International Journal of Law and Politics Studies
ISSN: 2709-0914
DOI: 10.32996/ijlps
Journal Homepage: www.al-kindipublisher.com/index.php/ijlps

IJLPS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Bridging the Proof Gap: Reconstructing Algorithmic Impact Assessment for Anti-Discrimination Litigation

Yishuo Guo
East China University of Science and Technology, Shanghai, China
Corresponding Author: Yishuo Guo, **E-mail**: yishuoguo@126.com

| ABSTRACT

Algorithmic Impact Assessment (AIA), as a core tool of algorithmic governance, currently serves mainly administrative regulation and enterprise compliance, and fails to effectively respond to the evidentiary needs of individual victims in anti-discrimination litigation. When a plaintiff claims that an algorithmic decision constitutes discrimination, it is usually necessary to overcome three barriers: blurred subject of responsibility, broken causal relationship and difficulty in obtaining evidence. This paper takes the Mobley v. Workday case as an entry point, analyzes the causes of the above three dilemmas, and proposes four institutional designs for the systematic reconstruction of Algorithmic Impact Assessment. Firstly, human-machine comparison benchmark is introduced to establish the admissibility standard for the algorithmic conclusions in the AIA report; Secondly, the adversarial assessment mechanism shall be established so that group statistical evidence can be used for individual causal proof; Thirdly, a technology-legal sandbox shall be constructed to provide a standardized platform for the generation and review of AIA reports; Fourthly, expand the functional boundaries of the Algorithmic Liability Insurance, link the insurance mechanism with the risk assessment conclusion of the AIA report, and open up the final relief channel from evidence to compensation. The four systems jointly form a closed loop system of standard setting, evidence generation, review accessibility, and practical relief, which promote the shift from prevention without relief to simultaneous prevention and relief, so that the rules of proof in anti-discrimination litigation can be dynamically adapted to the continuous evolution of AI technology.

| KEYWORDS

Algorithmic Impact Assessment, Algorithmic Discrimination, AI Accountability, Anti-Discrimination Litigation, Algorithmic Governance

| ARTICLE INFORMATION

ACCEPTED: 10 June 2026 **PUBLISHED:** 21 July 2026 **DOI:** 10.32996/ijlps.2026.8.6.5

1. Introduction

With the rapid development of artificial intelligence technology, algorithms have infiltrated all aspects of our lives. From recommending jobs and reviewing loans to assessing risks and screening resumes, their presence is everywhere. It can be said that algorithms are taking over more and more decisions that were once made by humans. But behind this efficiency, a troubling question quietly emerges: What can we do when the black box of algorithms makes unfair and even discriminatory decisions?

The Mobley v. Workday case [1] of 2025 has brought this issue into the spotlight. Five job seekers over the age of 40 applied for hundreds of positions through Workday, only to have their applications consistently rejected without a single interview. They sued Workday, alleging that the AI-powered recommendation system used in its hiring process systematically discriminates against workers over 40. Although the court allowed them to proceed as a class action, a deeper dilemma remains unresolved: Even if we can observe discrimination at the group level, how can an individual prove that the algorithm intentionally rejected them?

This case echoes French philosopher Jacques Ellul's profound insight into technology: algorithms have evolved into autonomous forces beyond human control, even as their value is still claimed to depend entirely on how humans use them [2]. This tension reveals the fragility of the principle of neutrality and underscores the urgent need to bring algorithms under legal scrutiny. At the same time, the case highlights three core dilemmas in algorithmic discrimination litigation: First, who is liable when decision-making power is outsourced to an algorithm provider? Second, how can causality be established—how do we move from group-level correlation to individual-level causation? Third, how can plaintiffs access evidence when algorithmic operations are shielded by trade secrets?

For a long time, the regulation of algorithms in the United States has relied on the path of administrative supervision. For example, the Algorithmic Accountability Law, which has been proposed many times but has not yet been implemented, requires enterprises to conduct Algorithmic Impact Assessment (AIA). Algorithmic Impact Assessment is widely defined as an accountability mechanism and governance tool aimed at anticipating, avoiding and mitigating the negative impact of the algorithmic decision system. Through a systematic documentation and evaluation process, the AIA examines its potential risks at key nodes in the development or deployment of algorithmic systems to provide means to mitigate possible hazards to vulnerable groups [3]. But enforcement is in the hands of the Federal Trade Commission (FTC), and individual victims have no right to sue at all. Anti-discrimination litigation, which can provide relief for individuals in theory, is also almost impossible because of the above-mentioned triple dilemma. How to break through these three predicaments and achieve relief for individuals? The answer, in this paper, lies not in starting afresh, but in the re-engineering of the AIA system itself —transforming it from a static compliance tool serving administrative regulation to a dynamic evidence system that is deeply integrated with anti-discrimination litigation.

The remainder of this paper proceeds as follows: The second part, based on summarizing the existing research, thoroughly analyses the causes of the triple dilemma; The third part focuses on the distribution of responsibilities, the proof of causal relationship, the disclosure of evidence and other dimensions, and launches four institutional designs for the systematic reconstruction of AIA; The fourth part summarizes the conclusion and looks forward to the future.

