State of the Evidence: Algorithmic Transparency

Matías Valderrama, María Paz Hermosilla and Romina Garrido | May 2023

Gob_Lab UAI
UNIVERSIDAD ADOLFO IBÁÑEZ

Open Government Partnership

Table of Contents

Introduction

Computational systems are everywhere mediating the fabric of contemporary social life, from public transportation and welfare provision to content moderation and criminal prosecution. Algorithmic decision-making (ADM) systems are increasingly intervening in government and business processes of all kinds. Search engines and social media platforms are deeply imbued by algorithmic systems with the power to rank, classify, moderate, or shape information and our social relations. Yet their decision-making processes are frequently opaque and inscrutable for citizens or even their developers.

Algorithmic systems can be opaque for many reasons. Following Burrell (2016), opacity can be intentional, to protect intellectual property, trade or state secrets, to conform to legal standards, or to avoid ways of gaming the system or violating other rights, such as privacy. They may also be opaque because of technical illiteracy or the lack of specialist knowledge of how to read the code underlying an algorithmic system. According to Burrell, algorithms can also be intrinsically opaque because of a mismatch between the level of complexity or high dimensionality and the human-scale reasoning. For whatever reason, the increasing inscrutability and opacity of algorithms has led to multiple voices calling attention to the growing power of algorithms and the need to hold them accountable (Diakopoulos, 2014, 2015; Pasquale, 2015).

This call for greater transparency of algorithms becomes especially relevant when considering ADMs in public services, where an automated decision could affect such sensitive issues as child protection or the allocation of social benefits. Given the impetus to move toward so-called smart cities, experts have warned of the potential dangers of introducing opaque, automated technologies as tools in urban governance (Tironi and Valderrama, 2022), which often contract private providers that do not offer detailed information or access to the code of their systems. This would entail the risk of algorithmic opacity leading to a "corporate capture of public power" (Brauneis and Goodman, 2018, p. 7). This is why for some authors, achieving more transparency in algorithms is an imperative to verify that they do not cause harm, escape legality, or end up reproducing socially unacceptable inequalities (Ada Lovelace Institute et al., 2021; Ada Lovelace Institute & DataKind UK, 2020).

Transparency is constantly mentioned in artificial intelligence (AI) principles and guidelines as an appeal to open the black boxes of AI and algorithmic systems and generate more trust in their applications. In fact, the concept of transparency is one of the main principles found across AI guidelines. For example, in a systematic review of 84 documents containing ethical principles or guidelines for AI, researchers found that the most repeated principle was transparency (Jobin et al., 2019). In another study conducted by the Berkman Klein Center for Internet and Society at Harvard University, transparency and explainability principles were present in 94 percent of documents researched in the data set, which covered 36 documents of prominent AI principles (Fjeld et al., 2020). As an example, one of the OECD principles for responsible stewardship of trustworthy AI[1] states that actors must “commit to transparency and responsible disclosure regarding AI systems” both to achieve a general understanding of how these systems work and to enable those affected by AI systems to challenge their results.

[1] For more information, see OECD.AI Policy Observatory, “Transparency and Explainability (Principle 1.3),” https://oecd.ai/en/dashboards/ai-principles/P7.

Meanwhile, multiple scientific studies, journalistic reports, and activist demands have pushed for laws and regulations to establish stricter measures for algorithmic transparency and accountability. (For a discussion of recent bills and regulations, see Oduro et al. 2022.) Examples are the proposed US Algorithmic Accountability Act of 2019 and 2022, the European Union AI Act proposed in April 2021,[2] the French Digital Republic Law of 2016, and the Canadian Directive on Automated Decision-Making of 2019. In addition, governments worldwide are developing algorithmic transparency guidelines or standards in the public sector, like the UK Algorithmic Transparency Standard of 2021 and the Chilean General Instruction of Algorithmic Transparency.

Despite the importance given to algorithmic transparency in all these documents and initiatives, there is still some ambiguity about its definition, mechanisms to implement it, and its impact on society. Moreover, it remains unclear how the regulations of algorithmic decision-making systems should be translated and put into practice, and how to implement many of the mechanisms promoted in these laws and regulations. This is why it is becoming increasingly important to start exploring the concepts, mechanisms, and results of algorithmic transparency.

To date, there are a few literature reviews that address the introduction of ADM in the public sector (Ada Lovelace Institute et al., 2021; Kroll et al., 2017; Levy et al., 2021). These reviews and other studies tend to separate transparency from accountability. The former is usually understood as the disclosure of information of algorithmic systems, while the latter is defined as a much more complex and multidimensional approach involving more policy mechanisms, such as audits or impact assessments.[3] Following the framework proposed by Bovens (2007), accountability encompasses actors both describing, justifying, or giving accounts of the use of an algorithmic system to a forum or the public, and such actors being held responsible and facing consequences for the misuse of such an algorithmic system (Metcalf et al. 2021; Wieringa, 2020). However, for other authors, transparency should not be limited to the disclosure of information, but should include mechanisms for the evaluation of algorithmic systems to achieve meaningful algorithmic transparency that allows different audiences to be able to approve or disapprove the use of an ADM system in the public sector (Brauneis and Goodman, 2018; Garrido et al., 2021).

