Technology and Regulation

High-risk AI transparency? On qualified transparency mandates for oversight bodies under the EU AI Act

Author(s)	Kasia Söderlund
Contact	katarzyna.soderlund@lth.lu.se
Affiliation(s)	Kasia Söderlund is a doctoral student at the Department of Technology & Society, Lund University, Sweden. The study was supported by the Wallenberg AI, Autonomous Systems and Software Program - Humanity and Society (WASP-HS), funded by the Marianne and Marcus Wallenberg Foundation. Open access funding provided by Lund University.

Keywords	AI Act, AI opacity, qualified transparency, EU law, enforcement		
Published	Received: 19 Nov 2024	Accepted: 17 Apr 2025	Published: 04 Jun 2025
Citation	Kasia Söderlund, High-risk AI transparency? On qualified transparency mandates for oversight bodies under the EU AI Act, Technology and Regulation, 2025, 96-113 • 10.71265/6bedar76 • ISSN: 2666-139X		

Abstract

The legal opacity of AI technologies has long posed challenges in addressing algorithmic harms, as secrecy enables companies to retain competitive advantages while limiting public scrutiny. In response, ideas such as *qualified transparency* have been proposed to provide AI accountability within the confidentiality constraints. With the introduction of the EU AI Act, the foundations for human-centric and trustworthy AI have been established. The framework sets regulatory requirements for certain AI uses and grants oversight bodies broad transparency mandates to enforce the new rules. This paper examines these transparency mandates under the AI Act and argues that it effectively implements qualified transparency, which may potentially mitigate the problem of AI opacity. Nevertheless, several challenges remain in achieving the Act's policy objectives.

1. Introduction

The significant advancements in artificial intelligence (AI)[1] over the past decades, driven in particular by the innovations in machine learning (ML)[2], have paved the way for the widespread adoption of AI across industries[3]. More recently, the emergence of generative AI, such as large language models (LLMs), has made AI broadly available to the public for personal and professional uses[4]. AI technologies are also increasingly adopted in automated decision-making (ADM) within the public sector[5], which in many instances may directly affect our fundamental rights[6]. However, while the enthusiasm surrounding AI continues to grow[7], so do the concerns over how it is developed, deployed, and used in certain cases. Numerous AI applications have been shown to negatively affect civil rights and democratic values, resulting in the proliferation of *algorithmic harms*[8].

Although the existing legal frameworks could arguably address many of the harmful effects associated with AI, the *legal opacity*[9] surrounding AI technologies often constitutes a significant obstacle for third-party compliance examination. Since AI and algorithmic systems do not easily fit within the conventional IP frameworks of patents or copyright laws, AI companies commonly rely on the trade secrecy or other confidentiality schemes to protect the sensitive information of their software[10]. This is problematic, as the legal frameworks designed primarily to encourage fair competition are now frequently used to prevent public scrutiny[11].

With the growing scale and gravity of algorithmic harms, addressing the issue of legal opacity of AI systems has become increasingly urgent[12]. One of the solutions proposed in this context has been the idea of a legal mechanism that assigns the responsibility for the comprehensive evaluation of algorithmic systems to trusted expert bodies. Such transparency intermediaries, on the one hand, would be legally bound by the duty of confidentiality to protect the interests of AI right-holders. On the other hand, they would be granted full transparency[13] to inspect and test these technologies in the interest of the public[14]. Inspired by the works of Frank Pasquale, this idea of mediated transparency is referred herein as *qualified transparency*[15].

[1] This article adopts the definition of "artificial intelligence" as articulated in the AI Act: "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Art.3(1) AIA).

[2] Stuart Russell and Peter Norvig, *Artificial Intelligence - A Modern Approach* ((4th Edition), 2021).

[3] See, for example, Ida Merete Enholm, Emmanouil Papagiannidis, Patrick Mikalef and John Krogstie, 'Artificial Intelligence and Business Value: A Literature Review' (2022) 24 *Information Systems Frontiers*.

[4] See, for example,. Natali Helberger and Nicholas Diakopoulos, 'ChatGPT and the AI Act' (2023) 12 Internet Policy Review; Oskar J Gstrein, Noman Haleem and Andrej Zwitter, 'General-Purpose AI Regulation and the European Union AI Act' (2024) 13 *Internet Policy Review*.

[5] See, for example, Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 *New Media & Society* 973; Madalina Busuioc, 'Accountable Artificial Intelligence: Holding Algorithms to Account' (2021) 81 *Public Administration Review*.

[6] See, for example, Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a Right to Explanation Is Probably Not the Remedy You Are Looking For' [2017] *SSRN Electronic Journal*.

[7] Peter Smith and Laura Smith, 'This Season's Artificial Intelligence (AI): Is Today's AI Really That Different from the AI of the Past? Some Reflections and Thoughts' [2024] *AI and Ethics*.

[8] Sylvia Lu, 'Regulating Algorithmic Harms' [2024] *Florida Law Review*.

[9] Charlotte Tschider, 'Legal Opacity: Artificial Intelligence's Sticky Wicket' [2021] *Iowa Law Review*.

[10] Charlotte Tschider, 'Beyond the "Black Box"' (2021) 98 *Denver Law Review;* Frank Pasquale, The Black Box Society : The Secret Algorithms That Control Money and Information (Harvard University Press 2015).

[11] See, for example, Pasquale (n 10).

[12] See, for example, Pasquale (n 10); Tschider (n 10); Busuioc (n 5).

[13] Paul B de Laat, 'Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?' (2018) 31 *Philosophy and Technology*.

[14] On the public interest in AI, see, for example,: Theresa Züger and Hadi Asghari, 'Introduction to the Special Issue on AI Systems for the Public Interest' (2024) 13 *Internet Policy Review*.

[15] Frank Pasquale, 'Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries' (2010) 104 *Northwestern University Law Review* 105; Pasquale (n 10).

In the European Union (EU), the need for addressing the negative implications of AI has been seen by the EU policymakers as instrumental in the *human-centric*[16] and *trustworthy* AI[17]. The AI Act[18] (AIA), enacted on 12 July 2024, lays down the foundations for this ambitious project. The complex scaffold of substantive rules is accompanied by the enforcement framework, in which AIA oversight bodies are responsible to ensure that the new rules are duly respected. To this end, they have been granted an array of competences, including the legal mandates to access – under the duty of confidentiality – the relevant information concerning AI systems.

This article examines the transparency mandates outlined above in light of the concept of qualified transparency. Drawing from the broader literature and policy documents on AI governance, the paper explores the meaning and role of qualified transparency theoretically, along with the conditions necessary for its effectiveness. Against this backdrop, the AI Act is analysed in terms of which governance bodies have been entrusted such transparency mandates under the AIA, and to what extent. Lastly, the paper discusses the potential challenges in rendering these transparency safeguards effective in meeting the AI Act's objectives.

It is argued that the AIA introduces the qualified transparency mechanism by vesting the appropriate legal mandates within the designated oversight bodies. Moreover, this high level of information disclosure is accompanied by the important factors needed for the sound functioning of qualified transparency, including the requirement of the technical and socio-legal expertise, due impartiality in the assessment process, and the capacity to apply appropriate enforcement measures. However, the article contends that the qualified transparency mechanism may be limited by the predominant reliance of internal conformity assessments of high-risk AI systems. Moreover, the enforcement of AIA may be challenging on the national level due to the extensive scope of oversight responsibilities and the remaining problem of the technical opacity in many AI systems. Furthermore, by requiring the disclosure of essential information concerning the AI systems, the AIA may curb the extent of the legal opacity stemming from the asserted AI proprietary claims.

The article is structured as follows. In Section 2, the risks associated with AI are outlined, along with the main problems linked to AI opacity. The section further characterises the concept of qualified transparency and identifies the key conditions necessary for its effective functioning. Section 3 analyses which national and EU institutions have been assigned the oversight transparency mandates vis-à-vis the AIA target actors, setting the stage for the discussion on the implementation of qualified transparency within the AI Act in Section 4. Section 5 highlights the possible challenges for the oversight bodies in rendering their transparency mandates operational and effective, with the conclusions drawn in Section 6.

