottom-up conformity assessment). For the MCDA, 11 experts were asked to score the 7 policy options against 17 sub-criteria on a scale from 0 to 10. The sub-criteria were clustered around the three main assessment criteria effectiveness, feasibility and cost-efficiency. To account for the different relative importance of the sub-criteria, they were weighted based on subjective relevance on a scale from 0 to 10. Scores were added together based on the weighted sum model to build comparable utility scores for each policy option.

The results of the MCDA were less indicative and significant than anticipated as differences between the policy options remained weak. However, some general conclusions may be drawn from this MCDA research. First, the notion of a civil society-based compliance mechanism was scored to be less effective, less feasible and more costly compared to all other compliance mechanisms. Second, external compliance mechanisms (by notified bodies) were believed to be more effective but also more difficult to implement compared to internal compliance mechanisms. Third, algorithmic auditing scored highest among all policy options. Fourth, despite its experimental nature, automated and real-time compliance mechanisms were not scored significantly lower, and in fact, were rated to be similar feasible and cost-efficient in comparison to the other compliance mechanisms.

This study has given an initial notion of experts' judgment and preferences for AI regulation's enforcement mechanisms. It has also established a set of criteria to compare the utility of compliance mechanisms against that can be built on in future. However, further research is needed. In particular, it remains to be analysed if internal compliance checks are more effective due to their access to the systems or if external compliance systems are more thorough due to their independence. Further, the potentials and risks of technical standardisation for AI systems by standardisation bodies should be analysed critically. It remains to be seen if they are qualified for a holistic view considering all potential risks to human and fundamental rights or if their perspective is too narrow and technical. Finally, civil society-based compliance and conformity assessments should be studied and tested for their feasibly, especially since they promise to pick up holistic perspectives of affected communities.

The impact of AI systems on human societies in the next decades is believed to be significant, including benefits and risks to our human and fundamental rights. At this crossroads, it remains important to decide on future-proof regulations. This study has shown that a deliberate choice for effective compliance mechanisms may impact the overall effectiveness of the AI regulation. This choice for reliable compliance mechanisms should consider the interests of all stakeholders, but especially the interests of affected communities and individuals.

6. Bibliography

Access Now. (2018). Human rights in the age of artificial intelligence. *Access Now November*.

Adiat, K. A. N., Nawawi, M. N. M., & Abdullah, K. (2012). Assessing the accuracy of GIS-based elementary multi criteria decision analysis as a spatial prediction tool – A case of predicting potential zones of sustainable groundwater resources. *Journal of Hydrology*, *440–441*, 75–89. https://doi.org/10.1016/j.jhydrol.2012.03.028

AI Ethics Impact Group. (2029). *AI Ethics Impact Group: From Principles to Practice - VDE*. Retrieved 19 May 2022, from https://www.ai-ethics-impact.org/en

AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/346720

Aizenberg, E., & van den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*, *7*(2), 2053951720949566. https://doi.org/10.1177/2053951720949566

Akkus, S. (2018). *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*.openresearch.amsterdam.http://openresearch.amsterdam/nl/page/37785/algorithmic-impact-assessments-a-practical-framework-for-public

Almeida, P., Santos Jr, C., & Farias, J. (2021). Artificial Intelligence Regulation: A framework for governance. *Ethics and Information Technology*, *23*. https://doi.org/10.1007/s10676-021-09593-z

Alter, A. (2017). *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*. Penguin Press.

Arner, D. W. (2017). FinTech, RegTech, and the Reconceptualization of Financial Regulation. *International Law*, 44.

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

Baldwin, R., Cave, M., & Lodge, M. (1999). Regulatory strategies. *Understanding Regulation. Theory, Strategy, and Practice*, 32–62.

Bandy, J. (2021). Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *ArXiv:2102.04256 [Cs]*. http://arxiv.org/abs/2102.04256

Bates, M. E., Grieger, K. D., Trump, B. D., Keisler, J. M., Plourde, K. J., & Linkov, I. (2016). Emerging Technologies for Environmental Remediation: Integrating Data and Judgment. *Environmental Science & Technology*, *50*(1), 349–358. https://doi.org/10.1021/acs.est.5b03005

Bates, M. E., Larkin, S., Keisler, J. M., & Linkov, I. (2015). How decision analysis can further nanoinformatics. *Beilstein Journal of Nanotechnology*, *6*(1), 1594–1600. https://doi.org/10.3762/bjnano.6.162

