

ORIGINAL RESEARCH

Open access

Designing Transparent Public Sector AI Systems: Lessons from the UK Algorithmic Transparency Standard

Maria Silva^{1*}, Joao Pereira¹, Bruno Costa², Ricardo Alves³

Abstract

With the introduction of the Algorithmic Transparency Standard, UK authorities have renewed their emphasis on increasing operational openness for automated tools within public administration. This framework demonstrates that the UK is progressing its artificial intelligence regulatory goals while ensuring these benchmarks are grounded in the practical, real-world experiences of operators to achieve uniform, widespread integration. By utilizing a dual-level reporting structure, the Standard encourages inclusive disclosure tailored to different groups, thereby strengthening confidence among varied institutional and public actors. Furthermore, deploying these protocols across public agencies will likely serve as a blueprint for broader corporate benchmarks, guiding compliance models within commercial markets. This paper delivers a structured analysis and critique of the policy document.

Keywords Algorithmic transparency standard, Operational openness, Artificial intelligence regulatory goals, Inclusive disclosure

*Correspondence:

Maria Silva

maria.silva@uc.pt

¹ Department of Public Sector AI and Transparency Standards, Faculty of Engineering, University of Coimbra, Coimbra, Portugal

² Department of Algorithmic Transparency and UK Standard, Faculty of Engineering, University of Porto, Porto, Portugal

³ Department of Transparent AI Systems Design, Faculty of Engineering, University of Lisbon, Lisbon, Portugal

Introduction

Policy, legal, and regulatory discussions surrounding the mitigation of threats and maximization of advantages associated with automated processes (applying "artificial intelligence" or AI in its most expansive sense to cover all algorithmic architectures) are accelerating rapidly. Indeed, both within UK public bodies and international agencies, the primary focus has transitioned away from conceptual anxieties regarding societal damage and ethical implications—which characterized the foundational scholarship on AI Ethics—toward tangible implementation through drafted statutes and operational criteria (reflecting the contemporary landscape [1]). In this evaluation, we offer our preliminary assessments of the recently finalized Algorithmic Transparency Standard [2]. We interpret this document as a progressive measure by the UK state to both direct the international AI regulatory landscape and ensure that emerging compliance rules are shaped by functional deployment—specifically through trial programs—ensuring they remain both realistic and impactful; see [3] for a critical review and [4, 5] for arguments supporting this operational strategy. We map out the core components of these recommendations and critique several notable characteristics. Our primary conclusions are:

- **Signal:** We view the document as an explicit declaration that this issue remains an urgent priority for state administrators. Even so, it is essential that this strategy advances through rigorous trial phases alongside mechanisms that incentivize commercial enterprises to adopt equivalent protocols.

- **Method:** A state-led administrative strategy to enforce compliance protocols; here we implicitly acknowledge the capacity of public procurement requirements to cascade operational benchmarks into commercial supply chains. We also suggest that avoiding over-regulation at an early stage (via broad statutory mandates) ensures that future legal frameworks will remain grounded in pragmatic, field-tested realities.
- **Public transparency:** The dual-level classification offers a sensible strategy for communicating with varied target groups. More broadly, we recognize the structural difficulties inherent in translating complex processes for distinct stakeholder groups, particularly given how challenging it is to render automated operations accessible and clear.
- **Avoids data protection debate:** Acknowledging that the broader statutory debate surrounding data privacy is highly fluid (specifically regarding how the UK might liberalize its domestic rules), finalizing the revised national data privacy architecture is not a mandatory prerequisite for deploying the Algorithmic Transparency Standard.

We conclude by examining what we categorize as an openness-first model for automated governance, alongside an endorsement of initiatives designed to trial this disclosure framework and share the trial data with the public. We argue for the necessity of empirically supported rules that are verified through case studies and direct consultation with industry practitioners. Additionally, we isolate several omissions and opportunities for future expansion. While initiating this process via administrator-led trials is logical, comprehensive openness ultimately requires clear explainability for everyday citizens—a dimension that remains under-developed and requires further maturation, especially given widening disparities in digital literacy.

Overview

This section delivers an analytical breakdown of the structural components comprising the Algorithmic Transparency Standard.