Rather than addressing the debate surrounding algorithmic accountability, this literature review focuses on transparency in a broader sense and seeks to bring more clarity to the concept of algorithmic transparency by identifying some of the main mechanisms that have been proposed in the literature to promote it. The main objective of this review is to examine the available evidence on the outcomes or results of such mechanisms. In this way, we want to contribute with an early stock-taking exercise of the research on impacts of algorithmic transparency in the public sector. This work is primarily intended for decision-makers to facilitate more precise and targeted policy discussions. Additionally, anyone who develops or works with algorithmic systems and wants to know the mechanisms under discussion to examine, evaluate, and make that system more transparent may find it useful.

2 For details, see Regulation of the European Parliament and the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM/2021/206 final), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

3 For a well-founded argument for such a distinction, see Wieringa, 2020; danah boyd, “Transparency ≠ Accountability,” Medium (blog), November 29, 2016, https://points.datasociety.net/transparency-accountability-3c04e4804504.

Defining Algorithmic Transparency

Just like the concept of transparency itself, algorithmic transparency can be an obscure concept. In the literature reviewed, there is a lack of a uniform vocabulary for algorithmic transparency and how it is operationalized in practice. In some cases, accountability or auditability appears as a dimension or facet of algorithmic transparency (Haataja et al., 2020; Springer & Whittaker, 2019), while in other cases, algorithmic transparency is often invoked as a dimension or mechanism to achieve algorithmic accountability. For example, in a report by the Ada Lovelace Institute, AI Now Institute, and Open Government Partnership (2021), transparency is conceptualized as one of eight policy mechanisms through which governments can pursue algorithmic accountability. They defined transparency mechanisms as the provision of "information about algorithmic systems to the general public (e.g. affected persons, media or civil society) so that individuals or groups can learn that these systems are in use, and demand answers and justifications related to such use" (Ada Lovelace Institute et al., 2021, p. 18).

While the demand for disclosures and more transparency in computational algorithms has existed for decades—for example, in web search engines (Grimmelmann, 2010; Introna and Nissenbaum, 2000)—the concept of algorithmic transparency gained popularity in the mid-2010s. The use of the term can be found in a 2012 paper by Gaffney and Puschmann (2014) on the lack of transparency of the Klout score calculation of an individual's level of influence on the internet. Another mention is found in historian Eden Medina's (2015) work on the Chilean Cybersyn project in the 1970s, emphasizing the importance of achieving not only greater transparency, but also democratic control over algorithmic systems, based on the work of cyberneticist Stafford Beer.

But perhaps the first scholar to use the concept more intensively was University of Maryland professor Nicholas Diakopoulos. Concerned about algorithms' growing power, Diakopoulos (2014, 2015, 2017) studies how journalists have adapted their traditional watchdogging and accountability functions to interrogate the power of algorithms and delineate their errors and biases in what he calls "algorithmic accountability reporting." Diakopoulos distinguishes between two main approaches: transparency and reverse engineering. On the former, Diakopoulos states that "Transparency can be a useful lever to bring to bear on algorithmic power when there is sufficient motive on the part of the algorithm's creator to disclose information and reduce information asymmetry" (Diakopoulos, 2015, p. 403). This transparency may be internally motivated by competition or public relations dynamics, or it may be externally imposed by government regulations that demand disclosure. "Such policies can improve public safety, the quality of services provided to the public, or have bearing on issues of discrimination or corruption that might persist if the information were not public" (Diakopoulos, 2015, p. 403).

However, as Diakopoulos (2014, 2015) argues, companies and governments usually have no obligation to make their algorithms transparent. Moreover, the objectives of "algorithm operators" may conflict with the desire for transparency. It has been argued that algorithmic transparency can expose trade secrets and undermine the competitive advantage of companies, damage their reputation, affect their business models, or create space for third parties to game and manipulate their algorithms. Diakopoulos therefore focused his attention on how journalists have taken a more adversarial approach by reverse engineering, analyzing five case studies of journalistic investigations that sought to hold algorithm developers accountable from the outside.

Early on, Diakopoulos (2014, 2015, 2016) organized a workshop on algorithmic transparency in the media at Columbia University's Tow Center for Digital Journalism in spring 2015. Then, he and Michael Koliska developed nine focus groups with experts from US news outlets and universities. From these workshops, the authors proposed to define algorithmic transparency as "*the disclosure of information about algorithms to enable monitoring, checking, criticism, or intervention by interested parties*" (Diakopoulos and Koliska, 2017; Koliska and Diakopoulos, 2018). This definition would be reproduced in other works, but one of the main questions that emerges from the definition of algorithmic transparency is: what kind of information can reasonably be made public about such algorithms? There is still no standard practice on what information should be documented and in what formats it should be made available to promote greater accountability and transparency (Ada Lovelace Institute et al., 2021, p. 45).