2. Addressing the risks posed by AI technologies

The field of AI, in the broad sense, encompasses a family of techniques that enable machines to "compute how to act effectively in a wide variety of novel situations"[19]. Machine learning (ML) methods, which constitute a subfield of AI, are increasingly replacing traditional hand-coded algorithms[20]. Their versatile capabilities have enabled many innovations, such as automating repetitive and time-consuming tasks, processing large datasets, identifying complex patterns, and reaching higher levels of accuracy.[21] In particular, deep learning

[16] European Commission, 'Communication: Building Trust in Human Centric Artificial Intelligence | Shaping Europe's Digital Future' COM (2018) 168 final.

[17] AI HLEG, 'Ethics Guidelines for Trustworthy AI' (Communication EC, 2019) accessed 6 October 2023; European Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' (2020).

[18] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 2024/1689

[19] Russell and Norvig (n 2) 19. However, note that the term "AI" is not easily defined, see e.g. Larsson and Heintz (n 48)

[20] Christopher M Bishop and Hugh Bishop, *Deep Learning: Foundations and Concepts* (Springer Nature 2023).

[21] Enholm and others (n 3).

methods have emerged as "the most successful paradigm" in machine learning[22], proving their potential in a wide range of contexts, such as speech and image recognition, or natural language processing (NLP)[23].

At the same time, the very features that allow these technologies to perform so well may themselves become the source of problems, inadvertently or otherwise. The most advanced and powerful AI systems may encroach our privacy, especially if they identify our vulnerabilities and intimate details[24]. Such information may be further used to distort our autonomy by highly personalised manipulation[25], impact our political views[26] and mental health[27], especially in the long run. On a larger scale, what has been referred to as *high-reach* AI systems[28], such as recommender systems and LLMs, may discriminate societal groups and exacerbate inequalities[29], amplify the spread of harmful or illegal content online[30], and disrupt democratic processes[31]. Moreover, Galaz et al. highlight the negative impact of AI on sustainability and ecological contexts[32]. In her analysis of algorithmic harms, Sylvia Lu observes that while AI systems may cause physical harm, most AI-related harms are *immaterial*. Since such harms typically do not cause obvious inconvenience or immediate suffering, they are difficult to track down and redress on ongoing basis. Often downplayed as minor secondary problems, they accumulate over time, leading to the gradual erosion of the civil rights and values[33].

As algorithmic harms to the large extent stem from their intangible and cumulative characteristics, addressing them necessitate systemic approach[34]. However, in order for such systemic approach to be effective, it should first tackle two aggravating factors – inadequate accountability frameworks, which Lu terms as *accountability paucity*, and the issue of *algorithmic opacity*, which further obstructs harm detection[35].

In the EU, the above corrective measures have been largely deficient on both accounts. Before the AI Act was adopted, the EU policy-makers stated themselves that national authorities responsible for compliance with safety and fundamental rights rules "do not have powers, procedural frameworks and resources to ensure and monitor compliance of AI development and use with applicable rules".[36] With regard to the accountability paucity, although AI technologies have already been subject to various EU legal frameworks, the existing rules have been inadequate to bridge the accountability gaps created by AI technologies. For example, the EU product safety legislation would apply "to products and not to services, and therefore in principle not to services based on AI technology either"[37]. Moreover, the governance frameworks would predominantly focus on safety risks present at the time of the product placement on the market, which is problematic for AI with dynamic ML components. Furthermore, AI poses new challenges to the liability rules, and creates new forms of risk, often eluding the oversight scrutiny[38]. In addition, the EU regulations

[22] Bishop and Bishop (n 20) V.

[23] Christian Janiesch, Patrick Zschech and Kai Heinrich, 'Machine Learning and Deep Learning' (2021) 31 *Electronic Markets*.

[24] See, for example, Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2019) 2019 *Columbia Business Law Review*.

[25] See, for example, Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (Profile Books 2019).

[26] See, for example, Hunt Allcott and others, 'The Welfare Effects of Social Media' (2020) 110 *American Economic Review* 629.

[27] See, for example, Betul Keles, Niall McCrae and Annmarie Grealish, 'A Systematic Review: The Influence of Social Media on Depression, Anxiety and Psychological Distress in Adolescents' 25 *International journal of adolescence and youth* 79.

[28] K Söderlund and others, 'Regulating High-Reach AI: On Transparency Directions in the Digital Services Act' (2024) 13 *Internet Policy Review*.

[29] See, for example, Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2018) *SSRN Electronic Journal*; Cathy O'Neil, Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy (2016);

[30] See, for example, Soroush Vosoughi, Deb Roy and Sinan Aral, 'The Spread of True and False News Online' (2018) 359 Science 1146; Manoel Horta Ribeiro and others, 'Auditing Radicalization Pathways on YouTube', FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020).

[31] See, for example, Vosoughi, Roy and Aral (n 30).

[32] Victor Galaz and others, 'Artificial Intelligence, Systemic Risks, and Sustainability' (2021) 67 Technology in Society 101741.

[33] Lu (n 8).

[34] Lu (n 8).

[35] Lu (n 8).

[36] European Commission, 'Impact Assessment of the Regulation on Artificial Intelligence' (2021) 13 accessed 6 October 2023.

[37] European Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' (n 17) 14.

[38] European Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' (n 17).

applicable to AI would primarily rely on individual remedies[39]. For instance, the EU General Data Protection Regulation (GDPR)[40] provides individuals with the right to *meaningful information about the logic involved* in automated decision-making which significantly affects the individuals concerned[41]. However, the exercise of such rights under the personal data and consumer protection laws have been explicitly restricted by the trade secrets or intellectual property rights.[42]

Nonetheless, the EU legal frameworks already include a rich body of regulations, governing the personal data and consumer protection, liability, anti-discrimination, competition laws, and digital services rules specifically targeting the large-scale effects of algorithms. It could therefore be argued that many of the AI-related harms could have been largely mitigated under the existing EU laws, were it not for the legal opacity of AI.

2.1 The opacity of AI systems

AI systems may be opaque due to the inherent technical complexity of certain classes of ML algorithms – the issue commonly referred to as the *black-box problem*[43]. The multilayered operations and/or the composite interactions of various algorithms working together render many AI systems difficult to examine, particularly deep learning neural networks[44]. This complexity in many cases extends the human cognitive capacity[45], including that of AI developers themselves. To prevent such limitations from impeding the uptake of AI technologies, fields like *eXplainable AI* (xAI) have emerged with the aim to develop more interpretable AI models and offer solutions to decipher opaque AI algorithms [46]. Thus, xAI algorithms are often used to “translate” the black-box algorithms, enabling humans to “effectively manage the emerging generation of artificially intelligent partners”[47].

Yet, the opacity of AI systems may go well beyond the issue of algorithmic opacity[48]. Although some AI systems are open-source[49], most of AI providers choose to keep the inner workings of AI systems confidential. Hence, even simplest algorithms – such as decision trees – may effectively be a “black-box” for third-parties[50]. This additional layer of opacity is provided by such legal mechanisms as the trade secrecy protection, business confidentiality, non-disclosure agreements (NDAs), or similar confidentiality agreements[51]. However, the trade secret protection is the most common approach used by AI rights holders to safeguard their proprietary interests.[52]

[39] European Commission, ‘Impact Assessment of the Regulation on Artificial Intelligence’ (n 36); Edwards and Veale (n 6); Mateusz Grochowski and others, ‘Algorithmic Transparency and Explainability for EU Consumer Protection: Unwrapping the Regulatory Premises’ (2021) 8 Critical Analysis of Law.

[40] Regulation 2016/679, ‘Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)’ OJ L 119.

[41] Art. 15 (1)(f) GDPR.

[42] See, for example, Rec. 63 GDPR states that the right of data access “should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software”. Grochowski and others (n 39).

[43] See, for example, Pasquale (n 10); Davide Castelvecchi, ‘Can We Open the Black Box of AI?’ (2016) 538 Nature.