The framework can be understood as a continuation of the policy trajectories established by the UK National Data Strategy [6] and the subsequent UK National AI Strategy [7]. Specifically, the framework addresses recommendations arising from the National Data Strategy public inquiry, which noted that to make oversight operational, public registers provide a highly tangible solution [8], pointing to the municipal authorities of Amsterdam and Helsinki as pioneers that introduced automated registries outlining how local agencies utilize software models to execute municipal programs. Furthermore, the framework serves as a direct application of the "state as an exemplar" doctrine [9], leveraging public administration as both an early adopter and a crucible for technological innovation [9], while utilizing state institutions to model secure, ethical deployments of automated systems [10, 11]. In short, our analysis suggests the UK is attempting to construct a localized ecosystem of institutional trust and oversight—potentially serving as a distinct counterweight to the EU regulatory model—that is driven by an innovation-led, market-friendly agenda intended to yield favorable geopolitical and economic advantages [9].

While this portion of the paper outlines the framework, it is constructed as an interpretive reading wherein the textual summary and our independent commentary are interwoven. We have divided this analysis into five thematic areas:

- Transparency tiers
- Methodology and purpose
- What exactly is the algorithmic part doing?
- Risk assessment
- Ownership

For each area, we present the mandatory criteria of the official document followed by our evaluation.

Transparency tiers

You must fill in both sections of the template. For the:

- *Tier 1 information, provide a short non-technical description of your algorithmic tool, and an overview of what the tool is and why the tool's being used*
- *Tier 2 information, provide more detailed technical information, such as specific details on how your tool works and the data the tool uses*

We interpret these disclosure categories as a direct intervention in the auditing phase of the automated oversight debate, which mandates verifiable confirmation that systemic threat assessments have been properly executed. From our perspective, institutional reporting involves two core metrics: the depth of disclosure and the target audience.

- **Level of reporting:** This maps directly to the systemic threat level of the software. Quite simply, elevated threat profiles trigger more exhaustive disclosure demands.
- **To whom:** External reporting is oriented toward several distinct groups of interested parties, including:
 - *Regulators:* Industry-specific and cross-market oversight authorities with the power to penalize compliance failures;
 - *Persons directly impacted:* End-users and prospective consumers who stand to be marginalized by outcomes driven by the system;
 - *Public:* This group encompasses individual citizens, general end-users, and prospective consumers;
 - *Other third parties:* Investigative journalists, civil society organizations, academics, and institutional researchers;
 - *Internal use:* System proprietors, internal risk management and compliance divisions, and corporate management structures.

Crucially, these reporting levels bifurcate automated disclosure based on the target consumer. The brief, high-level tier 1 disclosures are optimized for "persons directly impacted" and the general "public," whereas the provision of highly technical tier 2 variables—in conjunction with tier 1 summaries—is best reserved for "regulators," "internal use," and "other third parties." On the whole, we support the thesis that institutional communication must reflect the technical capacity of the recipient, and we view this dual-layer structure as a practical baseline for organizing external communications.

We note that this strategy departs from the regulatory path chosen by the EU [1], which imposes information disclosure mandates exclusively on "high-risk AI systems" without establishing distinct reporting paths for different audiences. Similarly, the US "Algorithmic Justice and Online Platform Transparency Act" [12] and the "Data Accountability and Transparency Act of 2020" [13] contain no such bifurcated disclosure rules. In the Asia-Pacific theater, China's "International Information Service Algorithm Recommendation Management Regulations" [14], though demanding greater structural clarity regarding how target systems operate [15], make no mention of adjusting disclosure depth based on the recipient. Australia, while lacking dedicated statutes governing automated administrative decisions [16], addresses these obligations via general disclosure principles within "Australia's AI Ethics Principles" [17], yet similarly fails to mandate audience-specific reporting variations. Likewise, the operational registries managed by the municipal governments of Amsterdam and Helsinki [18], alongside the federal framework used by the Government of Canada [19], do not incorporate a tiered presentation system based on user segments.