Initially, Diakopoulos proposed to consider five key "informational dimensions" that might be disclosed in a standard transparency policy for algorithms:

- **Human element:** involvement of developers, designers, or teams behind the algorithmic system;
- **Data:** databases used as input in the algorithmic system described in terms of accuracy, completeness, uncertainty, timeliness, representativeness, assumptions, and modes of collection, among other aspects;
- **Model:** the model of the algorithm itself with the main tools, features, or variables used as input, as well as the weights used in the algorithm;
- **Inferences:** the results of the systems, to enable benchmarking with standardized accuracy measurements (margin of error, confidence values, false positives, accuracy rate, etc.); and
- **Algorithmic presence:** the disclosure of when the algorithm or its outputs are used by end-users and how people are aware of them, something that he would later encompass under the concept of "interface."

In subsequent publications, Koliska and Diakopoulos make slight changes to this framework following the notion of the data pipeline, reorganizing the typology into four elements: data (inputs), model (transformation), inference (output), and interface (output). More recently, Diakopoulos (2020) has simplified and reordered this framework into three dimensions or layers: "including the level and nature of human involvement; the data used in training or operating the system; and the algorithmic model and its inferences" (Diakopoulos, 2020, p. 201). While this framework was intended for algorithmic transparency in news media, it reflects very well the aspects that could be disclosed about algorithmic systems in governments and businesses.

TABLE 1
Summary of transparency factors across four layers of algorithmic systems

Layer	Factors

Data	• Information quality.
	  ◦ Accuracy.
	  ◦ Uncertainty (e.g. error margins).
	  ◦ Timeliness.
	  ◦ Completeness.
	• Sampling method.
	• Definitions of variables.
	• Provenance (e.g. sources, public or private).
	• Volume of training data used in machine learning.
	• Assumptions of data collection.
	• Inclusion of personally identifiable information.

Model	• Input variables and features. • Target variable(s) for optimization. • Feature weightings. • Name or type of model. • Software modeling tools used. • Source code or pseudo-code. • Ongoing human influence and updates. • Explicitly embedded rules (e.g. thresholds).
Inference	• Existence and types of inferences made. • Benchmarks for accuracy. • Error analysis (including e.g. remediation standards). • Confidence values or other uncertainty information.
Interface	• Algorithmic presence signal. • On/off. • Tweakability of inputs, weights.

Source: Diakopoulos & Koliska, 2017, p. 817.

In another study, Diakopoulos (2017) applied this framework, which he called the algorithmic transparency model, to three cases. In the first case, the model was used to guide the disclosure of editorial information about a news bot that might be useful to other journalists as well as to end-users. In the second case, the model was applied in a more critical stance to problematize the biases embedded in a news product tied to Google search rankings. In the third case, the model was employed to detect gaps or opacities in an investigative journalism piece on Buzzfeed News. This would show the versatility and variety of dimensions that should be included when seeking to achieve algorithmic transparency.

The "transparency factors" proposed by Diakopoulos are similar to those proposed by Brauneis and Goodman (2018). In a study based on open records requests to federal and local governments in the United States, the researchers propose eight categories of information to disclose: "the algorithmic model's general predictive goal and application; relevant, available, and collectable data; considered exclusion of data; specific predictive criteria; analytic techniques used; principal policy choices made; results of validation studies and audits; and explanation of the predictive algorithm and the algorithm output" (Brauneis and Goodman, 2018, p. 66). As the authors argue, these categories of information do not necessarily point to perfect transparency but to what they conceptualize as "meaningful transparency" for the public or "knowledge sufficient to approve or disapprove of the algorithm's performance" (Brauneis and Goodman, 2018, p. 132). While this definition includes a number of the mechanisms discussed below, the authors are explicit in pointing out that giving access to the source code—the sign of full transparency according to the authors—does not necessarily lead to meaningful transparency. So intellectual property or trade secrets could be protected while achieving meaningful transparency. Furthermore, this definition of algorithmic transparency emphasizes the interpretability of information, rather than the disclosure of as much information as possible.

Another more recent definition of algorithmic transparency found in the literature is proposed by Utrecht University professor Stephan Grimmelikhuijsen, who incorporates procedural fairness theory and literature on government transparency into the discussion of algorithmic transparency. This definition includes two elements or dimensions: accessibility and explainability (Giest & Grimmelikhuijsen, 2020; Grimmelikhuijsen, 2022; Lepri et al., 2018). In this way, the author states that, "Algorithmic transparency is achieved when *external actors can access the underlying data and code of an algorithm and the outcomes produced by it are explainable in a way a human being can understand*" (Grimmelikhuijsen, 2022, p. 4). Under this definition, accessibility is not about simply making open to the public the source code, models, or underlying data. Even then, people—even experts—may not understand the impacts of the algorithmic system. Therefore, as the author points out, "accessibility means not just public availability, but accessibility means that external independent experts can access an algorithm for inspection and analysis to assess if it is compliant and does not violate any rules" (Grimmelikhuijsen, 2022, p. 4). Algorithmic transparency is thus a prerequisite for the performance of audits or inspections.

The second element is explainability, which can take many forms, from algorithms that operate transparently to systems that generate an explanation of why a particular output or decision was reached. Furthermore, such explainability can be aimed at different audiences, from experts to the general public.

TABLE 1 Core Dimensions of Algorithmic Transparency

	Description
Accessibility	Public availability of source code, model and/or data. This study focuses on access by external experts who can inspect and analyze an algorithm for bias and functionality.
Explainability	The outcomes of an algorithm can be explained in a way a human can understand how or why an algorithmic decision was reached. This study focuses on publishing underlying reasons of a decision.