[44] See, for example, Arun Rai, ‘Explainable AI: From Black Box to Glass Box’; J Kemper and D Kolkman, ‘Transparent to Whom? No Algorithmic Accountability without a Critical Audience’ (2019) 22 Information Communication and Society 2081.

[45] See, for example, Kemper and Kolkman (n 44).

[46] Alejandro Barredo Arrieta and others, ‘Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI’ (2020) 58 Information Fusion.

[47] Barredo Arrieta and others (n 46) 83.

[48] On the distinction between the terms “transparency of AI systems” and “algorithmic transparency” see Stefan Larsson and Fredrik Heintz, ‘Transparency in Artificial Intelligence’ (2020) 9 Internet Policy Review 1.

[49] See, for example, Kemper and Kolkman (n 44).

[50] Busuioc (n 5).

[51] See, for example,. Pasquale (n 10).

[52] See, for example,. Katarina Foss-Solbrekk, ‘Three Routes to Protecting AI Systems and Their Algorithms under IP Law: The Good, the Bad and the Ugly’ (2021) 16 Journal of Intellectual Property Law & Practice 247 ; Joshua A Kroll and others, ‘Accountable Algorithms’; Sharon K Sandeen and Tanya Aplin, ‘Trade Secrecy, Factual Secrecy and the Hype Surrounding AI’, Research Handbook on Intellectual Property and Artificial Intelligence (2022); Tschider (n 9).

The trade secrecy protection is regulated both internationally[53], and within the EU by the Trade Secrets Directive[54] (TSD). Following the TSD, a broad range of information may potentially be claimed by AI firms as trade secrets[55]. However, to be protected as such, the information concerned must meet three cumulative criteria stipulated in Article 2 TSD: it must be secret (not "generally known" or "readily accessible"), its commercial value must stem from its secrecy, and reasonable steps must have been taken by the right-holders to preserve the secrecy. Crucially, this kind of legal protection may last infinitely. The information may be disclosed lawfully only in certain cases enumerated in the TSD. One of such grounds includes the EU or national rules requiring to "disclose, for reasons of public interest, information, including trade secrets, (...) to administrative or judicial authorities for the performance of the duties of those authorities"[56].

In general, the trade secrecy laws have been introduced to foster commercial morality and technological progress[57]. From this perspective, the motivation to safeguard the software's competitive advantage may be considered as a valid interest within the scope of competition law, as uncontrolled disclosure of sensitive information would allow competitors to replicate the algorithms [58]. Moreover, maintaining a certain level of confidentiality has been seen as justified in preventing the attempts of gaming or manipulation of the systems by end-users[59]. Likewise, revealing too much information about the algorithmic workings could obstruct law enforcement[60], and in some cases, affect the privacy of individuals whose data had been used to train AI models[61].

However, the downside of the widespread use of the trade secrecy is that it makes harmful practices harder to detect and challenge. The confidentiality mechanisms have been shown to be commonly used as a form of self-protection of AI companies[62] from external scrutiny to hide discriminatory, anticompetitive, careless, or otherwise wrongful practices[63]. Apart from the legitimate interests pointed to above, in many cases confidentiality rules are being used to conceal, for instance, "under-representativeness of databases, programming deficiencies, cognitive biases instilled by AI developers, unauthorized collection of sensitive data"[64].

Notably, the fact that the trade secrecy is the most common avenue to safeguard sensitive information concerning AI systems exposes the more fundamental problem within the IP law. Legal scholars point out that although AI systems constitute intellectual property (IP) and should be protected as such, they usually not easily qualify for protection under the patent rules (as it is difficult, for example, to prove the novelty or the inventive step) or copyright frameworks (e.g. the condition of creativity may not be met). At the same time, as Foss-Solbrekk observes, trade secrecy appears to be the preferred form of legal protection by AI providers, since the criteria for trade secrecy may be easier to meet than those for patents and copyrights[65]. As the author argues, a "logical route" for protection of AI technologies should be the patent regime, which has been specifically designed for technical inventions[66]. This view is shared by Tschider, who highlights the

[53] WTO's Agreement on Trade-related Aspects of Intellectual Property Rights (the TRIPS Agreement); all Member States as well as the Union are part of this international treaty (Rec. 5 TSD).

[54] Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure [2016] OJ L 157/1.

[55] Lu (n 8); Wachter and Mittelstadt (n 24).

[56] Art. 1(2)(b) TSD.

[57] Peter S Menell, 'Tailoring a Public Policy Exception to Trade Secret Protection' (2017) 105 California Law Review 1.

[58] Tschider (n 9).

[59] Pasquale (n 10); de Laat (n 13).

[60] Pasquale (n 10); Tal Z Zarsky, 'Transparent Predictions' (2013) 2013 University of Illinois Law Review 1503 .

[61] Zarsky (n 60); Kroll and others (n 52); Pasquale (n 10); de Laat (n 13).

[62] Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society; Margot E Kaminski, 'Understanding Transparency in Algorithmic Accountability', *The Cambridge Handbook of the Law of Algorithms* (Cambridge University Press 2020).

[63] Burrell (n 62); de Laat (n 13); Nicholas Diakopoulos, 'Algorithmic Accountability Reporting: On the Investigation of Black Boxes' [2014] Tow Center for Digital Journalism; Pasquale (n 10).

[64] Lu (n 8) 38.

[65] Katarina Foss-Solbrekk and Ann Kristin Glenster, 'The Intersection of Data Protection Rights and Trade Secret Privileges in 'Algorithmic Transparency', Research Handbook on EU Data Protection Law (2022).

[66] Foss-Solbrekk (n 52) points specifically to the computer-implemented inventions (CIIs) as a potential alternative.

benefits of the patent framework as serving broader purposes than only economic incentives[67]. Meanwhile, however, the insufficiency of other legal tools to protect AI technologies renders trade secrecy a "convenient mechanism for companies to fill the gap where classic IP law fails." [68]

2.2 Qualified transparency as a proposed solution

Clearly, as the trade secrecy protection has become "a standard procedure in many business environments"[69], it is not an ideal legal instrument from the standpoint of seeking to safeguard other important values and interests, including those of end-users, competitors, and the general public[70]. In view of the growing negative implications of AI, regulatory intervention has been seen as necessary in calibrating the conflicting interests between the AI providers and other AI stakeholders[71]. However, since the possibility to protect AI systems under the umbrella of the traditional IP legal frameworks is largely excluded, most of the proposed solutions would centre around the potential transparency routes within the trade secrecy regime. For example, the "whistleblowing" activities[72] or reverse engineering[73] have been explicitly listed as lawful ways of information disclosure under the TSD. However, once disclosed through such mechanisms, the information can no longer be considered secret.

In response to the above issues, an alternative approach has been proposed, involving the delegation of compliance verifications to trusted third-party entities. In order to fulfil their responsibilities, such "transparency intermediaries" would be granted full access to the relevant information, also including the information normally classified as secret by the AI right-holders. Crucially, such third-party bodies would be legally bound by the duty of confidentiality to maintain the trade secrecy protection, in line with the trade secrecy laws. Frank Pasquale has termed this mechanism as *qualified transparency*, since this level of transparency "should be qualified in order to protect important intellectual property interests"[74]. While for the time being this kind of disclosure is in most cases granted only through court proceedings, Pasquale argues that regulators need to develop institutional capacity to create an alternative to the traditional litigation. After all, as he phrases it, "[a]gencies ought to be able to 'look under the hood' of highly advanced technologies"[75] when intentional opacity used by corporations precludes public scrutiny. A similar idea concerning the full disclosure to designated entities under the confidentiality regime has been proposed by Kaminski, who refers to such legal mandate as *systemic transparency*[76]. This could mean an information disclosure to a "board of technical experts", who would "get access to an algorithm's source code, training data sets, and interviews with the data scientists designing the system"[77]. To check private actors from acting only in their own self-interest[78], the aim of systemic transparency would be "to make visible error, bias, and discrimination in both machine and human systems, so they can be addressed and mitigated, if not corrected"[79]. European legal scholars, such as Wachter and colleagues, suggest that a *trusted third party* would be an "ideal solution", allowing "for examination of automated decision-making systems, including the rationale and circumstances of specific decisions"[80]. As the authors reason, this approach would limit "the risk to data controllers of exposing trade secrets, while also providing an oversight mechanism for data subjects that can operate when explanations are infeasible or too complex for lay comprehension." Further,

[67] Tschider (n 9).
[68] Foss-Solbrekk (n 52) 248.
[69] Lu (n 8).
[70] Tschider (n 9).
[71] Pasquale (n 10); de Laat (n 13); Kaminski (n 62).
[72] Menell (n 57).
[73] Diakopoulos (n 63).
[74] Pasquale (n 15) 161.
[75] Pasquale (n 10) 169.
[76] Kaminski (n 62).
[77] Kaminski (n 62).
[78] Kaminski (n 62).
[79] Kaminski (n 62) 129.
[80] Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 98.

related concepts have been put forth as a *trusted intermediary* by Menell[81], *an escrow third-party auditor* by Diakopoulos[82], or as a *neutral arbiter* by Crawford and Schultz[83].