Methodology and purpose

Tier 1 information

Clarify:
• How the system functions
• The specific issue intended to be addressed by the system, and its method for resolving it
• The core reasoning or underlying justification for deploying the system

Tier 2 information

Detail the operational parameters of the system, specifying its intended applications as well as its limitations.
Elaborate on the rationale for adopting the mechanism, such as:
• Delineating the capabilities and non-capabilities of the technology, including potential misconceptions regarding its application (2.1)
• Documenting the expected advantages—such as cost-effectiveness, operational efficiency, or user convenience (2.2)
• Cataloging non-automated alternatives evaluated during the planning phase, or outlining the procedural choices made prior to system implementation (2.3)
Enumerate the technical configurations of the system, which should encompass:
• The architectural model category, such as a rule-based expert system or a deep neural network (2.4)
• The frequency of system utilization—such as monthly transaction volume or the total number of individuals engaging with the software (2.5)

- The current lifecycle stage—categorized as conceptualization, design, engineering, deployment, or decommissioning, alongside creation timestamps and subsequent modifications (2.6)
- Protocols for maintenance and validation, specifying the precise mechanisms and timelines for human verification of automated outputs (2.7)
- The overarching system architecture (2.8)

These informational criteria, which investigate the core function and purpose of the framework, directly engage with the risk-mapping dimension of algorithmic oversight. This area focuses on identifying distinct vulnerabilities and evaluating the severity of potential hazards.

- **Determining appropriate domains:** Concerning the elements requiring identification, we argue that the domains of systemic bias, interpretability, technical resilience, and data privacy map onto the engineering dimensions of a framework. Conversely, fields such as fiscal, brand equity, and regulatory compliance risks align with broader institutional governance considerations [20]. Crucially, we contend that the divide between engineering and governance parameters vanishes when evaluating holistic AI assurance, an integration that is operationalized through standardized reporting mechanisms.
- **Severity of risk:** Evaluating the latent negative impacts of a system is of paramount importance. This evaluation is bound to the contextual application and ultimate objective of the technology; for instance, software calculating public welfare allocations carries an inherently elevated risk profile, whereas an automated spam filter does not. By extension, a fully decoupled autonomous mechanism executing final actions introduces higher operational vulnerability than an assistive model designed to augment human decision-making.

We suggest that the target application, alongside the specific engineering techniques deployed, establishes which risk categories demand the most intensive scrutiny. To illustrate, if a platform interfaces directly with ordinary citizens, mitigating algorithmic bias becomes critical, requiring granular data and source-code audits. If the infrastructure uses deep learning models, interpretability becomes the core priority to ensure diverse stakeholders can comprehend automated outputs. Ultimately, the intended deployment context serves as the primary metric for calculating overall risk, thereby dictating the necessary depth of public transparency.

What exactly is the algorithmic part doing?

Tier 1 information

Clarify:

The method by which the software is embedded within institutional decision-making pathways

Tier 2 information

Detail the integration mechanics of the software within the operational flow, explicitly noting the degree of influence the output exerts on final determinations. (3.1)
Elucidate the supervisory mechanisms exercised by human operators, covering:
• The volume and substance of data communicated by the system to the operator (3.2)
• The specific interventions and choices reserved for personnel, including protocols for auditing automated conclusions (3.3)
• Mandatory instructional programs required for workforce members tasked with deploying or managing the system, where applicable (3.4)
Outline the administrative review and contestation pathways, explaining the procedures available to citizens seeking to challenge or re-evaluate an automated outcome. (3.5)

The disclosure mandates regarding the system's role within administrative workflows concentrate heavily on the human-agency aspect of the broader AI accountability framework, while simultaneously intersecting with the previously detailed concepts of risk mapping.

Fundamentally, human-machine dynamics can be categorized into three operational frameworks:

- **Human in the loop:** No automated deductions or recommendations are executed without deliberate, proactive authorization by a human supervisor.
- **Human on the loop:** System outputs and decisions are automatically executed by default, but are monitored by an overseer who maintains the authority to intervene and override the process at any point.
- **Fully autonomous:** The system acts on its generated determinations and decisions automatically, entirely decoupled from real-time human intervention.