2. Literature Review

2.1. Governance Paths of Algorithmic Accountability and the Difficulties of Identifying Responsible Entities

The existing research on algorithmic governance mainly follows two paths. The public law approach takes administrative supervision as the core and emphasizes the prevention of systemic risks through institutional design. Early studies advocated the principle of transparency and attempted to achieve risk prevention by disclosing the logic of the algorithms [4.5]. But it is worth asking whether the principle of transparency can really prevent risks? Based on criticizing the transparency fallacy, Ananny & Crawford called for the establishment of alternative accountability mechanisms [6]. Scholars have further designed tools such as Algorithmic Impact Assessments and Data Protection Impact Assessments [7], and systematically put forward principles such as fairness, transparency and interpretability covering the whole process [8.9], trying to intervene in risks before damage occurs.

However, Busuioc points out that traditional administrative regulation presupposes that the regulator can see the logic of the algorithms and trace decision-making responsibilities, but the technical black box and continuous evolution of AI systems precisely undermine this presupposition [10]. When the government is unable to effectively supervise, it is even more difficult for individuals to obtain relief through administrative channels.

In response to this dilemma, Katyal proposes to move to internal accountability mechanisms in the private sector, such as codes of conduct and whistleblower protection, in the absence of government regulation [11]. Rahman et al. expanded the perspective of private law accountability from bilateral relations to multilateral ecosystems, and their framework of multilateral platforms requires multilateral accountability systems that emphasize the sharing of responsibilities among platforms, users, producers, and regulators [12]. But the problem is that no matter how much the scope of accountability is expanded, there is always a core issue that remains unresolved: who is responsible for the discriminatory decisions made by the algorithms?

In the Mobley v. Workday case, Workday argued that it was only a neutral tool provider and tried to shift the responsibility to the employer. There is a typical principal-agent problem between algorithm developers and deployers: developers can use technical complexity to advocate model uncertainty to evade liability [13], and enterprises weigh their own behavior in a vague regulatory environment [14]. This evasion of responsibility is not new. In the Uber case, the platform also advocated that the output of the algorithm is the automatic result of the preset code, rather than the discretionary behavior of the manager, to cut off the connection between the algorithm's instructions and the responsibilities of the employer [15].

The existing AIA system is mainly built on the logic of administrative supervision, and the system design itself does not provide

direct answers to the applicability problems that may arise in the determination of private law liability. Take the U.S. Algorithmic Accountability Act of 2022, for example, which delimits covered entities by setting financial thresholds and user sizes, requiring them to submit evaluation reports to the Federal Trade Commission, which shall be responsible for its implementation [16]. Therefore, it is impossible to respond to the fundamental question of whether an algorithm service provider constitutes an employer or a decision participant in private law litigation. This means that in order for AIA to truly serve anti-discrimination litigation, it is first necessary to redesign the identification mechanism of its responsible entities.

2.2. Causality and the Difficulty of Individual Proof

However, once the responsible entity has been identified, another fundamental question arises: How can individuals prove that they have been harmed by the algorithms? The review of Kordzadeh and Ghasemaghaei makes this point clear: the concept of algorithmic bias has been fully discussed in the academic community, but there is very limited empirical research on how individuals can prove that they are harmed by algorithms [17]. In other words, accountability requires evidence, and where does the evidence come from?

As the mainstream tool for identifying and mitigating algorithmic bias, the core logic of the Algorithmic Impact Assessment is based on the measurability of group statistical indicators. A large amount of research has been devoted to detecting the differential impact of algorithms on different groups through fairness indicators [18], and a series of mitigation strategies, including removing input variables and replacing variables, have been put forward [19].

However, this type of evidence supply based on group statistics faces fundamental limitations. Mayson has profoundly pointed out that the problem lies not in the algorithms themselves, but in prediction. Any method of predicting the future based on past data will project structural inequality in history into the future [20]. The group statistical indicators provided by the AIA are essentially quantitative descriptions of existing social inequality in the algorithm, but they cannot tell us to what extent the algorithm's predictions are determined by an individual's true characteristics, and to what extent they are the products of social structural biases. This is exactly the key fault line in moving from group statistical indicators to individual causal proof in anti-discrimination litigation.

Why can the existing AIA provide group statistical indicators, but find it difficult to support individual causal proof? The answer lies in the deep mismatch between technology supply and proof demand. At the group level, the existing technology has developed more mature statistical methods for group indicators [21], which provide tool support for common evidence in collective litigation. The court in the Workday case further clarified that the core issue of algorithmic discrimination - that is, whether the AI system has a differential impact on protected groups - is a common issue that can be proved by common evidence, thus opening the way for a class action.