While the information disclosure could be seen as the first step in the course of accountability objectives, seeing does not always produce *understanding*, in the sense of the "knowledge required to govern and hold systems accountable"[84]. Such third-party scrutiny of AI systems should obviously concern the technical aspects of the systems for which high level of technical literacy[85] is needed. However, the social processes that are influencing design decisions should be taken into account as well. Since such decisions often embed "human values and ideologies either inadvertently or by choice"[86], the deployment of AI systems on scale may have the significant societal impact on such issues as privacy, autonomy, non-discrimination, and/or the democratic discourse. Thus, a comprehensive assessment of AI systems should indeed consider AI systems as "algorithmic assemblages of humans and non-humans" working together.[87]

Moreover, it is important that the qualified transparency mandate is exercised in an independent way from external pressures[88]. In this context, Pasquale brought up an illustrative example of the US Federal Trade Commission (FTC) which in 2012 was tasked to determine whether Google had manipulated search results to increase the visibility of its own services while decreasing the visibility of actual or potential competitors. However, soon after the FTC recommended a suit against Google, the investigations were closed before the case reached the court. The FTC released only a short statement to the public, explaining that various "websites have experienced demotions (...) as a consequence of algorithm changes that also could plausibly be viewed as an improvement in the overall quality of Google's search results"[89]. Pasquale pointed out that this decision was made behind the closed doors, without providing the public with the details of the grounds of this conclusion, suggesting that the decision was "overruled by politically appointed commissioners"[90]. Conversely, in the EU, similar allegations prompted thorough investigations, leading the European Commission to impose a €2.42 billion fine on Google in 2017 for violating EU competition laws[91].

Finally, once reaching the level of understanding needed to "govern and hold accountable" the AI system, the oversight body should have at its disposal the capacity to impose regulatory sanctions. As "companies watch the past to predict the future"[92], the purpose of the administrative sanctions is not only to punish the offenders, but also to prevent the future, similar occurrences. Admittedly, sanctioning measures may be seen as going beyond the scope of qualified transparency in the strict sense. However, the *sanction of authority*[93] has been pointed to as an inseparable element of an effective AI governance framework, considering that the "idea of transparency can lose its purpose if it fails to produce meaningful effects"[94].

How the qualified transparency mechanism could be incorporated within the legal frameworks has been proposed in several ways, with approaches ranging from centralised to decentralised models, either as part of the government or independent from it. In the US context, Andrew Tutt advocates for the formation of an "FDA for algorithms" – that is, an equivalent for the US Federal Drug Administration – a single agency

[81] Menell (n 57).
[82] Diakopoulos (n 63).
[83] Kate Crawford and Jason Schultz, 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms' (2014) 55 Boston College Law Review.
[84] Ananny and Crawford (n 5) 974.
[85] Burrell (n 62); Pasquale (n 10).
[86] Riikka Koulu, 'Crafting Digital Transparency: Implementing Legal Values into Algorithmic Design' (2021) 8 Critical Analysis of Law.
[87] Ananny and Crawford (n 5); Kemper and Kolkman (n 44).
[88] Pasquale (n 10).
[89] FTC, 'Statement of the Federal Trade Commission Regarding Google's Search Practices'.
[90] Pasquale (n 10) 164.
[91] European Commission, 'Antitrust: Commission Fines Google €2.42 Billion - Press Release' (Brussels, 27 June 2017).
[92] W Gregory Voss and Hugues Bouthinon-Dumas, 'EU General Data Protection Regulation Sanctions in Theory and in Practice' (2021) 37 Santa Clara High Technology Law Journal.
[93] Bran Knowles and John T Richards, 'The Sanction of Authority: Promoting Public Trust in AI' (2021) Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.
[94] Ananny and Crawford (n 5).

responsible for regulation and enforcement of rules governing AI technologies used across the US. In the EU, an analogous idea of a centralised oversight body such as a "European Agency for Artificial Intelligence" has been suggested by Stahl [95]. In the same vein, Wachter, Mittelstadt, and Floridi propose that "a European regulator could be created specifically for auditing algorithms". Alternatively, the authors suggest that the role of trusted third parties could be incorporated within the existing national authorities[96]. Other legal scholars, such as Zarsky[97], as well as Edwards and Veale, consider also NGOs or civil society scrutiny organisations as viable options "to review the accuracy, lack of bias and integrity of a ML system"[98].

As will be shown below, the AI Act entrusts the qualified transparency mandates to various bodies – EU institutions, public authorities and private accreditation bodies governed by the public authorities – on both national and EU levels[99], depending on the type of AI technology.

3. Qualified transparency for oversight purposes in the AI Act

In the EU, the recognition of both opportunities and challenges posed by AI have underpinned European's own, "third way" approach. The EU policymaking has emphasised the need to develop a "solid European framework"[100] by fostering the development, deployment and use of AI which is in line with the EU's existing fabric of values and rules. The AI Act now constitutes the cornerstone for the governance framework for AI technologies, with the EU policy ideas of human-centric and trustworthy AI incorporated in Art. 1 AIA as the overarching objective. On the one hand, the new rules are intended to promote the uptake of AI technologies and support innovation, including through so-called *regulatory sandboxes*. On the other hand, the Regulation aims to safeguard a "high level of protection of health, safety, fundamental rights", including democracy, and environmental protection[101]. Notably, however, the emphasis in the AI Act is on the latter objective, i.e. on "addressing the risks associated with certain uses of such technology", so that "people can trust that the technology is used in a way that is safe and compliant with the law"[102].

While the scope of application of the AI Act encompasses all AI technologies operating in the Union, the framework classifies them into a few risk categories, with each category determining the corresponding regulatory implications. The AI uses which are deemed as to pose unacceptable risk to health, safety, environment and/or fundamental rights have been prohibited[103], other AI systems have been subject to limited transparency requirements[104]. A significant part of the AIA is dedicated to specifying the governance rules for AI systems considered as posing high risk to the above values, and these have been listed in Annex I[105] and Annex III[106] AIA. This category of AI systems has been subject to a set of requirements concerning the

[95] Bernd Carsten Stahl and others, 'A European Agency for Artificial Intelligence: Protecting Fundamental Rights and Ethical Values' (2022) 45 Computer Law and Security Review.

[96] Wachter, Mittelstadt and Floridi (n 80).

[97] Zarsky (n 60).

[98] Edwards and Veale (n 6) 23.

[99] The enforcement of EU laws may involve EU and/or national enforcement bodies. On the various ways of designing the EU law enforcement frameworks, see e.g. K. Söderlund and S. Larsson, 'Enforcement Design Patterns in EU Law: An Analysis of the AI Act' (2024) 3 Digital Society 2024 3:2 1.

[100] European Commission, 'Communication Artificial Intelligence for Europe' COM (2018) 237 final.

[101] Art. 1 AIA.

[102] European Commission, Explanatory Memorandum to the Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence.

[103] This group includes, for example, AI systems that use subliminal techniques to manipulate behaviour or cause harm, social scoring, scraping of facial images from the internet or CCTV, predictive policing of individuals, real-time remote biometric identification for law enforcement, except in narrowly defined situations (Art. 5 AIA).