Source: Grimmelikhuijsen, 2022, p. 3.

However, much ambiguity persists in the literature on how to define algorithmic transparency, the dimensions it integrates, its mechanisms to achieve it, and its impacts and benefits. While several authors refer to or discuss the definition offered by Diakopoulos and colleagues (Bitzer et al., 2021), the definition proposed by Grimmelikhuijsen has also been replicated in other studies (Criado et al., 2020). For its part, the idea of meaningful transparency seems to have permeated the discussion on algorithmic transparency standards in the UK (Ada Lovelace Institute, 2020; BritainThinks, 2021) [4] and Chile (Garrido et al., 2021). There also remains great ambiguity as to how transparency relates to other key concepts, such as accountability or fairness. Here, it is important to note that transparency is always "instrumental" or "merely a means" to accountability and in no way replaces it or guarantees it (Bovens, 2007; Brauneis and Goodman, 2018; Diakopoulos, 2020). Moreover, as Diakopoulos (2015, p. 403) has argued, algorithmic transparency is not a complete and well-defined solution to balancing algorithmic power, especially when companies or governments have no legal or other incentives to disclose information about their algorithms and account for their outcomes.

Nevertheless, several authors in this literature review suggest that algorithmic transparency is a key principle for advancing the regulation of ADMs in the protection of human rights (Brauneis and Goodman, 2018; Diakopoulos, 2020; Koliska and Diakopoulos, 2018; Springer and Whittaker, 2019). For the purposes of this review, algorithmic transparency will be understood as a relational achievement between different actors that can be internal and external to the algorithm development (algorithm's developers and controllers, affected individuals or communities, experts, journalists, governments, and the general public, e.g.). This achievement consists of the ability for actors to obtain information, monitor, test, critique, or evaluate the logic, procedures, and performance of an algorithmic system in order to foster trust and increase the accountability of the developers or controllers of the system. This means that algorithmic transparency is neither an inherent quality of algorithms nor a convention among a closed group of actors, but rather an achievement of a situation in which some actors are compelled to give accounts of an algorithmic system to others (Wieringa, 2020). Achieving that kind of accountability depends on—but is not exhausted by—algorithmic transparency. Furthermore, this implies that algorithmic transparency should not be seen as a fixed and static state of algorithms, but rather should be conceived as a dynamic and distributed achievement, requiring maintenance over time in the same way that algorithms themselves are constantly being changed and updated.

[4] For details, see Cansu Safak and Imogen Parker, "Meaningful transparency and (in)visible algorithms," Ada Lovelace Institute (October 15, 2020) https://www.adalovelaceinstitute.org/blog/meaningful-transparency-and-invisible-algorithms/.

Methodology

Scoping Review Process

To capture the state of the evidence in algorithmic transparency, a scoping review was conducted. A scoping review is a type of literature review method aimed at mapping, collating, and synthesizing existing literature on emerging, heterogeneous, and complex areas of research and debate (Čartolovni et al., 2022). As Arksey and O'Malley (2005) define it, a scoping review neither seeks to evaluate the quality of the included studies nor begins with a clearly stated research question and then restricts its search to publications with a predetermined study design to address it. Instead, it seeks to quickly "map" the significant research within an area of interest. In other words, it aims to provide a thorough coverage (breadth) of the publications that are already accessible. Among the reasons for choosing a scoping review to study algorithmic transparency, we can point out the novelty of this issue and the need to examine the extent of research and summarize and disseminate relevant research findings to policy makers and decision-maker bodies.

The general research question of the literature review was as follows: What is known from the existing literature about the mechanisms and outcomes of algorithmic transparency in the public sector? In other words, what do we know about the implementation of algorithmic transparency in the public sector? From a practical point of view, it was decided that the coverage of the review would be from 2015 to 2022 and would be explored in English and Spanish (the languages spoken by the researchers). To identify relevant documents, we used a variety of search engines, citations from works in the field, and expertise and personal recommendations from GobLab UAI and the Open Government Partnership. Because not all available evidence corresponds to academic publications, we combined searches in academic databases (Scopus, Web of Knowledge, ACM) with other search engines such as Google.

To our knowledge, there is no unified database on algorithmic transparency, but there are relevant publications in different repositories and sources. Likewise, there is highly relevant gray literature to consider policy mechanisms of algorithmic transparency. For these reasons, we developed a protocol for discovery and eligibility, adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework.

Query

After an iterative process of keyword searches, we decided to restrict the search to the concept of transparency and a series of concepts linked to computational technologies. We restricted the search to titles, abstracts, or keywords that included the concepts. We first started with very specific searches focusing on the concept of algorithmic transparency, getting only a couple hundred results. We then separated the search between the word *transparency* and several computational technology terms. The final search consisted of the following: [(algorit* OR "automated decision-making" OR "algorithmic decision-making" OR "machine learning" OR "artificial intelligence") AND transparenc*].