But proving causation at the individual level remains problematic. The counterfactual explanations proposed by Wachter et al. can bypass the technical and legal obstacles to opening the black box [22], but the confidence interval of the explanations generated by the current technology is too wide to meet the requirements of the Daubert standard [23]. The study of Salih et al. has revealed that the results of mainstream explanation methods are highly dependent on model selection and vulnerable to feature collinearity interference, resulting in extremely unstable explanations [24]. Huang and Marques-Silva further fundamentally demonstrated that interpretations based on Shapley values necessarily have the possibility of misleading information [25]. The significance of this proof in evidentiary law is quite far-reaching: if a certain interpretation method has systematic defects at the theoretical level, the evidence it produces will naturally be difficult to meet the requirements of the Daubert standard.

However, the technical literature has mostly focused on the improvement of the accuracy of the explanation methods, but has not asked the key legal question: What degree of inaccuracy is acceptable under the Daubert standard? This question points to the necessity of a human-machine comparison benchmark. As emphasized by Ross et al in their work on medical evidence, the error rate of the algorithm's interpretation needs to be compared with the error rate of the judgment of a human expert before a reasonable threshold of its legal admissibility can be established [26]. Therefore, only by comparing the accuracy of the algorithm's interpretation with the accuracy of the judgment of human experts can the technical bottleneck be transformed into a measurable evidentiary standard, thus closing the gap between legal rigor and technical feasibility.

2.3. Barriers to Evidence Acquisition and Technical Admissibility

Even if individual causal evidence is technically possible, the plaintiff still faces two barriers that are difficult to cross: first, how to obtain the algorithmic information protected by trade secrets; Second, how can the information obtained be effectively examined in court?

For the first dilemma, scholars are generally concerned about the conflict between the protection of trade secrets and the right to know of individuals. Foss-Solbrekk challenges the idea of absolute exemption of algorithmic information and advocates a disclosure path based on the public interest [27]. Lu pointedly noted that, even with the relevant provisions of the GDPR, the technical complexity of the proprietary algorithmic system, the legal claim of trade secrets and the opacity of management jointly build a black box that is inaccessible [28]. de Laat further warned that the increasingly strengthening trend of intellectual property protection in the field of machine learning would make the trade secret claims of enterprises weightier, thus seriously weakening the rights and interests of individuals in practice [29].

For the second dilemma, that is, how to achieve effective review in court after obtaining information, the study goes deep into the level of procedural mechanisms and evidentiary standards. Grozdanovski has analyzed the EU AI liability legislative proposal, confronted the contradiction between evidence disclosure and trade secret protection in litigation, and explored how to ensure the effective participation of the parties procedurally [30]. Bloch-Wehba and Bunnell have proposed solutions at the institutional level. The former advocates the activation of the Freedom of Information Act to obtain information on public sector algorithms, and the latter suggests the establishment of independent institutions for professional supervision [31.32]. However, in the face of the black box of algorithms, even if a court-appointed expert system, such as Rule 706, is introduced, it is faced with the dilemma of having the system but lacking the resources due to the extreme shortage of experts and high audit costs [33.34].

Faced with these difficulties, the legal profession has begun to explore ways to respond, but it is equally difficult. Ahern's systematic exposition of forward-looking governance tools such as regulatory sandboxes and policy laboratories provides a blueprint for bridging the mismatch between legislative and technological iterations, but these frameworks are currently more at the level of macro policy experiments and fail to address the procedural details of individual litigation relief [35]. The reflection of Bernstein et al. on bureaucratic accountability directly challenges the simplified model of principal-agent theory, pointing out that accountability needs to be embedded in the whole process of algorithm design, deployment, and operation, rather than relying only on ex post punishment [36]. Although these explorations have pointed out the problems acutely, they mostly remain at the level of institutional concepts and have not yet formed a set of integrated practical schemes that can respond to the dual constraints of strategic behavior and engineering cost at the same time, which is also the direction that this article attempts to promote.

2.4. Summary

To sum up, the existing research has revealed the problems in algorithmic discrimination relief from various dimensions: evidence is required for accountability in private law, and the confidence interval of counterfactual explanations is too wide; The strategic behavior of enterprises further aggravates the difficulty of proof; Hard constraints at the engineering level make it difficult to scale up any scheme that relies on in-depth review of individual cases. Then, how can we design a workable Algorithmic Impact Assessment system so that it can provide evidence for anti-discrimination litigation and thus systematically respond to these three dilemmas? The following four institutional designs are just preliminary answers to this question.

3. Institutional Design: An Engineering-Coordinated Solution to the Triple Dilemma

The deep logic of the four institutional designs put forward in this paper points to a fundamental value choice: the difficulty in presenting evidence in algorithmic discrimination litigation is not an isolated phenomenon, but a universal and typological dilemma. In real litigation, the specific circumstances are different, and strictly abiding by a set of pre-determined rules to distribute risks and burdens among the parties will inevitably bring about injustice in individual cases, resulting in unfair problems such as the conflict between the adjudicative outcome and substantive justice. There is a huge disparity in evidentiary capacity between algorithm service providers and individual plaintiffs. If we still abide by the general rules of burden of proof and require the victim to bear the adverse consequences of failing to provide evidence, it is contrary to substantive justice and cannot achieve the normative purpose. Therefore, it is necessary to consider the alleviation of the burden of proof for the typological difficulties commonly existing in algorithmic discrimination litigation cases [37]. The risk of algorithmic discrimination is a product of society, which cannot be borne entirely by users or service providers but should be allocated in layers through the dynamic transfer of the burden of proof. The following four institutional designs are precisely an institutionalized embodiment of this value judgment.