[104] Art. 50 AIA concerns such applications as chatbots, deepfakes and emotion recognition systems, requiring the disclosure that the AI technology is (or has been) used.

[105] Annex I contains the list of the sectors covered by Union harmonisation legislation (now also the AIA), for instance safety components of vehicles and medical devices.

[106] Annex III includes such AI uses as biometric identification systems, critical infrastructure (e.g. energy, transport), access to employment, education, public services, and law enforcement.

quality and risk management systems, data governance, technical documentation, transparency and provision of information to deployers, human oversight, and record-keeping (logs)[107].

Moreover, the rapid and widespread adoption of generative AI, such as Chat GPT, has prompted the creation of " a general-risk category in its own right"[108], during the last year of works on the AI Act. Such systems have been termed *General Purpose AI* (GPAI), and those which may have a significant impact on the EU market due to their reach or their potential negative effects on public health, safety, public security, fundamental rights[109], have been designated as *GPAI with systemic risk*. It should be noted, however, that the majority of the currently deployed AI systems would fall within the low-risk category[110], which is not subject to any binding rules.

The AIA framework has been largely embedded within the broader context of the EU harmonised product legislation – the New Legislative Framework (NLF) – with the Regulation on market surveillance[111] (MSR) incorporated and directly applicable as *lex generalis* to the AIA[112]. In general, the NLF covers certain groups of harmonised products, such as toys, medical devices, machinery, lifts, and protective equipment. The Framework constitutes one of the EU measures seeking to protect the EU citizens as consumers, with the aim to ensure that unsafe products do not circulate within the EU.[113]

Before they may enter the EU market, harmonised products need to comply with *essential requirements* or *harmonised standards.*[114] Compliance with the relevant harmonised standards establishes the *presumption of conformity* with the essential requirements. Notably, certain products are subject to the conformity assessment by notified bodies (i.e. accreditation organisations operating under the NLF), as in case of certain classes of medical devices. The EU declaration of conformity (i.e. CE-marking) is affixed to the product once it is compliant with the stipulated requirements, and the product may be introduced in the EU market. The adherence to the applicable rules is overseen thereafter by the relevant *market surveillance authorities* (MSAs).

The AIA harmonised standards, which at the time of writing are currently being developed by the EU standard harmonisation bodies (CEN and CENELEC)[115], will be applicable to many of the requirements concerning high-risk AI, including the quality and risk management systems, technical documentation, data governance, etc. Moreover, situating the AIA within the NLF to the large extent reflects the choice of responsible oversight bodies, their enforcement powers and transparency mandates.

Before diving into the analysis, it is important to highlight that the AIA imposes the duty of confidentiality on entities carrying out the enforcement tasks. Art. 78 AIA requires each of the enforcement bodies and "any other natural or legal person involved" to "respect the confidentiality of information and data obtained" so as to protect, in particular, "the intellectual property rights and confidential business information or trade secrets (...), including source code".

[107] See Chapter 3 AIA.

[108] Helberger and Diakopoulos (n 4) 3.

[109] Art. 3 (65) AIA.

[110] European Commission, 'AI Act | Shaping Europe's Digital Future' (2024) <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> accessed 23 October 2024.

[111] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011 [2019] OJ L 169/1

[112] Art. 74 AIA.

[113] Charter of Fundamental Rights of the European Union [2000] OJ C364/1, Art. 38, "Union policies shall ensure high-level of consumer protection". See also in general: Stephen Weatherill, 'Product safety regulation' (2014).

[114] See, for example, Sybe de Vries, Olia Kanevskaia and Rik de Jager, 'Internal Market 3.0: The Old "New Approach" for Harmonising AI Regulation' (2023) 2023 8 European Papers - A Journal on Law and Integration 583.

[115] See Annex I to the Commission's standardisation request available at https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en

3.1 The AI Act enforcement on the national level

The national enforcement frameworks of the AIA will consist of at least two authorities – one notifying authority (overseeing the notified bodies), and one market surveillance authority (MSA)[116]. However, the Member States have the flexibility to designate a few MSAs, including by integrating the oversight responsibilities within their existing regulatory structures. In such cases, one of the MSAs should act as a *single point of contact*[117].

3.1.1 *Market surveillance authorities*

Following Article 74 AIA, the MSAs have been entrusted the responsibility to oversee all AI operators and AI applications on the national level. This would therefore involve the supervision of the high-risk AI systems, the GPAI models operating on the national territory, as well as the ongoing monitoring of the market with respect to the prohibited practices and the limited transparency obligations.

Under the MSR, the investigation powers of the MSAs to realise their responsibilities are very broad. These may include the authority to "require economic operators to provide relevant documents, technical specifications, data or information on compliance and technical aspects of the product", the power to carry out unannounced on-site inspections, as well as to enter any premises in the course of investigations[118].

Moreover, with regard to the high-risk AI, the AI Act further stipulates in Article 74 (12) AIA that MSAs are granted "full access by providers to the documentation" produced for the purposes of the AIA, as well as access to the "training, validation and testing data sets used for the development of high-risk AI systems". Details of the information required are further listed in the relevant annexes to the AIA. Following Annex IV AIA for example, the technical documentation should include the methods and steps performed for the development of the AI system, the key design choices including the rationale and assumptions made, the validation and testing procedures used. Moreover, when testing and auditing procedures based on the provided information prove insufficient, upon a "reasoned request", the MSAs may access the source code if it is necessary to assess the conformity of the high-risk AI system with the AIA[119].

The Member States are responsible for ensuring that the MSAs operate "independently, impartially and without bias so as to safeguard the objectivity"[120] of their activities, and that the MSAs have at their disposal "adequate technical, financial and human resources" to conduct their enforcement tasks. The competences of their personnel should cover a broad array of expertise, including "in-depth understanding of AI technologies", personal data protection, fundamental rights, health and safety risks[121].

The enforcement measures on the national level provided within the AIA range from the ones that are handled by the MSAs, to the ones that are imposed as penalty sanctions by the Member States. For example, in case of lacking CE-marking or technical documentation the MSA may request the AI provider to recourse such non-compliance within the specified time. If the issue persists, the MSA may restrict or prohibit the availability of the high-risk AI system on the market, or ensure that it is recalled or withdrawn[122]. The Member States are tasked to specify the rules on the penalties and other enforcement measures (including warnings and non-monetary penalties), which should be "effective, proportionate and dissuasive". However, the AIA states that the monetary penalties should vary between 1% and 7% of the total worldwide annual turnover for the preceding financial year.[123]

[116] The Member States should report their domestic enforcement arrangements by August 2025. See e.g.: Kai Zenner, 'The EU AI Act: Responsibilities of the Member States' accessed 9 September 2024.

[117] Art. 70 AIA.

[118] Art. 14 (4) MSR.

[119] Art. 74 (13) AIA.

[120] Art. 70 (1) AIA.

[121] Art. 70 (3) AIA.

[122] Art. 83 AIA.

[123] Art. 99 AIA.

3.1.2 Other national authorities

Apart from the MSAs, the AIA provides that the national public authorities supervising the obligations to respect fundamental rights protection laws have also been granted similar transparency powers with regard to the high-risk AI systems listed in Annex III. Hence, on the basis of Article 77 AIA, authorities such as Data Protection Authorities and national ombudsmans may access any documentation created for the AIA purposes when it is necessary to effectively fulfill their mandates. Where the information provided is insufficient to determine whether an infringement concerning fundamental rights has occurred, these authorities may request the MSAs to organise additional testing of the AI system.

3.1.3 Notified bodies

Under the AIA, the involvement of notified bodies (NBs) will in some cases be required, and in some cases optional for high-risk AI providers. The engagement of NBs is compulsory for high-risk AI systems falling within the scope of Annex I, and which are already subject to third-party conformity assessments under the NLF[124]. For instance, this concerns AI systems qualified as medical devices under the Medical Devices Regulation[125] (MDR). Such third-party conformity assessment would be conducted before the product enters the market, and periodical reviews may follow as part of the post-market monitoring activities. Although NBs are accredited and supervised by national notified authorities, the AI providers are free to choose any of the notified bodies within the EU[126].