Scientific Databases

Using the specified query, searches were carried out in three scientific databases during December 2022. Querying the Web of Science database generated 3,598 results for the period 2015–2022. Querying the SCOPUS database resulted in 5,569 hits for the same period. Querying the ACM Digital Library generated 1,064 results for the same period. After merging the corpora, this resulted in 10,391 documents. As the query was broken down, 3,964 duplicates and errors were removed, leaving 6,427 unique titles. Noticing that similar documents continued to exist, the R packages Quanteda and RNewsFlow were used to filter out similar documents. After several tests, it was decided to exclude documents with a similarity of 80 percent, resulting in a sample of 6,006 documents.

We then selected the most relevant documents by establishing three eligibility criteria. First, empirical studies were favored over conceptual ones. Second, preference was given to articles dealing with the impact or consequences of the implementation and use of algorithmic systems, rather than their design or early development. And third, publications on policy mechanisms or cases of the public sector were favored. To apply these criteria, we searched the database, defining as eligible documents that included two or more key concepts, such as algorithmic transparency, empirical, impact, assessment, evaluation, implementation, disclosure, audit, register, oversight, accountability, explainability, and their derivatives. We did not include in this step very specific concepts that yielded no results (e.g., moratoria) or very general concepts that left an unmanageable number of results. In this step, we found that no small number of articles corresponded to off-topic studies (e.g., water transparency using computational methods). Once we filtered titles and abstracts against our first and second inclusion criteria, 157 documents were included in the full-text analysis. Of these, a significant number corresponded to papers in the areas of media and communication and health care (see Čartolovni et al. 2022). The first set of papers addresses the lack of transparency of algorithms in digital platforms, while the second discusses how to achieve greater legitimacy and trust in automated or AI tools in healthcare practice. In applying our third criteria, we focused on including articles that touched on public sector or policy issues but also included articles that allowed us to describe and elaborate further on the transparency mechanisms discussed below. From this process, 25 documents were included in the full-text analysis and in the final corpus.

Google Search

Second, a keyword-based web search of the Google.com search engine was performed in clean mode, after logging out from personal accounts and erasing all web cookies and history. The search was performed using the following keywords: “algorithmic transparency,” “AI transparency,” “automated decision-making transparency,” “algorithmic decision-making transparency,” and “machine learning models transparency.” Every link in the first 30 search results was followed and screened for articles or policy documents mentioning algorithmic transparency mechanisms, leading to the identification of seventeen non-duplicated results that were not present in the database in the previous step, of which five were included in the final review. Theses, position papers, news articles, principles and guidelines, blog posts, or publications unrelated to transparency in the public sector were excluded.

Citation Chaining

Third, after identifying relevant documents through the two processes described, we used citation chaining to manually screen the full texts and, if applicable, reference lists of all eligible sources in order to identify other relevant documents. Eleven additional sources were identified. To ensure theoretical saturation, we exhausted the citation chaining within all identified sources until no additional relevant document could be identified.

Final Corpus and Limitations

Combining all these document sources, we created a corpus of 41 documents on which this literature review is based. To briefly characterize these documents, they consist of 19 journal articles, nine reports, five book chapters, and eight conference papers. In addition, the recentness of these documents is striking. More than half of the documents were published in the last three years (25), with 2022 being the year with the most documents (11). We tried to establish the countries of origin of the documents, but several of them do not specify or do not focus on a specific country. What is clear is that most of the documents reviewed come from authors or institutions, cases, or data samples from the United States (19) and Europe (16). Only three documents correspond to experiences in Latin America.

We acknowledge several limitations to our study. First, we restricted our search method to general concepts like artificial intelligence and algorithmic decision-making. Our review's scope was quite broad and included all the available evidence on algorithmic transparency, even if additional terms (such as neural networks, deep learning, and so on) would have further expanded our search. Due to time and resource constraints for this study, we chose to focus primarily on umbrella words since they are often used in the literature rather than expanding our search to include other terms. Furthermore, we could have missed significant publications available in other languages because the keywords were restricted to English words. This may help to explain the predominance of evidence from the Global North. Future reviews could include concepts and search engines in other languages to cover more evidence on algorithmic transparency and to enable a broader view of how the definitions, mechanisms, and results of algorithmic transparency are being discussed in various countries.

Findings

What Does Algorithmic Transparency Look Like?

In the literature reviewed, we found multiple mechanisms that promote algorithmic transparency. These actions vary in terms of who makes the account of the algorithmic systems (system developers or external actors), the stages of the algorithmic life cycle (before, during, or after the system is put into use), and the actors or forums to which such actions are oriented (policy makers, civil servants, citizens, experts, affected communities, e.g.).

The following paragraphs present three broad categories of methods for algorithmic transparency: disclosures, explanations, and evaluations. Regardless of this categorization, these mechanisms should not be thought of in isolation or as mutually exclusive. As stated in a report by the Ada LoveLace Institute (2020), we currently have a "fragmented landscape of mechanisms for transparency that, taken individually or combined in the limited ways currently possible, leave us far from ensuring that we are capable of scrutinizing and evaluating the functions, or effects on communities and individuals, of ADM systems in use or under consideration in central or local government" (Ada Lovelace Institute, 2020, p. 1). Therefore, these mechanisms must be implemented in combination and dynamically to achieve algorithmic transparency.

DISCLOSURES

A first type of algorithmic transparency mechanism is the direct disclosure of information about algorithmic systems. However, there are important differences in who has to make that disclosure, how it is enforced, what kind of documentation it requires, and to which audiences it is directed.