3.1. Human-Machine Comparison Benchmark

As mentioned earlier, the greatest difficulty in the current algorithmic discrimination litigation is the ambiguity of the responsible entity. In the Workday case, the defendant's defence as a neutral tool provider took advantage of the lack of criteria for determining liability [1]. This article advocates that the clauses on decision-making participation in service agreements and development documents be listed as mandatory disclosure items in the Algorithmic Impact Assessment. If the results of algorithmic recommendations play a decisive role in decision-making (such as scoring and ranking), the service provider constitutes a decision-making participant that can be included in the anti-discrimination lawsuit. This determination standard extends the responsibility from the algorithm deployer to the service provider, which just responds to the dilemma of evading responsibility caused by the principal-agent problem.

After the responsible entity is clearly defined, the problem of proving the causal relationship immediately follows. The experiment of Senoner et al. shows that in visual inspection tasks, the provision of interpretable AI assistance with heat maps can significantly improve the decision-making accuracy of human experts, and its effect is better than that of pure black boxes [38]. This finding reveals a key fact: the utility of AI interpretation does not depend on its absolute precision, but on whether it can help humans make more reliable decisions than independent judgment. Along this line of thinking, this paper proposes the human-machine comparison benchmark as the relative standard for evidentiary admissibility: by comparing the width of the confidence interval of counterfactual explanations with the range of judgment errors of human experts in similar tasks, if the reliability of the AI's interpretation is not inferior to the standardized human judgment benchmark, it can be preliminarily determined that it possesses evidentiary admissibility.

This standard can be developed in practice along two paths. The preferred path is to build a human decision-making surrogate model based on large-scale historical data, with its judgment error range as the reference benchmark [39], which must disclose its training data and limitations to be subject to cross-examination. In areas where there is insufficient data, human performance in the test can be benchmarked with the use of certified industry-standardized test banks. On this basis, this paper further proposes the concept of risk calibration coefficient to build a bridge between technical feasibility and legal rigor. The core logic of the concept is that the error rate of the algorithm's interpretation does not have to be lower than the theoretical optimal value of the human expert, if it does not exceed the range of the human benchmark error multiplied by the specific risk coefficient. The right to set risk coefficients can be handed over to institutions such as NIST or FTC to publish standardized coefficient tables through annual rulemaking procedures. For example, in high-risk scenarios involving basic rights such as recruitment and credit, a strict coefficient of 0.8 times is applied; In low-risk scenarios such as ordinary product recommendation, the more lenient coefficient of 1.2 times can be applied.

This benchmark, once established, should be an integral part of the AIA report. Any service provider who claims that there is no discrimination in its algorithmic decision shall provide counterfactual explanations generated by certified tools in the AIA report and prove that their confidence interval falls within the allowable range of the human-machine comparison benchmark. An AIA report that fails to meet this requirement will not have presumptive evidentiary weight in the proceedings.

3.2. Adversarial Assessment

Although the human-machine comparison benchmark solves the problem of identifying the responsible entity and the admissibility of evidence, the plaintiff still faces a key obstacle in algorithmic discrimination litigation: how to prove systemic discrimination exists. The adversarial assessment system introduces red team testing from cybersecurity into algorithm governance, with the aim of transforming latent discrimination patterns into actionable systematic evidence. Red team testing systematically detects model vulnerabilities by simulating an attacker's perspective and has been widely used in the field of large language model security assessment [40.41]. However, existing practices focus more on vulnerabilities at the model level, paying insufficient attention to the socio-technical system in which the algorithm is embedded, and easily devolve into a security theater: that is, testing is only a formality, and its function is more to create a sense of security than to substantively uncover risks [42], For this reason, this system draws on the two-tiered red team concept proposed by Majumdar et al. and requires mandatory red team testing for high-risk systems [43]. Here, a high-risk system can refer to the definition standard of the Algorithmic Accountability Act, that is, affecting more than 1 million people, or involving fundamental rights such as employment, credit, and housing. This test aims to systematically detect whether there are reproducible discriminatory patterns within the model across a specific input range, to link group-level statistical evidence with individual harm and address the problem of individual causal proof.

At the practical level, the system draws on the idea of Casper et al. to verify defects by implanting backdoors [44], using the stable discriminatory patterns against specific groups found in the test as objective evidence that enterprises should know that their systems are at risk before deployment. The plaintiff only needs to prove that they belong to the vulnerable group identified by the red team test, which can preliminarily establish a causal link, thus transferring the burden of proof to the defendant. To balance trade secret protection, the design also establishes a hierarchical disclosure mechanism: filing only complete technical details with the regulatory authority and disclosing to litigants only the feature distribution of the failure modes (for example, the false positive rate for female users aged 30-35 increases by 20%). This not only guarantees the litigation rights of the parties involved but also keeps the core algorithmic logic of the enterprise within the scope of trade secret protection. The conclusions of the adversarial assessment should be a core component of the AIA report. Service providers are required to disclose in the AIA report all reproducible discriminatory patterns uncovered by the red team test and their distributed impact on protected groups. Failure to disclose or inaccurate disclosure may constitute an adverse presumption in subsequent litigation.