It is widely recognized that shifting from an assistive setup toward full autonomy increases systemic risk as human agency diminishes. Nevertheless, even within human-in-the-loop architectures, the reliability of human intervention can be compromised by intense data throughput, particularly when optimal performance depends on processing velocities that preclude manual review (e.g., algorithmic high-frequency trading). Consequently, while cataloging human oversight is a positive step toward identifying critical choice nodes and autonomous thresholds, there is a danger that increasing human management will be treated as an uncritical remedy for risk mitigation. Indeed, heightened human intervention can inadvertently re-embed subjective cognitive bias [21], and human cognitive processing itself lacks structural transparency [22].

Risk assessment

Tier 2 information

Data

Enumerate and outline:

- The data repositories utilized for model training
- The data repositories upon which the system is currently or will be implemented

Provide hyperlinks to these data repositories where possible

Incorporate:

- The specific designations of the utilized data repositories, where applicable (4.1)
- A structural overview of the data used for training and running the tool, detailing the specific variable types employed during the optimization, validation, or operational phases—such as 'chronological age', 'geographic address', and equivalent features (4.2)
- The web addresses (URLs) for the data repositories utilized, if accessible (4.3)
- The protocols and rationale governing data acquisition, or the historical protocols and rationale under which the data was originally amassed by a external entity (4.4)
- The formal data-sharing instruments and covenants currently in effect (4.5)
- Granular details regarding the entities authorized to access this information, alongside data retention windows and the precise conditions governing storage (4.6)

Impact assessments

Catalog the compliance and impact evaluations executed, such as:

- Data Protection Impact Assessment
- Algorithmic impact assessment
- Ethical evaluation
- Equality impact assessment

For each individual evaluation, append:

- The formal designation of the assessment (5.1)
- A narrative describing the specific evaluative process undertaken (5.2)
- The completion date of the evaluation (5.3)
- A hyperlink to the full evaluation document or an executive summary, where accessible (5.4)

Risks

Supply a comprehensive exposition of standard vulnerabilities associated with the technology, including:

- The nomenclature of these standard vulnerabilities (5.5)
- A detailed breakdown of each isolated vulnerability (5.6)

Illustrative examples encompass:

- Latent negative outcomes arising from the infrastructure being applied in a manner discordant with its intended design or scope
- The generation of asymmetrical or biased outputs, particularly through the use of training distributions that lack statistical representation or contain systemic biases
- Procedural arbitrariness and functional failure, such as the system generating inequitable or erroneous determinations

Mitigations

Supply a thorough account of the active interventions engineered to neutralize the vulnerabilities detailed in the 'Risks' section (5.7)

We interpret these disclosure parameters, which are mandated exclusively under the tier 2 framework, regarding 'Data' and 'Impact assessments' as providing necessary granularity to supplement the vulnerabilities mapped out in earlier risk-modeling stages. The 'Data' component clarifies potential threats to the civil liberties and statutory protections of end-users regarding information privacy. Concurrently, it exposes whether sensitive demographic features are integrated into the algorithm's construction or if such features precipitate discriminatory system tendencies.

Meanwhile, the sections covering 'Impact assessments' (for a highly relevant exploration of an equality impact assessment within an AI framework,[23]), 'Risks', and 'Mitigations' uncover the internal apprehensions identified by the system proprietor. These encompass vulnerabilities related to privacy, ethical alignment, equity, and the algorithm's core functionality, alongside the corresponding countermeasures designed to defuse them. In this context, system proprietors frequently observe and report operational trade-offs between data privacy and ethical fairness [24, 25], as well as mathematical friction between divergent definitions of fairness itself [21, 26-28]. The core of the information structured within this section can be systematically verified through independent, third-party technical audits (a mechanism that has notably been integrated into recent iterations of the drafted EU AI Act [1]).