In contrast, once the relevant harmonized standards are in place, the involvement of NBs for conformity assessment of the high-risk AI listed under Annex III AIA will be merely optional[127]. Instead, the AI providers may choose to follow the internal conformity assessment procedure[128].

The AIA stipulates that NBs should be organized and operated “so as to safeguard the independence, objectivity and impartiality” of their activities[129]. Furthermore, they should have “sufficient administrative, technical, legal and scientific personnel” who possesses “experience and knowledge relating to the relevant types of AI systems, data and data computing”[130].

When the employment of notified bodies is required or opted for, NBs would have access to largely the same documentation and information as MSAs above[131] (see Annex VII). Yet, in contrast to the competences of the MSAs, the NB have not been provided the option of accessing the source code[132]. This is notable, particularly in light of the fact that the initial AIA proposal stated otherwise[133]. Still, NBs may require that the provider supply further information and/or carry out additional tests. If the NB is not satisfied with the tests performed by the provider, it may conduct the necessary tests itself[134].

3.2 The AI Act enforcement on the EU level

While there are other institutions involved in the AIA governance framework on the EU level[135], the investigative powers vis-à-vis AI providers have been vested in the AI Office with regard to the GPAI

[124] Art. 43(3) AIA.

[125] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [2017] OJ L 117/1.

[126] Art. 43(1) AIA.

[127] Art. 43(1) AIA.

[128] Annex VI AIA.

[129] Art. 31(6) AIA.

[130] Art. 31(11) AIA.

[131] Annex VII (4.3) AIA.

[132] Following Para. 4.5. Annex VII AIA, NBs have been granted access to relevant parameters instead.

[133] Cf. Para 4.5. Annex VII, European Commission, ‘Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence’ (n 102).

[134] Annex VII (4.4) AIA.

[135] More broadly, the AIA governance framework on the EU level consists of the Commission, AI Office, the Board of AI, the panel of independent experts, and the advisory forum.

models[136]. In addition, the panel of independent experts ("scientific panel") has been created to support the AI Office and the national authorities in their enforcement tasks.

3.2.1 AI Office

The AI Office, apart from the broad range of other tasks under the AIA[137], has been granted equivalent market surveillance powers in cases when the GPAI models and systems are developed by the same provider[138]. In this capacity, AI Office may access any documentation created under the AIA concerning the GPAI[139] (including GPAI with systemic risk). Such documentation would include, inter alia, technical documentation of the model, training and testing process, and the results of its evaluation[140]. Moreover, the AI Office may access "any additional information that is necessary for the purpose of assessing compliance of the provider"[141].

Eventually, the investigations concerning GPAI by the AI Office may lead to adoption of appropriate enforcement measures. If the evaluations give rise to "serious and substantiated concern of a systemic risk at Union level", the AI provider is required to adopt appropriate mitigation measures, or the AI product may be restricted from the market, withdrawn or recalled (as in the case of high-risk AI systems)[142]. Moreover, the GPAI providers may face administrative penalties for non-compliance with up to 3 % of their annual total worldwide turnover in the preceding financial year[143].

3.2.2 Panel of independent experts

The Scientific Panel, consisting of experts "having particular expertise and competence and scientific or technical expertise in the field of AI"[144], has been established primarily to support the enforcement activities of the AI Office. In particular, this regards the monitoring of the GPAI models, where the experts may provide qualified alerts to the AI Office which could potentially trigger the aforementioned investigations. The experts may also be appointed to carry out evaluations of GPAI to assess their compliance or to investigate systemic risks of the GPAI models[145]. Additionally, in the course of their enforcement activities, the Member States will be able to request support from the panel's experts[146]. Thus, when engaged in the supporting activities of the AI Office and/or the MSAs, the experts may also have access to the same documentation concerning AI technologies under the AIA when it is necessary to perform their tasks.[147]

4. Qualified transparency under the AI Act

Given that the AI Act has been primarily adopted to address the harmful effects of AI systems in the Union, the above analysis suggests that it indeed shows promise in this regard. The AIA challenges the problem of *accountability paucity* by imposing regulatory obligations on certain groups of AI systems, in accordance with their level of risk. Moreover, it addresses the issue of *legal opacity* of AI by granting oversight bodies the legal mandates and enforcement tools to ensure the AIA compliance. At the same time, the MSAs, the notified bodies, the AI Office, and anyone obtaining the information in the course of the investigations are bound by the duty of confidentiality. The AIA further stipulates that the oversight bodies should be equipped with a broad range of expertise, act in an impartial way, and their investigations may be followed by appropriate, dissuasive enforcement measures. The way the AIA approaches the problem of AI opacity can thus be seen as reflecting the idea of *qualified transparency*. It is also noteworthy that the AIA incorporates

[136] Art. 88(1) AIA.

[137] Kai Zenner, Philipp Hacker and Sebastian Hallensleben, 'A Vision for the AI Office: Rethinking Digital Governance in the EU' (2024) <https://www.euractiv.com/section/artificial-intelligence/opinion/a-vision-for-the-ai-office-rethinking-digital-governance-in-the-eu/> accessed 24 July 2024.

[138] Art. 75(1) AIA.

[139] Art. 53(1) AIA, with the exeption of open-source GPAI (see Art. 53(2)AIA).

[140] Art. 53(1) AIA.

[141] Art. 91(1) AIA.

[142] Art. 93 AIA.

[143] Art. 101 AIA.

[144] Art. 68(2) AIA.

[145] Art. 92 AIA.

[146] Art. 69 AIA

[147] Arts. 68 and 91 AIA.

transparency mandates for national competent authorities monitoring the fundamental right frameworks. This potentially creates an opportunity for national competent authorities to close the accountability gaps in other EU legal frameworks.

It appears, however, that the qualified transparency under the AI Act provides varying levels of disclosure for oversight bodies. On the basis of both AIA and MSR, the MSAs have been granted the most far-reaching enforcement powers and their responsibilities for market surveillance would concern all AI risk categories. Regarding the high-risk AI, the MSAs may access the documentation required under the AIA, and as the last resort, they have been granted the possibility to access the source code of the software. The AI Office has similar competences, but the scope of its market surveillance responsibilities is limited to GPAI in cases when the model and system are developed by the same provider. Both NBs and the competent authorities responsible for fundamental rights oversight may request access to the AIA documentation concerning high-risk AI systems, in accordance with their scope of competences. In contrast to the MSAs, however, the final text of AIA does not provide the possibility for NBs to access the system's source code. Arguably, this limitation has been introduced in response to the concerns of AI providers regarding the risk of insufficient confidentiality safeguards in the NBs[148].

Moreover, it is interesting to note that the AIA does not deal directly with the issue whether the specific information is covered by the trade secrecy or other forms of legal protection. In fact, the AIA stresses that the application of the new rules should be performed without preclusion of the trade secrets and business confidentiality[149]. Instead, it requires that the details specified under the AIA (and the harmonized standards, when applicable) concerning the AI systems should be disclosed to the oversight authorities upon their request. Hence, it could be argued that the AI Act cuts through the legal opacity conundrum by requiring the AI providers to disclose information that is necessary to assess the conformity with binding rules, regardless of the kind of legal protection invoked. This solution thus appears to follow the route opened up by the TSD, with the AIA providing the legal basis for disclosure of information (including trade secrets) to administrative authorities for reasons of public interest[150].

Furthermore, the transparency requirements of the AIA could result in the AIA effectively limiting the scope of confidentiality claims asserted to protect AI systems, at least with regard to the high-risk AI and GPAI. The trade secrecy regime would not be appliable to the extent the AI providers follow the relevant harmonised standards. This would be in accordance with the TSD's provisions which exclude the applicability of trade secrecy for information that is "generally known" or "readily accessible"[151].

In light of the above analysis it appears that the AIA establishes foundations for regulatory mechanisms which could mitigate the negative implications of AI outlined in Section 2. However, the effectiveness of the new rules will largely depend on their application, including the extent to which the qualified transparency mandate will be utilised by the oversight bodies. In this context, some of the potential obstacles in exercising the above transparency mandates will be pointed to below.