3.3. Technology-Legal Sandbox

To truly enter legal proceedings and fulfil an evidentiary function, the evidence produced by the adversarial assessment must cross a key threshold: How can the court examine such evidence efficiently and cost-effectively? The scarcity of expert resources, the protection of trade secrets and the high audit costs make it difficult to sustain the mode of relying on in-depth, case-by-case review [33.34]. An investigation of the AI ethics audit ecosystem by Schiff et al. has revealed the root cause of this dilemma: the current audit is highly dependent on the personal ability of a small number of experts, and lacks standardized and reproducible tools to support it [45]. The criticism of Adams and others on the reliability of expert testimony further shows that even if experts are introduced, if the methodology itself is not standardized and verified, the focus of controversy will only shift from the algorithmic black box to the expert black box [46].

To resolve this structural tension, this paper proposes a technology-legal sandbox system, which aims to establish a review mode of tool-first, expert-supplemental by scaling up expert judgment capabilities through automated tools. The core of the system is to translate the legal requirements for the reliability of evidence into operational technical standards and to achieve dynamic updates of tools with the help of a controlled environment for regulatory sandboxes [47].

Specifically, this article proposes that NIST, drawing on the assurance audit framework of Lam et al. [48], establish benchmark test suites and conduct hierarchical certification of audit tools: Tier 1 basic certification can meet the needs of routine screening, Tier 2 enhanced certification is used for high-risk scenarios, and Tier 3 experimental tools are provided for judicial review reference. The court will only appoint experts when the case involves specific defects that cannot be covered by Tier 2 or above tools and there are no lower-cost alternatives. To ensure that standards keep pace with the times, the law may require conforming to the technical specifications issued by national accreditation bodies as the meta-standard, and the specific indicators shall be dynamically updated by NIST through the annual rulemaking procedure and shall enter the scope of judicial notice after the public comment period. This kind of citation update mechanism not only avoids the legislative lag inherent in frequent amendments, but also entrusts the professional judgment of technical standards to specialized agencies. The Technology-Legal Sandbox provides institutionalized tool support for the generation and review of AIA reports. An AIA report generated by a sandbox-certified tool may be presumed to be prima facie evidence in the proceedings; Reports generated by uncertified tools are subject to additional expert review before they can enter the proceedings.

3.4. Algorithmic Liability Insurance

Even if the problems of responsibility determination, causal relationship, and evidence admissibility are solved through the system design above, individual victims may still be unable to obtain actual compensation due to the defendant's insufficient solvency or prohibitively high litigation costs. The risk of Algorithmic Discrimination is the product of the co-evolution of socio-technical systems, and its cost should not be borne entirely by individuals, nor can it be absorbed only by service providers. Therefore, this paper proposes the Algorithmic Liability Insurance" system, which expands the insurance mechanism from the traditional ex-post compensation to a whole-process risk management tool of pre-monitoring-in-process intervention, and post-relief. The insurance covers endogenous technical risks such as algorithmic black boxes and discrimination [49], as well as exogenous cybersecurity risks such as data leakage and system intrusion [50]. The dual structure of parallel product liability insurance and cybersecurity insurance can correspond to different types of damages [51].

In order to realize the precise underwriting of liability and the dynamic adjustment of premiums, this paper draws on the idea of vehicle black box data storage, mandates that high-risk algorithm systems install embedded monitoring probes to record human-computer interaction processes and key decision-making instructions in real time, and traces and reconstruct accident scenarios through multimodal data fusion technology [52]. The premium adjustment mechanism follows the dynamic principle: assessments

of the risks associated with algorithmic operations are conducted quarterly. When the risk level rises, the premium shall be adjusted from the following month, with a single increase of not more than 20%, and a 30-day investigation period shall be set for in-depth repair [53]. The core logic of this dynamic adjustment mechanism is to embed economic incentives into the daily operation of enterprises. If an enterprise wants to continue to enjoy a lower premium, it must internalize compliance as a routine part of algorithmic iteration, rather than passively responding to litigation. In addition, an Algorithmic Risk Reinsurance Fund shall be established. When the total annual claims of the industry exceed the set threshold, the fund shall intervene to trigger compensation to smooth systemic risks [54]. Algorithmic liability insurance and AIA form a closed-loop system: the risk assessment conclusion in the AIA report can be used as the basis for premium pricing; The monitoring data of the insurance probe can be fed back to inform AIA's iterative updates. Both jointly ensure that a favourable judgment in an anti-discrimination lawsuit is not merely a paper victory but can be converted into actual compensation for the victim.