Ownership

Tier 1 information

Clarify:
The mechanisms through which stakeholders can acquire additional information or submit inquiries—encompassing traditional offline channels alongside a dedicated contact email for the accountable institution, department, or designated representative

Tier 2 information

Catalog the entities legally accountable for the implementation of the system, including:
The parent institution (1.1)
The specific organizational unit managing the tool (1.2)
The senior executive designated as the responsible proprietor (1.3)
External contractors or third-party entities involved, provided the infrastructure was engineered externally (1.4)
The official corporate registration number (Companies House identifier) of the external contractor (1.5)
The functional responsibilities delegated to the external contractor (1.6)
The formal legal parameters governing their access to state-controlled data repositories (1.7)

For the sake of structural completeness, we append those components of the text that supply the target audiences of both reporting tiers with the necessary administrative contact details and asset ownership data to facilitate external inquiries. Of particular importance for the tier 2 audience is the requirement to disclose the legal terms regulating third-

party access to public sector records. We view this explicit disclosure as an indispensable transparency measure required to cultivate public confidence in the stewardship of state-held data assets.

Commentary

In this section, we provide focused commentary organized by key themes.

Signal

The Algorithmic Transparency Standard serves as a component of the National Data Strategy [6] and builds upon the National AI Strategy [7]. Both of these initiatives pledge to enhance algorithmic transparency across the public sector. In fact, the National AI Strategy identifies transparency as a core characteristic of AI that demands 'a unique policy response' [29]. It also outlines a National AI Research and Innovation Programme designed to support, among other goals, the cultivation of new capabilities centered on the 'transparency of AI technologies' [29].

By launching the Standard, the UK Government has reaffirmed this commitment. This dedication, combined with the rapid rollout of the initiative (occurring a mere two months following the National AI Strategy), signals the substantial weight the government places on this segment of the AI framework. Nevertheless, as discussed below, to transform this initiative from a positive signal into meaningful action, we recommend accelerating the piloting phase and opening doors for voluntary industry participation so that private sector actors can also submit entries.

Method

We endorse the strategy adopted by the UK Government regarding the Algorithmic Transparency Standard, particularly because it implicitly stimulates broader debates and agendas surrounding industry standards. We highlight three specific points:

- The Algorithmic Transparency Standard was explicitly created as a mechanism for UK public sector institutions. While the private sector is not its direct target, the widespread implementation of the Standard within the UK's massive public sector will naturally shape broader standard-setting. This is expected to influence best practices in private enterprise through the eventual adoption of identical or highly similar algorithmic transparency frameworks. We view this methodology as a clear application of the 'public sector as an exemplar' model [9-11], which aligns with the National AI Strategy [7].
- The Algorithmic Transparency Standard is currently undergoing a pilot phase, which provides the necessary room for iteration and refinement. This strategy leverages practitioner-led, empirically-grounded feedback that can be integrated into the framework before it is finalized. We anticipate that this iterative process will make the Standard more actionable and promote superior long-term adoption.
- We observe that this pilot-based strategy stands in contrast to the legislator-led methodologies pursued by both the EU [1] and the US [12]. However, it shares common ground with China, which utilizes 'AI pilot zones' to develop practical and effective policy mechanisms [30], as well as Canada, which solicits user feedback for its 'Algorithmic Impact Assessment Tool' [19]. Both of these nations similarly rely on a practitioner-driven approach.

Public transparency

We believe the tiered structure offers an effective mechanism for communicating relevant algorithmic transparency details to varied audiences. Specifically, tier 1 information provides accessible and succinct insights tailored to a broad, non-

technical audience, including the general public. We consider this level of accessibility a vital prerequisite for establishing widespread public trust in algorithmic systems.

Conversely, the tier 2 information criteria ensure that more technical audiences receive satisfactory details regarding the purpose, design, deployment, and risks of an algorithmic tool. Crucially, the granular detail required in tier 2 incentivizes model owners to follow established best practices by adhering to a structural 'processes' approach [31].

At present, we view this audience-tiering framework as a unique approach within the landscape of AI standard-setting, particularly when compared to existing algorithmic registers [18, 32].

Data protection

The Algorithmic Transparency Standard does not establish or rely on any presuppositions regarding the specific data protection framework governing an algorithm owner's operations. We interpret this neutrality as a deliberate design choice that allows the Standard to be deployed immediately, even while UK data standards remain a subject of active regulatory debate [33].

Conclusion

We conclude by reflecting on both the transparency-first approach to algorithmic accountability and the decision to pilot the transparency standard openly. Our observations underscore the critical value of standards that are empirically informed, verified by real-world case studies, and shaped by practitioner engagement.