Bauer, J. (2017). *The Necessity of Auditing Artificial Intelligence Algorithms* (SSRN Scholarly Paper No. 3218675). Social Science Research Network. https://doi.org/10.2139/ssrn.3218675

Belton, V., & Stewart, T. J. (2002). *Multiple criteria decision analysis: An integrated approach*. http://copac.ac.uk/search?rn=1&ti=Multiple+Criteria+Decision+Analysis%3A+An+Integrated+Approach+&sort-order=rank

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., … Anderljung, M. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *ArXiv:2004.07213 [Cs]*. http://arxiv.org/abs/2004.07213

Butler, T., & O'Brien, L. (2019). Understanding RegTech for Digital Regulatory Compliance. In T. Lynn, J. G. Mooney, P. Rosati, & M. Cummins (Eds.), *Disrupting Finance: FinTech and Strategy in the 21st Century* (pp. 85–102). Springer International Publishing. https://doi.org/10.1007/978-3-030-02330-0_6

CAHAI, A. H. C. O. A. I. (2020). *Feasibility Study*. Council of Europe.

Canadian Government. (2021). *Algorithmic Impact Assessment—Évaluation de l'incidence algorithmique*. Retrieved 15 May 2022, from https://open.canada.ca/aia-eia-js/?lang=en

Cantero Gamito, M. (2021). *From Private Regulation to Power Politics: The Rise of China in AI Private Governance Through Standardisation* (SSRN Scholarly Paper No. 3794761). Social Science Research Network. https://doi.org/10.2139/ssrn.3794761

CDEI. (2021a). *The need for effective AI assurance—Centre for Data Ethics and Innovation Blog*. https://cdei.blog.gov.uk/2021/04/15/the-need-for-effective-ai-assurance/

CDEI. (2021b). *The roadmap to an effective AI assurance ecosystem*. GOV.UK. https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem

Cegan, J. C., Filion, A. M., Keisler, J. M., & Linkov, I. (2017). Trends and applications of multi-criteria decision analysis in environmental sciences: Literature review. *Environment Systems and Decisions*, *37*(2), 123–133.

CEPS, Directorate-General for Communications Networks, C. and T. (European C., ICF, Wavestone, Renda, A., Fanni, R., Laurer, M., Agnes Sipiczki, Yeung, T., Maridis, G., Fernandes, M., Gabor Endrodi, G., Milio, S., Devenyi, V., Georgiev, S., Pierrefeu, G. de, & Arroyo, J. (2021). *Study to support an impact assessment of regulatory requirements for*

Artificial Intelligence in Europe: Final report. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/523404

Cihon, P. (2019). *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*.

Council of Europe. (2019). *Unboxing artificial intelligence: 10 steps to protect human rights*. Curran, R. W., Bates, M. E., & Bell, H. M. (2014). Multi-criteria Decision Analysis Approach to Site Suitability of U.S. Department of Defense Humanitarian Assistance Projects. *Procedia Engineering*, *78*, 59–63. https://doi.org/10.1016/j.proeng.2014.07.039

Da Silva, F. O. (2022). European Consumer Organisation*Personal Interview*.

Dafoe, B. Z. and A., & Oxford, C. for the G. of A., Future of Humanity Institute, University of. (2019). *Artificial Intelligence: American Attitudes and Trends*. Retrieved 7 May 2022, from https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/

Danish Institute for Human Rights,. (2016). *Human rights compliance assessment quick check | The Danish Institute for Human Rights*. https://www.humanrights.dk/publications/human-rights-compliance-assessment-quick-check

Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products, and repealing Council Decision 93/465/EEC (Text with EEA relevance), 218 OJ L (2008). http://data.europa.eu/eli/dec/2008/768(1)/oj/eng

Deloitte. (2021). *RegTech Universe*. Deloitte Luxembourg. https://www2.deloitte.com/lu/en/pages/technology/articles/regtech-companies-compliance.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

Diakopoulos, N. (2017). *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms: FAT ML*. https://www.fatml.org/resources/principles-for-accountable-algorithms

Din e.V. and DKE. (2020). *Artificial Intelligence Standardization Roadmap*. https://www.dke.de/standardization-roadmap-ai

Directorate-General for Justice and Consumers (European Commission), & Trasys International. (2020). *Study on the use of innovative technologies in the justice field: Final report*. Publications Office of the European Union. https://data.europa.eu/doi/10.2838/585101