3.5. Summary: The Collaborative Logic of the Four Systems

To sum up, the four institutional designs proposed in this paper jointly constitute a systematic reconstruction of Algorithmic Impact Assessment. The human-machine comparison benchmark compares AI explanations with human decision-making benchmarks, establishes a relative standard for the admissibility of evidence in anti-discrimination litigation, responds to the difficulties caused by the excessively wide confidence intervals of counterfactual explanations at their source, and qualifies the AIA report to serve as evidence. Adversarial assessment introduces red team testing into algorithm governance, which on the one hand mandates the identification of systemic defects, and on the other hand connects the plaintiff with vulnerable groups, thus transforming group statistical evidence into individual causal proof at the institutional level, facilitating the dynamic transfer of the burden of proof, and thus establishing the obligation to disclose systemic defects that must be fulfilled by the AIA report. The technology-legal sandbox effectively solves the dilemma of scarce expert resources and high costs in the evidence review process by scaling up expert judgment capabilities through automated means via hierarchical certification and citation updates of tools and provides platform support for the efficient generation and review of AIA reports. Algorithmic liability insurance expands the insurance mechanism from traditional ex-post compensation to a governance tool that operates through the whole process of pre-monitoring, in-process intervention, and post-relief. Through embedded probes, dynamic premium adjustment, and reinsurance funds, it establishes the ultimate relief pathway from evidence to compensation for the prevailing plaintiff, becoming the key guaranteed mechanism for the implementation and effectiveness of the AIA system.

These four institutional designs are intertwined, linking standard setting, evidence generation and accessible review with actual relief, and forming a closed loop system throughout the whole process of Algorithmic Impact Assessment. It is through this closed loop that Algorithmic Impact Assessment is transformed from a static administrative compliance tool into a dynamic evidence system deeply integrated with anti-discrimination litigation.

4. Conclusion

Algorithmic discrimination litigation faces three dilemmas: a blurred responsible entity, a broken causal relationship, and difficulty in obtaining evidence. This is not an isolated problem that occasionally arises in the process of technological development, but a structural contradiction that will inevitably be triggered after artificial intelligence is embedded in social decision-making. To address this problem, this paper proposes a systematic reconstruction of Algorithmic Impact Assessment, transforming it from an administrative compliance mechanism into a dynamic evidentiary system integrated with anti-discrimination litigation.

The reconstruction relies on four institutional mechanisms. Human–machine comparison benchmarks establish admissibility standards for algorithmic explanations by comparing confidence intervals against calibrated human error baselines. Adversarial assessment (red-team testing) converts group-level statistical bias into presumptive evidence for individual causation by anchoring plaintiffs to identified vulnerable populations. Technology–legal sandboxes standardize the generation and judicial review of AIA reports through certified tools, making expert capacity scalable. Algorithmic liability insurance connects evidentiary findings with effective compensation through embedded monitoring probes and dynamic premium adjustments.

Together, these mechanisms produce three direct effects: first, through mandatory disclosure requirements, they transform the identification of responsible entities from a vague "neutral tool provider" defence into a verifiable fact of "decision participant"; second, when algorithmic explanations meet human–machine comparison benchmarks and the plaintiff belongs to a vulnerable population identified through adversarial assessment, a causal presumption is triggered, shifting the burden of proof onto the

defendant; and third, they reduce evidentiary barriers through layered disclosure and certified tools, making AIA reports accessible, admissible, and contestable in court, while algorithmic liability insurance uses embedded monitoring probes to record fairness metrics in real time, establishing the specific point at which the enterprise became aware of the risk.

This framework repositions AIA as a bridge between ex-ante governance and ex-post legal relief—an institutional hub integrating technological auditing, evidentiary standards, and market-based risk allocation. Through this transformation, algorithmic discrimination claims can evolve from technologically intractable individual disputes into standardized and manageable legal processes, enabling algorithmic systems to move from opacity to accountability**.**