- **Transparency first:** The Transparency Standard mandates that algorithm owners submit comprehensive details to the Algorithmic Transparency Standard collection regarding the purpose, ownership, architecture, and deployment of their algorithms, alongside assessments of data utilization and risk mitigation. Exposing the depth of an owner's diligence in this public manner creates external pressure to demonstrate that best practices have been strictly maintained. Specifically, owners must prove they have executed a rigorous process to thoroughly evaluate, investigate, and mitigate any risks or concerns linked to the algorithm. Furthermore, we anticipate that a degree of healthy competition among contributors to the repository could generate a beneficial ratchet effect, continuously driving up the baseline for best practices.
- **Pilot and Empiricism:** Framing standards through pilots to collect empirical data has previously been identified as a critical priority for ensuring proper AI governance [3]. We explicitly note that frameworks tested and refined through hands-on deployment by diverse practitioners turn out to be far more functional. This functionality, in turn, yields stronger compliance once the standards are officially finalized. Additionally, such practicality improves the clarity and standardization of the disclosures submitted by model owners to the Algorithmic Transparency Standard collection. By conducting this transparency pilot in the public eye, all stakeholders can observe its evolution, which fosters valuable feedback for both individual contributors and UK governmental or regulatory authorities. Making the standard-setting process itself transparent ultimately deepens public confidence in algorithmic systems.

In closing, we point out that the current template does not address how citizens will be equipped to genuinely comprehend these algorithms and their direct impacts on daily life. A significant digital divide persists and remains unaddressed. While we agree that the template marks a clear advancement for transparency, it has not yet fully mastered the underlying complexities of achieving it.

Acknowledgements

We would like to thank both Denise Almeida (UCL) and Elizabeth Lomas (UCL) for kindly providing comments and suggestions during the writing of this paper.

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 20 Jan 2025

Revised: 11 Mar 2025

Accepted: 09 Apr 2025

Published online: 15 September 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

European Commission. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts [Internet]. Brussels: European Commission; 2021 Apr 21 [cited 2021 Dec 7].

Central Digital and Data Office. Algorithmic transparency recording standard [Internet]. London: Cabinet Office; 2021 [cited 2021 Dec 7].

Polle R, Kazim E, Carvalho G, Koshiyama A, Inness C, Knight A, et al. Towards AI standards: thought-leadership in AI legal, ethical and safety specifications through experimentation. SSRN Electron J. 2021.
<https://doi.org/10.2139/ssrn.3935987>.

Bandy J. Problematic machine behavior: a systematic literature review of algorithm audits. Proc ACM Hum Comput Interact. 2021;5(CSCW1):1-34.
<https://doi.org/10.1145/3449148>.

Pagallo U, Aurucci P, Casanovas P, Chatila R, Chazerand P, Dignum V, et al. AI4People—on good AI governance: 14 priority actions, a S.M.A.R.T. model of governance, and a regulatory toolbox. SSRN. 2019.
<https://doi.org/10.2139/ssrn.3486508>.

Department for Digital, Culture, Media and Sport. UK national data strategy [Internet]. London: UK Government; 2020 [cited 2021 Dec 7].

Office for Artificial Intelligence. National AI strategy [Internet]. London: UK Government; 2021 [cited 2021 Dec 7].

Open Data Institute. Getting data right: perspectives on the UK National Data Strategy 2020 [Internet]. London: Open Data Institute; 2020 [cited 2022 Jan 6].
<https://doi.org/10.61557/JRJT2485>.

Kazim E, Almeida D, Kingsman N, Kerrigan C, Koshiyama A, Lomas E, et al. Innovation and opportunity: review of the UK's National AI Strategy. *Discov Artif Intell*. 2021;1(1):14.
<https://doi.org/10.1007/s44163-021-00014-0>.

Zapisetskaya B, De Silva S. National AI strategy: step change for the AI economy in the UK [Internet]. London: CMS; 2021 [cited 2022 Jan 24].

Yaros O, Prinsley MA, Vanryckeghem V, Randall R, Hajda O, Hepworth E. UK government publishes National Artificial Intelligence Strategy [Internet]. London: Mayer Brown; 2021 [cited 2022 Jan 24].