INFORMATION REQUESTS

In the public sector, a first mechanism that has been discussed to increase the transparency of algorithmic systems is disclosure through requests for information. Following Levy et al. (2021, p. 321), this would be a reactive mechanism that "may sometimes be too little too late." Furthermore, it is still not clear what kind of information should be made public and when governments should not make public information about the algorithms they use in their processes. Moreover, this mechanism often clashes with the trade secrets of private providers of public services (Ada Lovelace Institute, 2020), meaning that in many cases, these requests for information are rejected.

Studying the US case, Fink (2018) analyzes the Freedom of information Act (FOIA) and related regulations and court rulings on the public disclosure of the government's algorithms. FOIA does not explicitly address the issue of government algorithms. Only certain regulations and policies of individual agencies, as well as court rulings, have provided recommendations and guidance on what and how information should be provided about their algorithms. From this analysis, Fink suggests that there are two reasons why it has been argued that the US government has no obligation to disclose information about the algorithms it uses. First, there is great ambiguity as to whether government algorithms—and even more so those developed by private companies—would correspond to "agency records" or "records" that fall under FOIA. And second, some interpretations and rulings have established that algorithms would fall under the FOIA exemptions, either because they correspond to internal rules or because they are trade secrets.

Using different approaches, Fink managed to collect 73 Freedom of Information requests that included the word or words "algorithm" or "source code." In only 21 of the responses to these requests, the algorithm or source code was provided. Reviewing the responses to these requests, Fink found that denials cited very different FOIA exemptions. None mentioned the internal rules exemption, citing reasons related to national security, trade secrets, privacy, and law enforcement investigations: “The variation in responses, as well as the disconnect between documented policies and actual practices, suggests a need for clarity on the applicability of freedom of information laws to government algorithms” (Fink, 2018, p. 1466). In this regard, Fink notes that the exemption from disclosure of information on public algorithms based on trade secret grounds should be rethought. Because government software is often outsourced rather than developed in-house, algorithms and their potential biases can remain opaque by decision. According to Fink, “Open government advocates have for years argued that the privatization of public services jeopardizes accountability (see, e.g., Feiser, 1999; Sullivan, 1987). Considering the increasing role algorithms play in government decision-making, however, those concerns gain new urgency” (Fink, 2018, p. 1466).

Another research effort that included FOIA requests is the work of Brauneis and Goodman (2018). They assembled a portfolio of 42 open records requests for different algorithms developed by foundations, corporations, and government entities for use by the government. These requests focused on six programs: Public Safety Assessment, Eckerd Rapid Safety Feedback, Allegheny Family Screening Tool, PredPol, HunchLab, and New York City Value-Added Measures. Of these, only one program (the program from Allegheny County) was able to provide both the prediction algorithms and substantial details on how they were developed. Multiple obstacles were encountered with the open records request, which led to frustration among the researchers. Some requests were rejected on the grounds that they were exempt from providing information or that the information was confidential and/or part of trade secrets. They received information on the development of the algorithmic systems for very few programs. This is probably because the vendors were left to provide it, and governments did not have such crucial information.

From this research, Brauneis and Goodman highlight three obstacles to this transparency mechanism that must be overcome in order to achieve true algorithmic transparency: “(1) the absence of appropriate record generation practices around algorithmic processes; (2) insufficient government insistence on appropriate disclosure practices; and (3) the assertion of trade secrecy or other confidential privileges by government contractors” (Brauneis and Goodman, 2018, p. 8). Hence, the authors suggest as fixes that governments use their "contracting powers" and establish in the contracting of ADMs “the proper creation, provision, and disclosure of records” (p. 164).

Brandão et al. (2022) also conducted FOIA requests in their research. Between August and October 2021, the researchers requested information from 30 municipalities of Brazil about the use of facial recognition (FR) systems in the public transportation system. In response to news about the use of these systems to prevent fraud in the provision of discounts or free services in public transportation guaranteed by law to specific demographic groups, the researchers wanted to learn more about the operation of these systems and their performance. Specifically, the researchers asked the municipalities to respond to a questionnaire with 40 questions via the Access to Information Law (Lei de Acesso à Informação). The questions of the questionnaire focused on six axes: “(i) general information about the use of FR systems, such as the starting date of use; (ii) general characteristics of the FR system used; (iii) measures adopted prior to the employment of the system, such as offering training to public agents to use it; (iv) measures adopted to make the

use of FR aligned with the purposes of LGPD [Brazilian General Data Protection Law]; (v) how the information generated by the FR system is supervised by humans; (vi) number of frauds identified by the system and how the holder of public transportation benefits [receives communications] when she/he has allegedly committed fraud" (Brandão et al., 2022, p. 570). Interestingly, the researchers argue that while Brazil does not have a law or directive on automated decision-making systems, the Brazilian General Data Protection Law (Lei Geral de Proteção de Dados Pessoais) does establish mechanisms that could move toward greater algorithmic transparency in the public transportation system: the requirement of free and informed consent and the elaboration of personal data protection impact reports.