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] *Mobley v. Workday, Inc., No. 23-cv-00770-RFL, 2025 WL 1611791 (N.D. Cal. May 16, 2025).*
[2] *Ellul, J. (1964). The technological society. Vintage Books.*
[3] *Moss, E., Watkins, E. A., Metcalf, J., & Elish, M. C. (2020). Governing with algorithmic impact assessments: Six observations. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3584818*
[4] *Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. Digital Journalism, 3(3), 398–415. https://doi.org/10.1080/21670811.2014.976411*
[5] *Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. Digital Journalism, 5(7), 809–828. https://doi.org/10.1080/21670811.2016.1208053*
[6] *Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645*
[7] *Cartolovni, A., Tomicic, A., & Mosler, E. L. (2022). Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. International Journal of Medical Informatics, 161, Article 104738. https://doi.org/10.1016/j.ijmedinf.2022.104738*
[8] *Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., Colón-Rodríguez, C. J., Dullabh, P., Duran, D. G., Fair, M., Hernandez-Boussard, T., Hightower, M., Jain, A., Jordan, W. B., Konya, S., Moore, R. H., Moore, T. T., Rodriguez, R., Shaheen, G., Snyder, L. P., Srinivasan, M., Umscheid, C. A., & Ohno-Machado, L. (2023). Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Network Open, 6(12), Article e2345050. https://doi.org/10.1001/jamanetworkopen.2023.45050*
[9] *Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2023). Trustworthy artificial intelligence: A review. ACM Computing Surveys, 55(2), Article 39. https://doi.org/10.1145/3491209*
[10] *Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. Public Administration Review, 81(5), 825–836. https://doi.org/10.1111/puar.13293*
[11] *Katyal, S. K. (2019). Private accountability in the age of artificial intelligence. UCLA Law Review, 66(1), 54–141.*
[12] *Rahman, H. A., Karunakaran, A., & Cameron, L. D. (2024). Taming platform power: Taking accountability into account in the management of platforms. Academy of Management Annals, 18(1), 251–294. https://doi.org/10.5465/annals.2022.0090*
[13] *Kemp, A. (2024). Competitive advantage through artificial intelligence: Toward a theory of situated AI. Academy of Management Review, 49(3), 618–635. https://doi.org/10.5465/amr.2022.0045*
[14] *Zhang, L., Shao, Z., Chen, B., & Benitez, J. (2025). Unraveling generative AI adoption in enterprise digital platforms: The effect of institutional pressures and the moderating role of internal and external environments. IEEE Transactions on Engineering Management, 72, 335–348. https://doi.org/10.1109/TEM.2024.3512345*
[15] *Tomassetti, J. (2021). Algorithmic management, employment, and the self in gig work. In D. D. Acevedo (Ed.), Beyond the algorithm: Qualitative insights for gig work regulation (pp. 123–145). Cambridge University Press.*
[16] *U.S. House of Representatives. (2022). Algorithmic Accountability Act of 2022 (H.R. 6580, 117th Congress). https://www.congress.gov/bill/117th-congress/house-bill/6580/text*
[17] *Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212*
[18] *Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024). Unmasking bias in artificial intelligence: A systematic review of bias detection and mitigation strategies in electronic health record-based models. Journal of the American Medical Informatics Association, 31(5), 1172–1183. https://doi.org/10.1093/jamia/ocae023*

[19] *Siddique, S. M., Tipton, K., Leas, B., Jepson, C., Aysola, J., Cohen, J. B., Flores, E., Harhay, M. O., Schmidt, H., Weissman, G. E., Fricke, J., Treadwell, J. R., & Mull, N. K. (2024). The impact of health care algorithms on racial and ethnic disparities: A systematic review. Annals of Internal Medicine, 177(4), 484–496. https://doi.org/10.7326/M23-2960*

[20] *Mayson, S. G. (2019). Bias in, bias out. Yale Law Journal, 128(8), 2218–2300.*

[21] *Zliobaite, I. (2017). Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31(4), 1060–1089. https://doi.org/10.1007/s10618-017-0506-1*

[22] *Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.*

[23] *Guidotti, R. (2024). Counterfactual explanations and how to find them: Literature review and benchmarking. Data Mining and Knowledge Discovery, 38(5), 2770–2824. https://doi.org/10.1007/s10618-022-00831-6*

[24] *Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1), Article 2400304. https://doi.org/10.1002/aisy.202400304*

[25] *Huang, X., & Marques-Silva, J. (2024). On the failings of Shapley values for explainability. International Journal of Approximate Reasoning, 171, Article 109112. https://doi.org/10.1016/j.ijar.2023.109112*

[26] *Ross, D. E., Seabaugh, J., Seabaugh, J. M., Barcelona, J., Seabaugh, D., Wright, K., Norwind, L., King, Z., Graham, T. J., Baker, J., & Lewis, T. (2022). Updated review of the evidence supporting the medical and legal use of NeuroQuant® and NeuroGage® in patients with traumatic brain injury. Frontiers in Human Neuroscience, 16, Article 715807. https://doi.org/10.3389/fnhum.2022.715807*

[27] *Foss-Solbrekk, K. (2023). Searchlights across the black box: Trade secrecy versus access to information. Computer Law & Security Review, 50, 105811. https://doi.org/10.1016/j.clsr.2023.105811*

[28] *Lu, S. (2022). Data privacy, human rights, and algorithmic opacity. California Law Review, 110(6), 2087–2147. https://doi.org/10.15779/Z38804XM07*

[29] *de Laat, P. B. (2022). Algorithmic decision-making employing profiling: Will trade secrecy protection render the right to explanation toothless? Ethics and Information Technology, 24(2), 17. https://doi.org/10.1007/s10676-022-09642-1*

[30] *Grozdanovski, L. (2025). My AI, my code, my secret - Trade secrecy, informational transparency and meaningful litigant participation under the European Union's AI Liability Directive Proposal. Computer Law & Security Review, 56, 106117. https://doi.org/10.1016/j.clsr.2025.106117*

[31] *Bloch-Wehba, H. (2020). Access to algorithms. Fordham Law Review, 88(4), 1265–1314.*

[32] *Bunnell, N. (2021). Remedying public-sector algorithmic harms: The case for local and state regulation via independent agency. Columbia Journal of Law and Social Problems, 54(2), 261–303.*

[33] *Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 1–15. https://doi.org/10.1177/2053951720927946*

[34] *Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., & Ookden-Rayner, L. (2022). The medical algorithmic audit. The Lancet Digital Health, 4(5), e384–e397. https://doi.org/10.1016/S2589-7500(22)00024-5*