Duberry, J., Büchi, M., Berryhill, J., Dormeier Freire, A., Garzia, D., Ghernaouti:, S., Hanifa, V., Hamidi, S., George Jain, A., Kosmerlj, A., Leander, A., Leclère, O., Lorenzini, J., Stauffer, M., Stern, N., Verma, H., & Welp, Y. (2021). *Promises and Pitfalls of Artificial Intelligence for Democratic Participation. Workshop Proceedings.CCDSEE, GSI, University of Geneva,*

December 10 – 11, 2020, Virtual Event (SSRN Scholarly Paper No. 3817666). Social Science Research Network. https://doi.org/10.2139/ssrn.3817666

Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H., & Steinrötter, B. (2021). The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *J*, *4*(4), 589–603. https://doi.org/10.3390/j4040043

ECP, P. voor de I. (2019, January 20). Artificial Intelligence Impact Assessment (English version). *ECP | Platform voor de InformatieSamenleving*. https://ecp.nl/publicatie/artificial-intelligence-impact-assessment-english-version/

Edelman. (2019). *2019 Edelman Trust Barometer*. Edelman. https://www.edelman.com/trust/2019-trust-barometer

EU Comission. (2016). *GDPR compliance checklist*. GDPR.Eu. https://gdpr.eu/checklist/

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, (2021) (testimony of EU Comission). https://eur ex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A52021PC0206*Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*, (2021) (testimony of EU Comission). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Comission. (2020). *Notified bodies*. https://ec.europa.eu/growth/single-market/goods/building-blocks/notified-bodies_de

European Parliament Think Tank. (2019). *A governance framework for algorithmic accountability and transparency | Think Tank | European Parliament*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262

European Union Agency for Fundamental Rights. (2020, November 18). *Getting the future right – Artificial intelligence and fundamental rights*. European Union Agency for Fundamental Rights. https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights

Fanni, C. F. K., Joshua P. Meltzer, Andrea Renda, Alex Engler, and Rosanna. (2021, October 25). Strengthening international cooperation on AI. *Brookings*. https://www.brookings.edu/research/strengthening-international-cooperation-on-ai/

Fanni, R., Steinkogler, V. E., Zampedri, G., & Pierson, J. (2020). Active Human Agency in Artificial Intelligence Mediation. *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*, 84–89. https://doi.org/10.1145/3411170.3411226

Felderer, M., & Ramler, R. (2021). Quality Assurance for AI-based Systems: Overview and Challenges. *ArXiv:2102.05351 [Cs]*. http://arxiv.org/abs/2102.05351

Fitzpatrick, T. (2012). Evaluating legislation: An alternative approach for evaluating EU Internal Market and Services law. *Evaluation*, *18*(4), 477–499. https://doi.org/10.1177/1356389012460439

Floridi, L. (2019). Translating Principles Into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy and Technology*, *32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

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 Scholarly Paper No. 4064091). Social Science Research Network. https://doi.org/10.2139/ssrn.4064091

Fransen, L. (2015). The politics of meta-governance in transnational private sustainability governance. *Policy Sciences*, *48*(3), 293–317. https://doi.org/10.1007/s11077-015-9219-8

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Gesley, J., Ahmad, T., Soares, E., Levush, R., Guerra, G., Martin, J., Buchanan, K., Zhang, L., Umeda, S., Grigoryan, A., Boring, N., Hofverberg, E., Feikhert-Ahalt, C., Rodriguez-Ferrand, G., Sadek, G., & Goitom, H. (2019). *Regulation of Artificial Intelligence in Selected Jurisdictions*.

Gibney, E. (2020). The battle for ethical AI at the world's biggest machine-learning conference. *Nature*, *577*(7792), 609–610.

Global Partners Digital. (2021). *National Artificial Intelligence Strategies and Human Rights: A Review (second edition) - Publication | Global Partners Digital*. Retrieved 18 May 2022, from https://www.gp-digital.org/publication/national-artificial-intelligence-strategies-and-human-rights-a-review-second-edition/

Goldhaber-Fiebert, J. D., & Prince, L. (2019). Impact evaluation of a predictive risk modeling tool for Allegheny county's child welfare office. *Pittsburgh: Allegheny County*.

Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Illustrated Edition). Houghton Mifflin Harcourt.

Hajkowicz, S., & Higgins, A. (2008). A comparison of multiple criteria analysis techniques for water resource management. *European Journal of Operational Research*, *184*(1), 255–265.