By December 2021, 20 of the 30 municipalities had responded fully or partially to the questionnaire. Among them, 14 municipalities responded that they were indeed using facial recognition tools to avoid fraud in the provision of discounts and free services. Of these, only six municipalities responded completely to the questionnaire. From these responses, the researchers were able to contrast the different levels of information handled and disclosed by the municipalities, which could be linked to the resource levels of each municipality. However, they found that even in municipalities with a lot of resources, there was reluctance to provide information on FR systems. Thus, the researchers developed an algorithmic transparency score, analyzing practices that encourage, are neutral to, or hinder algorithmic transparency, based on the responses of the six municipalities. In this way, they were able to compare the municipalities and establish that in general there is a "very low" level of algorithmic transparency in the municipalities when using facial recognition systems. They also found that except in one case, most municipalities do not ask for consent when using these systems and present interpretations of the current law and justifications as to why they would not have to collect consent from citizens. In addition, personal data protection impact reports have been carried out in only two municipalities and were under development in another two. Overall, the researchers conclude that based on these requests for information, "the level of algorithmic transparency is low in the sector studied, which increases the chances that mistakes made by FR technologies are not challenged by citizens or other stakeholders" (Brandão et al., 2022, p. 577).

These studies, in two very different contexts (the United States and Brazil) show similar patterns in terms of obstacles to accessing information using this transparency mechanism. Likewise, the questionnaires and information requested vary, which may make it difficult to compare countries using this mechanism.

MODEL CARDS

Other transparency mechanisms include proposals for model or algorithm reporting. Unlike reactive disclosures made upon a request for information, researchers have proposed different formats for documenting algorithmic systems. These proposals are inspired by previous initiatives focused on datasets, such as Datasheets for Datasets (Gebru et al., 2021) or the **Dataset Nutrition Label** created by the **Data Nutrition Project**.

One example is a proposal by Mitchell and other researchers (2019). The authors argue that the documentation accompanying machine learning models (if provided) usually provides very little information on model performance, expected use cases, limitations and potential bias, or other information that would help users evaluate the suitability of these systems for their context. To remedy this situation, Mitchell et al. (2019) propose that released machine learning models should be accompanied by brief documentation that they called "model cards" to increase transparency

about how AI technologies work. In a conference paper, they set out the standardized procedures and contents that such model cards should include. The model card template includes basic information about the model; the intended uses during its development; the most relevant factors, including cultural, demographic, or phenotypic groups; metrics of the real-world impacts of the model (performance measures, decision thresholds, variability, and so on) or how the model performs differently when considering these groups; details of the datasets used in the evaluation and construction of the card; information from training data (if possible); results of quantitative analyses; and ethical considerations and recommendations.

The authors provide two examples: a smiling detection model and a public toxicity detection model. In both cases, one can see how well or poorly the models perform with marginalized groups or groups categorized by age or gender and learn important details of how the model works. “Model cards provide a way to inform users about what machine learning systems can and cannot do, the types of errors they make, and additional steps that could create more fair and inclusive outcomes with the technology” (Mitchell et al., 2019, p. 221).

This mechanism, although focused on the developers of the algorithmic systems, can be implemented in government procurement processes. Model cards not only give more transparency to an algorithmic system but also, by applying them to several models, allow comparison of performance for more informed decision making on ethical and fairness grounds. Developers, policy makers, organizations, and impacted individuals can better understand how different models work and perform according to different metrics and considerations, and have known benchmarks for the actual (un)suitability of a model in each context, such as in child welfare or migration services.

SOURCE CODE

A mechanism of great relevance in recent years is the publication of the source code of algorithmic systems. “Underlying source code can be an important mechanism for the technical transparency of algorithmic systems, and in general, is accepted as a best practice” (Ada Lovelace Institute et al., 2021, p. 47). Instead of focusing on descriptions of a model's performance and results, this may allow interested parties to review how the algorithmic systems work directly. While this action is more complex when trade secrets or security issues are involved, it can be a very useful mechanism for building public trust and legitimacy in high-criticality systems, such as contact tracing systems. This would ultimately allow for greater trust in the actions of governments, for example in the management of the pandemic or in the delivery of public benefits (Ada Lovelace Institute, 2020). These initiatives are part of the promotion of open standards in the publication of databases, operating systems, and pieces of code in non-proprietary formats, which not only provide transparency but also can enable innovation and experimentation by third parties. Examples of such mechanisms are Canada's Directive on Automated Decision-Making or France’s Digital Republic Law, which require that the source code developed by the government be made public, subject to certain exemptions for confidentiality or security.[5]

[5] For more examples from France, see Mission logiciels libres et communs numériques (Free platforms and digital commons mission), https://code.gouv.fr/#/repos; Guide juridique logiciels libres (Free platforms legal guide), https://guide-juridique-logiciel-libre.etalab.gouv.fr.