[35] *Ahern, D. (2025). The new anticipatory governance culture for innovation: Regulatory foresight, regulatory experimentation and regulatory learning. European Business Organization Law Review, 26(2), 241–283. https://doi.org/10.1007/s40804-025-00341-2*

[36] *Bernstein, A., & Rodríguez, C. (2023). The accountable bureaucrat. Yale Law Journal, 132(6), 1600–1690.*

[37] *Navas, S. (2020). Producer liability for AI-based technologies in the European Union. International Law Research, 9(1), 77. https://doi.org/10.5539/ilr.v9n1p77*

[38] *Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human–AI collaboration. Scientific Reports, 14(1), 31150. https://doi.org/10.1038/s41598-024-82501-9*

[39] *Gignac, G. E., & Ilic, D. (2025). Psychometrically derived 60-question benchmarks: Substantial efficiencies and the possibility of human-AI comparisons. Intelligence, 110, 101922. https://doi.org/10.1016/j.intell.2025.101922*

[40] *Jabbar, M. S., Al-Azani, S., Alotaibi, A., & Ahmed, M. (2025). Red teaming large language models: A comprehensive review and critical analysis. Information Processing & Management, 62(6), 104239. https://doi.org/10.1016/j.ipm.2025.104239*

[41] *Trabilsy, M., Prabha, S., Gomez-Cabello, C. A., Haider, S. A., Genovese, A., Borna, S., Wood, N., Gopala, N., Tao, C., & Forte, A. J. (2025). The PIEE cycle: A structured framework for red teaming large language models in clinical decision-making. Bioengineering, 12(7), 706. https://doi.org/10.3390/bioengineering12070706*

[42] *Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., & Heidari, H. (2024). Red-teaming for generative AI: Silver bullet or security theater? In Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (pp. 421–437). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732028*

[43] *Majumdar, S., Pendleton, B., & Gupta, A. (2025). Red teaming AI red teaming. In E. Raff & E. M. Rudd (Eds.), Proceedings of the 2025 Conference on Applied Machine Learning for Information Security (Vol. 299, pp. 66–85). Proceedings of Machine Learning Research.*

[44] *Casper, S., Li, Y., Li, J., Bu, T., Zhang, K., Hariharan, K., & Hadfield-Menell, D. (2023). Red teaming deep neural networks with feature synthesis tools. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems 36 (pp. 1–47). Curran Associates, Inc.*

[45] *Schiff, D. S., Kelley, S., & Ibáñez, J. C. (2024). The emergence of artificial intelligence ethics auditing. Big Data & Society, 11(4). https://doi.org/10.1177/20539517241299732*

[46] *Adams, I. T., Stoughton, S., del Pozo, B., Geary, I. T. J., Olson, M., & Alpert, G. P. (2025). Forced science: A critical appraisal of the scientific rigor of 'force science' policing research. Police Quarterly. Advance online publication. https://doi.org/10.1177/10986111251357498*

[47] *Morgan, D. (2023). Anticipatory regulatory instruments for AI systems: A comparative study of regulatory sandbox schemes. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 980–981). Association for Computing Machinery. https://doi.org/10.1145/3600211.3604732*

[48] *Lam, K., Lange, B., Blili-Hamelin, B., Davidovic, J., Brown, S., & Hasan, A. (2024). A framework for assurance audits of algorithmic systems. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1078–1092). Association for Computing Machinery. https://doi.org/10.1145/3630106.3658957*

[49] *Li, S., & Faure, M. (2025). The insurability of AI-related risks: Implications from the recent legislation in the European Union. Geneva Papers on Risk and Insurance - Issues and Practice, 50(3), 524–544. https://doi.org/10.1057/s41288-025-00352-5*

[50] *Uuganbayar, G., Yautsiukhin, A., Martinelli, F., & Massacci, F. (2021). Optimisation of cyber insurance coverage with selection of cost effective security controls. Computers & Security, 101, 102121. https://doi.org/10.1016/j.cose.2020.102121*

[51] *Buiten, M. C. (2024). Product liability for defective AI. European Journal of Law and Economics, 57(1–2), 239–273. https://doi.org/10.1007/s10657-024-09794-z*

[52] *Parlak, M., Altunel, N. F., Akkas, U. A., & Arici, E. T. (2022). Tamper-proof evidence via blockchain for autonomous vehicle accident monitoring. In 2022 IEEE 1st Global Emerging Technology Blockchain Forum: Blockchain & Beyond (iGETblockchain). IEEE. https://doi.org/10.1109/iGETblockchain56591.2022.10087067*

[53] *Antonio, Y., Indratno, S. W., & Saputro, S. W. (2021). Pricing of cyber insurance premiums using a Markov-based dynamic model with clustering structure. PLOS ONE, 16(10), e0258867. https://doi.org/10.1371/journal.pone.0258867*

[54] *Erdélyi, O. J., & Erdélyi, G. (2021). The AI liability puzzle and a fund-based work-around. Journal of Artificial Intelligence Research, 70, 1309–1334. https://doi.org/10.1613/jair.1.12663*