5. The potential limitations of qualified transparency

While the AIA provides the legal grounds for information disclosure concerning AI systems, primarily for the MSAs, the transparency mandate ultimately depends on their discretion whether to intervene. Moreover, as previously mentioned, information disclosure does not equate to "knowledge required to govern and hold

[148] For example, Veale and Borgesius point to the lack of transparency in NBs activities, particularly due to frequent outsourcing of their tasks. Michael Veale and Frederik Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach' (2021) 22 Computer Law Review International 97.

[149] Art. 78 AIA.

[150] Art.1 (2)(b) TSD

[151] See, for example, Ulla Maija Mylly, 'Transparent AI? Navigating Between Rules on Trade Secrets and Access to Information' (2023) 54 IIC International Review of Intellectual Property and Competition Law 1013 accessed 7 August 2024.

systems accountable"[152]. In turn, transparency – even qualified transparency – does not automatically lead to accountability and sanctioning measures. Hence, although the AIA establishes the foundations for the human centric and trustworthy AI, its adoption is merely the first step on the path towards achieving these objectives.

The preceding sections focused on the growing problem of algorithmic harms, which has been attributed to the issues of accountability paucity and AI opacity. The qualified transparency mechanism as a potential solution appears to be incorporated within the AIA, in line with the TSD. The upcoming three sections will be an attempt to look ahead and discuss the potential obstacles that may influence the effectiveness of the new transparency rules in serving AIA's aims.

5.1 The limited involvement of notified bodies

As stated earlier, while NBs are an integral part of the EU's harmonised product safety framework, their involvement under the AIA is mandatory only with regard to Annex I AI systems for which third-party oversight is already required. This would concern the conformity assessments conducted before the AI systems are deployed on the EU market and in the routine checks applicable in post-marketing monitoring. Such transparency measures on both stages could potentially reduce the instances when the MSAs would need to initiate ex-post investigations. However, since the AI systems listed in Annex III may undergo internal controls, the scope of qualified transparency applied to those systems as a "preventive approach" is likely to be marginal. This is despite that many of the high-risk AI systems listed in Annex III might arguably pose similar levels of risk as compared to the Annex I AI systems. For instance, under the Medical Devices Regulation (MDR)[153], the third-party conformity assessment for any recommendation- and decision-making software is mandatory. Thus, the question arises whether at least some AI systems in Annex III used in such areas as biometrics, critical infrastructure, social benefits, or law enforcement, could pose comparable risks.

On the current terms, however, the implementation of the AIA rules will mostly rely on their adherence by AI providers and deployers. This (over-)reliance on the AI operators' compliance and the ex-post approach by regulators has drawn substantial criticism from legal scholars[154]. Veale and Borgesius observe that "notified bodies play important roles beyond assurance, for example in translating rules, providing know-how to targets of regulation, and providing feedback to regulators and standard-setters".[155] Absence of such transparency intermediaries as NBs could result in "big gaps in knowledge flows" regarding how the AI Act framework is interpreted and applied by the AI operators[156]. Along the same lines, Ebers et al. have suggested that the EU Commission "should explore whether at least certain high-risk AI systems should be subject to an independent ex ante control"[157]. Even stricter approach has been presented by Malgieri and Pasquale, who argue that high-risk AI systems should be subject to third-party *licensure schemes*, as "there are some wrongs that can arise out of AI that are too serious to be recompensed ex-post"[158].

It could be speculated that the reason for the limited scope of the third-party conformity assessments could have been influenced by the concerns that such onerous regulatory requirements might jeopardise the AI innovation and uptake in the EU. The Commission explicitly states in the Explanatory Memorandum to the proposed AI Act that the "conformity assessment approach aims to *minimise the burden for economic operators* as well as *for notified bodies*, whose capacity needs to be progressively ramped up over time"[159]. Importantly, Recital 125 AIA indicates that the requirement of third-party conformity assessment for high-

[152] Ananny and Crawford (n 5).

[153] Regulation 2017/745 (n 125).

[154] Martin Ebers and others, 'The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)' (2021) 4 J; Gianclaudio Malgieri and Frank Pasquale, 'Licensing High-Risk Artificial Intelligence: Toward Ex Ante Justification for a Disruptive Technology' (2024) 52 Computer Law and Security Review; Andrew Tutt, 'An FDA for Algorithms' (2017) 69 Administrative Law Review 83.

[155] Veale and Borgesius (n 148) 16.

[156] Veale and Borgesius (n 148) 16.

[157] See, for example, Ebers and others (n 154).

[158] Malgieri and Pasquale (n 154) 3.

[159] European Commission, 'Explanatory Memorandum to the Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence ' (n 102) (emphasis added).

risk AI systems listed in Annex III has been intentionally left open for future consideration, thus the current status quo might change[160]. This is particularly noteworthy given that the Commission is authorised to adopt delegated acts in this area[161].

Still, the pre-marketing conformity assessments by NBs would have certain limitations. For example, the scrutiny by NBs would be limited to the model operation on specific datasets, which would therefore not provide a guarantee that the same model could "fail egregiously" in other data-contexts[162]. Moreover, some ML systems do not evolve after being deployed, but other systems continually change their internal structures, and/or are subject to decision processes by the system designers[163]. This feature makes the ex-ante conformity assessment impossible to "future-proof" the dynamic AI systems, although such challenges could be addressed to some extent by regular post-marketing surveillance procedures. Under the MDR, for instance, NBs should "periodically, at least once every 12 months, carry out appropriate audits and assessments", when applicable.[164] Thus, the range of the monitoring measures within the EU product safety regime vary depending on the safety risks involved in concrete contexts. It is therefore plausible that the EU regulators may adjust the AIA enforcement framework as well, if deemed necessary.

5.2 The MSAs enforcement duties - high expectations or high hopes?

Since the role of NBs will be limited under the AI Act (at least for the time being), the vast majority of the AIA enforcement duties on the national level will fall upon the MSAs. The scope of their responsibilities may pose a challenge, as the MSAs will need to oversee the AIA compliance concerning high-risk AI systems, the conformity with limited transparency requirements, the GPAI on the national territory, and monitor the market against the prohibited practices. The effectiveness of the national enforcement frameworks may be further complicated due to the dissemination of the oversight tasks between various MSAs, depending on the national arrangement. The Member States should therefore ensure the clear delineation of the oversight tasks between the MSAs, if that is the case, so that the enforcement is effective across the entire AIA framework.

Apart from the setting up comprehensive and well-coordinated national AIA enforcement frameworks, the Member States are tasked with ensuring the appropriate expertise and funding of AIA oversight bodies – both of the notifying authorities and the MSAs. However, the experiences with underfunded national authorities responsible for the enforcement of the GDPR present a worrying precedent[165], and Veale and Borgesius consider the budgetary estimation of the AIA enforcement by the Member States as "dangerously optimistic"[166]. Moreover, similar challenges may be encountered by the newly established AI Office. According to Zenner and Hacker, it faces a difficult mission, "an almost endless list of responsibilities, harsh deadlines, and a limited budget"[167]. The shortage of relevant expertise in the EU is a widespread problem affecting most European companies, as over 60% of EU enterprises report difficulties in filling such vacancies[168]. Ensuring the appropriate organisational and technical resources can therefore be an uphill task for the public sector as well.[169]

The differences in funding among the national MSAs may lead to disparities in the way the AIA is interpreted, and result in an uneven level of enforcement and protection across the EU. Moreover, depending on their

[160] As the Rec. 125 AIA states, "given the current experience of professional pre-market certifiers in the field of product safety and the different nature of risks involved, it is appropriate to limit, *at least in an initial phase of application of this Regulation,* the scope of application of third-party conformity assessment for high-risk AI systems other than those related to products" (emphasis added).

[161] Art. 43 (6) AIA.

[162] Margot E Kaminski, 'Regulating the Risks of AI' (2022) 103 Boston University Law Review 1347, 1350 accessed 16 September 2024.

[163] Kroll and others (n 52).

[164] Annex IX (3.3) MDR.