Markey EJ. Algorithmic Justice and Online Platform Transparency Act [Internet]. Washington (DC): United States Congress; 2021 [cited 2022 Jan 6].

Brown S. Data Accountability and Transparency Act: discussion draft [Internet]. Washington (DC): United States Senate; 2020 [cited 2022 Jan 6].

Cyberspace Administration of China. Provisions on the administration of algorithm recommendation services for internet information services [Internet]. Beijing: Cyberspace Administration of China; 2022 [cited 2022 Jan 24].

China Briefing News. China passes sweeping recommendation algorithm regulations [Internet]. 2022 [cited 2022 Jan 24].

Cox J, Lewih A, Halferty I. AI, machine learning & big data laws and regulations: Australia [Internet]. London: Global Legal Group; 2021 [cited 2022 Jan 24].

Department of Industry, Science, Energy and Resources. Australia's AI ethics principles [Internet]. Canberra: Australian Government; 2021 [cited 2022 Jan 24].

Haataja M, van de Fliet L, Rautio P. Public AI registers: realising AI transparency and civic participation in government use of AI [Internet]. Amsterdam/Helsinki; 2020 [cited 2022 Jan 6].

Government of Canada. Algorithmic Impact Assessment tool [Internet]. Ottawa: Government of Canada; 2021 [cited 2022 Jan 24].

Koshiyama A, Kazim E, Treleaven P, Rai P, Szpruch L, Pavey G, et al. Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms. *R Soc Open Sci*. 2024;11(5):230859.
<https://doi.org/10.1098/rsos.230859>.

Kazim E, Barnett J, Koshiyama A. Automation and fairness: assessing the automation of fairness in cases of reasonable pluralism and considering the black box of human judgment. SSRN. 2020.
<https://doi.org/10.2139/ssrn.3698404>.

Bernini M. The opacity of fictional minds: transparency, interpretive cognition and the exceptionality thesis. In: Garratt P, editor. *The cognitive humanities: embodied mind in literature and culture*. London: Palgrave Macmillan; 2016. p.35-54.
https://doi.org/10.1057/978-1-137-59329-0_3.

Institute for the Future of Work. Building a systematic framework of accountability for algorithmic decision making [Internet]. London: Institute for the Future of Work; 2021 [cited 2021 Dec 7].

Kazim E, Koshiyama A. The interrelation between data and AI ethics in the context of impact assessments. *AI Ethics*. 2021;1(4):219-25.
<https://doi.org/10.1007/s43681-020-00029-w>.

Agarwal S. Trade-offs between fairness and privacy in machine learning. In: *IJCAI 2021 Workshop on AI for Social Good*; 2021.

Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. *arXiv*. 2016;arXiv:1609.05807.
<https://doi.org/10.48550/arXiv.1609.05807>.

Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*. 2017;5(2):153-63.
<https://doi.org/10.1089/big.2016.0047>.

Berk R, Heidari H, Jabbari S, Kearns M, Roth A. Fairness in criminal justice risk assessments: the state of the art. *Sociol Methods Res*. 2021;50(1):3-44.
<https://doi.org/10.1177/0049124118782533>.

Keeling E. The UK's National AI Strategy: setting a 10-year agenda to make the UK a "global AI superpower" [Internet]. London: Allen & Overy; 2021 [cited 2022 Jan 6].

State Council of the People's Republic of China. China promotes local AI pilot zones [Internet]. Beijing: State Council of the People's Republic of China; 2020 [cited 2022 Jan 24].

Kazim E, Koshiyama AS. A high-level overview of AI ethics. *Patterns* (N Y). 2021;2(9):100314.
<https://doi.org/10.1016/j.patter.2021.100314>.

Delacroix S, Olhede SC, Blacklaws C, Bhatti SA. Algorithm use in the criminal justice system report [Internet]. London: Law Society; 2019 [cited 2022 Jan 6].

Almeida D, Shmarko K, Lomas E. The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks. *AI Ethics*. 2022;2:377-87.
<https://doi.org/10.1007/s43681-021-00077-w>.