Hamilton, M. C., Nedza, J. A., Doody, P., Bates, M. E., Bauer, N. L., Voyadgis, D. E., & Fox-Lent, C. (2016). Web-based geospatial multiple criteria decision analysis using open software and standards. *International Journal of Geographical Information Science*, *30*(8), 1667–1686. https://doi.org/10.1080/13658816.2016.1155214

High-Level Expert Group on AI (AI HLEG). (2020). *European AI Alliance—ALTAI - The Assessment List on Trustworthy Artificial Intelligence*. https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence

Hill, J. (2018). *Fintech and the Remaking of Financial Institutions* (1° edizione). Academic Press.

Hill, M. J., Braaten, R., Veitch, S. M., Lees, B. G., & Sharma, S. (2005). Multi-criteria decision analysis in spatial decision support: The ASSESS analytic hierarchy process and the role of quantitative methods and spatially explicit analysis. *Environmental Modelling & Software*, *20*(7), 955–976.

Howard, A. F. (1991). A critical look at multiple criteria decision making techniques with reference to forestry applications. *Canadian Journal of Forest Research*, *21*(11), 1649–1659. https://doi.org/10.1139/x91-228

ICO, I. C. O. (2020). *Guidance on the AI auditing framework Draft guidance for consultation*. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

IEEE. (2019). *Ethically Alliged Design*. Ethics In Action | Ethically Aligned Design. https://ethicsinaction.ieee.org/

IEEE 7000™ Projects | IEEE Ethics In Action in A/IS - IEEE SA. (n.d.). *Ethics In Action | Ethically Aligned Design*. Retrieved 19 May 2022, from https://ethicsinaction.ieee.org/p7000/

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: The global landscape of ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johansson, E., Sutinen, K., Lassila, J., Lang, V., Martikainen, M., & Lehner, O. M. (2019). *REGTECH- A NECESSARY TOOL TO KEEP UP WITH COMPLIANCE AND REGULATORY CHANGES?* 15.

Jotterand, F., & Bosco, C. (2020). Keeping the 'Human in the Loop' in the Age of Artificial Intelligence: Accompanying Commentary for 'Correcting the Brain?' by Rainey and Erden. *Science and Engineering Ethics*, *26*(5), 2455–2460. https://doi.org/10.1007/s11948-020-00241-1

Kaminski, M. E., & Malgieri, G. (2019). *Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations* (SSRN Scholarly Paper No. 3456224). Social Science Research Network. https://doi.org/10.2139/ssrn.3456224

Kazim, E., & Koshiyama, A. (2020). *AI Assurance Processes* (SSRN Scholarly Paper No. 3685087). Social Science Research Network. https://doi.org/10.2139/ssrn.3685087

Kazim, E., & Koshiyama, A. (2021). The interrelation between data and AI ethics in the context of impact assessments. *AI and Ethics*, *1*(3), 219–225. https://doi.org/10.1007/s43681-020-00029-w

Kim, P. (2017). *Auditing Algorithms for Discrimination* (SSRN Scholarly Paper No. 3093982). Social Science Research Network. https://papers.ssrn.com/abstract=3093982

Koene, A., Dowthwaite, L., & Seth, S. (2018). IEEE P7003TM Standard for Algorithmic Bias Considerations. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 38–41. https://doi.org/10.23919/FAIRWARE.2018.8452919

Konrad, K., Rip, A., & Schulze Greiving, V. (2017). *Constructive Technology Assessment – STS for and with Technology Actors*.

Koshiyama, A., & Engin, Z. (2019). *Algorithmic Impact Assessment: Fairness, Robustness and Explainability in Automated Decision-Making*. Data for Policy 2019: Digital Trust and Personal Data (Data for Policy 2019) (DFP), London. https://doi.org/10.5281/zenodo.3361708

Kurth, M. H., Larkin, S., Keisler, J. M., & Linkov, I. (2017). Trends and applications of multi-criteria decision analysis: Use in government agencies. *Environment Systems and Decisions*, *2*(37), 134–143. https://doi.org/10.1007/s10669-017-9644-7

Lachaud, E. (2020). What GDPR tells about certification. *Computer Law & Security Review*, *38*, 105457. https://doi.org/10.1016/j.clsr.2020.105457

Latonero, M., & Agarwal, A. (2021). Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar. *Carr Center Discussion Paper Series*.

Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M., & Rincon, C. (2022). Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. *ArXiv:2202.02776 [Cs]*. https://doi.org/10.5281/zenodo.5981676

Lewis, D., Filip, D., & Pandit, H. J. (2021). An Ontology for Standardising Trustworthy AI. In *Factoring Ethics in Technology, Policy Making, Regulation and AI*. IntechOpen. https://doi.org/10.5772/intechopen.97478

Linkov, I., & Moberg, E. (2012). *Multi-Criteria Decision Analysis: Environmental Applications and Case Studies*. CRC Press. https://doi.org/10.1201/b11471

Linkow, I., & Moberg, E. (2011). *Multi-Criteria Decision Analysis: Environmental Applications and Case Studies* (1° edizione). CRC Press.

Lovelace, A., & DataKind, U. (2020). *Examining the black box: Tools for assessing algorithmic systems*. Technical report, AdaLovelace Institute, https://ico. org. uk/media/about ….

Makarov, V. O., & Davydova, M. L. (2020). *On the concept of regulatory sandboxes*. 1014–1020.

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, *34*(4), 754–772. https://doi.org/10.1016/j.clsr.2018.05.017

Mattu, J. A., Jeff Larson,Lauren Kirchner,Surya. (2016). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=gl4jHLt-6ZxkcB55q8h_B25ydpK2Tm56

McGee, A., & Weatherill, S. (1990). The Evolution of the Single Market – Harmonisation or Liberalisation. *The Modern Law Review*, *53*(5), 578–596. https://doi.org/10.1111/j.1468-2230.1990.tb01826.x

McGregor, L., Murray, D., & Ng, V. (2019). INTERNATIONAL HUMAN RIGHTS LAW AS A FRAMEWORK FOR ALGORITHMIC ACCOUNTABILITY. *International & Comparative Law Quarterly*, *68*(2), 309–343. https://doi.org/10.1017/S0020589319000046

Miller, K. (2021). *Radical Proposal: Third-Party Auditor Access for AI Accountability*. Stanford HAI. https://hai.stanford.edu/news/radical-proposal-third-party-auditor-access-ai-accountability

Mittelstadt, B. (2019). *Principles Alone Cannot Guarantee Ethical AI* (SSRN Scholarly Paper No. 3391293). Social Science Research Network. https://doi.org/10.2139/ssrn.3391293

Mökander, J., & Axente, M. (2021). *Ethics-Based Auditing of Automated Decision-Making Systems: Intervention Points and Policy Implications* (SSRN Scholarly Paper No. 3958887). Social Science Research Network. https://papers.ssrn.com/abstract=3958887

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2021). *Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation* (SSRN Scholarly Paper No. 3959746). Social Science Research Network. https://papers.ssrn.com/abstract=3959746

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*, *27*(4), 44. https://doi.org/10.1007/s11948-021-00319-4

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), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020b). 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), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Nativi, S., & De Nigris, S. (2021, July 14). *AI Standardisation Landscape: State of play and link to the EC proposal for an AI regulatory framework*. JRC Publications Repository. https://doi.org/10.2760/376602

Nicoletti, B. (2018). *The Future of FinTech*. https://link.springer.com/book/10.1007/978-3-319-51415-4

Nordström, M. (2021). AI under great uncertainty: Implications and decision strategies for public policy. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01263-4

OECD. (2019). Recommendation of the Council on Artificial Intelligence (OECD). *International Legal Materials*, *59*(1), 27–34. https://doi.org/10.1017/ilm.2020.5

Parker, C. (2012). Unexpected challenges in large scale machine learning. *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, 1–6. https://doi.org/10.1145/2351316.2351317

Pesce, M., Terzi, S., Al-Jawasreh, R. I. M., Bommarito, C., Calgaro, L., Fogarin, S., Russo, E., Marcomini, A., & Linkov, I. (2018). Selecting sustainable alternatives for cruise ships in Venice using multi-criteria decision analysis. *The Science of the Total Environment*, *642*, 668–678. https://doi.org/10.1016/j.scitotenv.2018.05.372

Poustie, M. S., Deletic, A., Brown, R. R., Wong, T., de Haan, F. J., & Skinner, R. (2015). Sustainable urban water futures in developing countries: The centralised, decentralised or hybrid dilemma. *Urban Water Journal*, *12*(7), 543–558. https://doi.org/10.1080/1573062X.2014.916725

PwC. (2019). *A practical guide to Responsible Artificial Intelligence (AI)*. 20.