[165] See, for example, Ido Sivan-Sevilla, 'Varieties of Enforcement Strategies Post-GDPR: A Fuzzy-Set Qualitative Comparative Analysis (FsQCA) across Data Protection Authorities' [2022] Journal of European Public Policy; European Parliament, 'Resolution on the Commission Evaluation Report on the Implementation of the General Data Protection Regulation Two Years after Its Application'.

[166] Veale and Borgesius (n 155) 25.

[167] Zenner, Hacker and Hallensleben (n 137) 1.

[168] Eurostat, 'Enterprises That Recruited or Tried to Recruit ICT Specialists by Size Class of Enterprise' (2024).

[169] European Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' (n 31); Kemper and Kolkman (n 44).

socio-political objectives, some Member States may be strategically motivated to scale down the MSAs' activities, thereby facilitating the deployment of non-compliant AI systems.[170] Furthermore, insufficient resources of specific MSAs may slow the regulatory oversight, which would have a negative impact on "start-ups and companies driven by innovation", for whom "the speed at which they can enter the market is crucial"[171]. With regard to this obstacle, however, the possibility of consulting with the experts from the scientific panel could be seen as a useful option for the national authorities.

5.3 The remaining issue of the black-box AI

Although the AIA holds the potential to address the problem of legal opacity of AI, the issue of the technical opacity of AI – or the "black-box problem" – appears to remain opaque. It should be kept in mind that while the AIA contains certain transparency requirements vis-à-vis the high-risk AI and GPAI, the low-risk AI systems (that is, the majority of AI systems) are not subject to any minimum transparency standards. This could potentially pose a challenge for oversight bodies monitoring the market against the suspected use of prohibited practices, for instance concerning the manipulative AI systems affecting end-users' privacy and autonomy.

With respect to the AI deployed in contexts requiring full accountability, the use of black-box algorithms has been contested [172]. For example, many legal scholars consider the use of inscrutable AI models as not justified in the public sector, in particular when interpretable alternatives are available. Busuoic states that "where decisions have high-stakes (individual-level) implications, algorithms can neither be secret (proprietary) nor uninterpretable"[173]. Likewise, de Laat argues in the same context that algorithmic transparency should be ensured "either by making model outcomes understandable ex post, or by choosing models ex ante that are intelligible by design"[174]. As explained in Section 2.1, given the inherent transparency limitations of some highly complex AI, their application would necessitate the use of other algorithms (such as xAI) to examine their workings. Therefore, the question is whether these methods would provide a sufficient level of confidence to, effectively, entrust them with the task of "conformity assessment" under the AIA.

Following the EU's Ethics Guidelines for Trustworthy AI, it appears that the bottom-line of the conformity assessments is that AI systems should be interpretable, at least in the context of *trustworthy* AI[175]. This reasoning seems to be in accordance with the general rationale of the NLF, which requires all producers be able to demonstrate full compliance of the products with the relevant standards. Moreover, the requirement of explainability of automated decisions with significant consequences for individuals – unsuccessfully implemented under the GDPR[176] – has been reiterated in the AIA[177], thus further supporting the argument for application of interpretable models in such contexts. However, Ebers et al. state that the requirement to fully understand the capacities and limitations is "currently not feasible for some AI systems", which could lead to an "indirect ban of opaque high-risk AI systems"[178].

Notably, the AIA does not explicitly forbid the use of highly advanced AI, as long as the compliance with the AIA can be demonstrated, including the adherence to the relevant transparency requirements[179], ensuring

[170] Novelli and others (2024) point out that the AIA lacks the mechanism to revise the decisions of the MSAs on the EU level, such as in case of the European Data Protection Board under the GDPR. The authors argue that "without the ability to correct or harmonize national decisions, there's a heightened risk that AI could be misused in some Member States", including applications in sensitive areas like biometric surveillance. See: Claudio Novelli and others, 'A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities' [2024] SSRN Electronic Journal 27.

[171] Novelli and others (n 170) 23.

[172] de Laat (n 13); Busuioc (n 5); Tschider (n 9).

[173] Busuioc (n 5) 834.

[174] de Laat (n 13) 540.

[175] As reiterated in Rec. 27 AIA, stating that AI systems should be "developed and used in a way that allows appropriate traceability and explainability".

[176] On the contested "right to explanation" under the GDPR, see, for example, Wachter, Mittelstadt and Floridi (n 80).

[177] Art. 86 (1) AIA provides that the person affected "shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken."

[178] Ebers and others (n 154) 596.

[179] Art. 13 (1) AIA provides that the high-risk AI systems "shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately."

the effective human oversight[180], risk management systems, quality of datasets, and accuracy. At the same time, many of the high-risk AI provisions contain unsharp clauses. For example, the requirement to “properly understand” the AI capacities and limitations, and of AI being “sufficiently transparent”, may be difficult to interpret in concrete contexts. The harmonised standards, once adopted, may offer more clarity in this respect, as these have been requested to serve as technical benchmarks for AIA adherence[181]. Following the harmonised standards would provide the legal certainty for AI providers, as compliance with the standards entails the presumption of conformity with the AIA. Consequently, this could imply that the EU’s standard-setting bodies (CEN-CENELEC) will play the decisive role in defining what constitutes a black-box AI and what qualifies as sufficiently transparent AI[182]. Hence, it remains to be seen where the line between AI transparency and opacity will be drawn in the harmonised standards, which will directly affect the AIA compliance evaluations by both AI operators and oversight bodies alike.

6. Conclusions

With AI being widely adopted and algorithmic harms on the rise as a consequence, the need for systemic intervention has been seen as increasingly pressing. However, addressing the negative implications of AI would first necessitate confronting the issue of accountability paucity and the opacity surrounding AI technologies. In response, ideas such as qualified transparency have been proposed as potential solutions to balance the interests and rights of all AI stakeholders.

The article argues that the AI Act introduces the qualified transparency for certain AIA oversight bodies, primarily the MSAs, NBs and the AI Office in its market surveillance capacity. These oversight bodies have been granted – under the duty of confidentiality – access to the information necessary to scrutinise AI systems’ compliance with the AIA. Moreover, by defining the information required for regulatory scrutiny, the AIA may effectively limit the scope of the trade secrecy claims, often asserted by AI companies. However, the effectiveness of the AIA transparency rules depends on the extent they will be operationalised by the oversight bodies. In preparation for the upcoming enforcement responsibilities, the significant task ahead for the AIA oversight bodies is to build up their expertise and capacity for the broad range of enforcement tasks. Some of the oversight challenges may still arise from the technical complexity inherent in the most advanced AI systems. Although the transparency requirements for high-risk AI and GPAI have partially addressed the black-box problem, the interpretation of these provisions will be significantly shaped by the forthcoming harmonised standards.

Emerging from the above discussion is the notion that the effectiveness of the qualified transparency and AIA framework as a whole in reaching its objectives still depends on many factors. The AIA governance mechanisms may require future adjustments, such as the introduction of the conformity assessment by NBs for certain high-risk AI categories listed in Annex III. Thus, it is yet to be determined whether the AIA regulatory framework is sufficiently well-balanced to safeguard the policy objectives of the human-centred and trustworthy AI, as well as whether it is flexible enough to keep pace with the AI progress.

[180] Art. 14 (4)(a) AIA states that ”the natural persons to whom human oversight is assigned are enabled (...) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance”.

[181] See, for example, Carlo Tovo, ‘Judicial Review of Harmonized Standards: Changing the Paradigms of Legality and Legitimacy of Private Rulemaking under EU Law’ (2018) 55 Common Market Law Review 1187; Rossana Ducato, ‘Why Harmonised Standards Should Be Open’ (2023) 54 IIC International Review of Intellectual Property and Competition Law 1173.

[182] Notably, the delegation of such significant regulatory powers to the private standardisation bodies has been broadly questioned, see e.g. Federica Paolucci, ‘Shortcomings of the AI Act - Evaluating the New Standards to Ensure the Effective Protection of Fundamental Rights’ [2024] Verfassungsblog accessed 22 July 2024, Ducato (n 181); Tovo (n 181);