QC, R. A., & Dee, M. (2020). *Meeting the new challenges to equality and non-discrimination from increased digitisation and the use of Artificial Intelligence*. https://equineteurope.org/wp-content/uploads/2020/06/ai_report_digital.pdf

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

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435. https://doi.org/10.1145/3306618.3314244

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. *ArXiv:2001.00973 [Cs]*. http://arxiv.org/abs/2001.00973

Ranchordas, S. (2021). *Experimental Regulations for AI: Sandboxes for Morals and Mores* (SSRN Scholarly Paper No. 3839744). Social Science Research Network. https://doi.org/10.2139/ssrn.3839744

Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council Text with EEA relevance, 316 OJ L (2012). http://data.europa.eu/eli/reg/2012/1025/oj/eng

Reisman, D., Jason Schultz, Kate Crawford, & Meredith Whittaker. (2018). *ALGORITHMIC IMPACT ASSESSMENTS - ... / algorithmic-impact-*. https://pdf4pro.com/view/algorithmic-impact-assessments-53bf9.html

Renda, A. (2019). *Artificial Intelligence. Ethics, governance and policy challenges*. CEPS Centre for European Policy Studies. https://www.ceeol.com/search/book-detail?id=829907

Review, M. T. (2019, March 12). This Is How A.I. Bias Really Happens—And Why It's So Hard to Fix. *MIT Technology Review*. https://medium.com/mit-technology-review/this-is-how-a-i-bias-really-happens-and-why-its-so-hard-to-fix-369a864b4be7

Rogeberg, O., Bergsvik, D., Phillips, L. D., van Amsterdam, J., Eastwood, N., Henderson, G., Lynskey, M., Measham, F., Ponton, R., Rolles, S., Schlag, A. K., Taylor, P., & Nutt, D. (2018). A new approach to formulating and appraising drug policy: A multi-criterion decision analysis applied to alcohol and cannabis regulation. *The International Journal on Drug Policy*, *56*, 144–152. https://doi.org/10.1016/j.drugpo.2018.01.019

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

Schell-Busey, N. (2022). Using Meta-Analysis/Systematic Review to Examine Corporate Compliance. In B. van Rooij & M. Rorie (Eds.), *Measuring Compliance: Assessing*

CorporateCrime and Misconduct Prevention (pp. 264–284). Cambridge University Press. https://doi.org/10.1017/9781108770941.015

Scherer, M. U. (2015). *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*. https://doi.org/10.2139/SSRN.2609777

Schulam, P., & Saria, S. (2019). *Can you trust this prediction? Auditing pointwise reliability after learning*. 1022–1031.

Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170362. https://doi.org/10.1098/rsta.2017.0362

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). *Release Strategies and the Social Impacts of Language Models* (arXiv:1908.09203). arXiv. https://doi.org/10.48550/arXiv.1908.09203

Spielkamp, M. (2019, January). *Automating Society: Taking Stock of Automated Decision-Making in the EU. BertelsmannStiftung Studies 2019* [Other]. http://aei.pitt.edu/102677/

Stewart, T. J., & Durbach, I. (2016). Dealing with Uncertainties in MCDA. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple Criteria Decision Analysis: State of the Art Surveys* (pp. 467–496). Springer. https://doi.org/10.1007/978-1-4939-3094-4_12

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

Treasury Board of Canada, T. B. of C. (2021, March 22). *Algorithmic Impact Assessment Tool* [Guidance]. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & SOCIETY*, *35*(1), 147–163. https://doi.org/10.1007/s00146-018-0845-5

US GAO, U. S. G. A. (2021). *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*. https://www.gao.gov/products/gao-21-519sp

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

Veale, M., & Zuiderveen Borgesius, F. (2021). *Demystifying the Draft EU Artificial Intelligence Act* (SSRN Scholarly Paper No. 3896852). Social Science Research Network. https://papers.ssrn.com/abstract=3896852

Walsh, T., Levy, N., Bell, G., Elliott, A., Maclaurin, J., Mareels, I., & Wood, F. *The effective and ethical development of artificial intelligence: An opportunity to improve our wellbeing*.

Warofka, A. (2018, November 6). An Independent Assessment of the Human Rights Impact of Facebook in Myanmar. *Meta*. https://about.fb.com/news/2018/11/myanmar-hria/

Yatsalo, B., Sullivan, T., Didenko, V., & Linkov, I. (2011). Environmental risk management for radiological accidents: Integrating risk assessment and decision analysis for remediation at different spatial scales. *Integrated Environmental Assessment and Management*, *7*(3), 393–395. https://doi.org/10.1002/ieam.229