However, this form of disclosure can also be restrictive and may "potentially distract from other important disclosures" (Ada Lovelace Institute et al., 2021, p. 47). Only experts with a certain degree of knowledge can understand and work with the source code, and many of the models cannot operate without access to the database. As stated in different papers, the release of source code as transparency presupposes expert knowledge. And even experts in the position of inspecting the code would not be able to examine or evaluate the impacts of the algorithmic system without full documentation (Burrell, 2016; Kroll et al., 2017; Sloan & Warner, 2018). For authors such as Diakopoulos (2015) or Brauneis and Goodman (2018), an algorithmic transparency policy must develop an effective and meaningful user experience for transparency information. Rather than simply disclosing and publishing the source code, assuming that all users will be able to understand it, ideally the information disclosed should take into account levels of knowledge and ways to integrate it into end-users' decisions (Diakopoulos, 2015, p. 411).

ALGORITHM OR ARTIFICIAL INTELLIGENCE REGISTERS

A final mechanism of information disclosure that has become noteworthy in recent times is "algorithm registers" or "AI registers." These are consolidated repositories or directories of information on the algorithmic systems used by governments for free consultation (Ada Lovelace Institute et al., 2021). This type of mechanism can be considered "proactive" or active transparency, in contrast to the reactive transparency of Freedom of Information requests (Levy et al. 2021). These registers can be a good mechanism to "shed light on aspects of ADM systems, and the types of processing where it is not deemed appropriate to make public the source code or full datasets. They can help contextualize the function of ADM systems. Standardizing and making available documentation on the data produced in support of ADMs in a systematic and intelligible way could make a significant contribution to the transparency of ADM systems" (Ada Lovelace Institute, 2020, p. 12). Haataja et al (2020) from the Finnish company Sadot, which designed the AI registers in Helsinki and Amsterdam, describe these registers as follows:

> [A] standardised, searchable and archivable way to document the decisions and assumptions that were made in the process of developing, implementing, managing and ultimately dismantling an algorithm. With this, transparency, and when applicable, explainability, can be given for public debate, independent auditors, and individuals [sic] citizens. For civil society, it is a window into the artificial intelligence systems used by a government organisation. Ultimately, we hope it will become a catalyst for meaningful democratic participation and a platform for fostering mutual trust. (Haataja et al., 2020, p. 3)

Likewise, such records may come closer to what Kroll and others (2017) emphasize as disclosing the commitments of the algorithmic system, rather than revealing its source code from the outset. These registers have been implemented in cities such as New York, Ontario, Amsterdam, Helsinki, Antibes, and Nantes. More recently, the Eurocities network has been promoting the creation of these registers in different European cities. These "registers," "inventories," or "directories" vary in whether they are reports, spreadsheets, or continuously updated online repositories. These records, as in the case of Ontario, can be added to data or asset lists of government entities or be designed exclusively for the catalog of algorithms. In terms of

information, they also vary. In a quick review of these registries, one can find information ranging from the general purpose and logic of the algorithmic system and its uses or implementations, up to repositories with the source code of the system.

These registers have not only been promoted at city level but are also beginning to be established as part of algorithmic transparency standards at the state level.[6] For instance, given the mandate of the Digital Republic Law, the French government is in the process of making their algorithms more transparent. In 2020, the government created an open data task force, the Etalab, to implement the Digital Republic Law. Since then, it has been working on putting together a list of the decision-making algorithmic tools of the government and publishing their rules. Etalab published a first version of a guide in February 2021 with four categories of information: agency responsible; global context and how the algorithm is embedded in the decision-making process; impact of the decision; the algorithm's technical workings. The guide also includes contact information. "The scope and type of information of this register were designed to strike a balance between a maximum level of transparency and the agencies' resources to build registers" (Pénicaud, 2021).

Another example connected to algorithm registers is the UK's Algorithmic Transparency Standard, which is currently in a pilot phase. This standard was **launched** in November 2021 by the Cabinet Office's Central Digital and Data Office (CDDO) to "empower experts and the public to engage with the data and provide external scrutiny. Greater transparency will also promote trustworthy innovation by providing better visibility of the use of algorithms across the public sector, and enabling unintended consequences to be mitigated early on." To date, this standard consists of 1) an Excel/CSV spreadsheet created as a standardized method of gathering and presenting information on the algorithmic systems of the government and 2) a template that corresponds to the spreadsheet's sections and serves as a guide for public sector organizations in gathering the data necessary to complete the spreadsheet (Oswald et al., 2022). Since then, different algorithmic systems have been published in this register. According to the analysis of Kingsman et al. (2022, p. 2), this standard would be "establishing an ecosystem of trust (governance)—potentially an alternative to the EU's regime—that is underpinned by an innovation and opportunity agenda with a view to driving positive geopolitical and economic outcomes" (Kingsman et al., 2022, p. 2).

These registries are usually created within governments or with a private counterpart but can also be created externally to ensure that all critical information is included. An interesting case of a repository is **Algorithm Tips**, a web-based database developed by Diakopoulos, which is updated weekly by searching Google for documents on new uses of algorithms by the US government, automatically scoring their importance, and then evaluating the most important documents in terms of their potential negative impact.

Another example is the repository of public algorithms managed by GobLab UAI, the public innovation lab of the School of Government at the Adolfo Ibáñez University in Chile, which has been built with information obtained through desk research (Garrido et al., 2021; GobLab Universidad Adolfo Ibáñez, 2022). Additionally, and with the support of IDB Lab, GobLab has been **advising** the Chilean National Transparency Council in the creation of a Chilean algorithmic transparency standard that is currently under discussion by that public agency. This proposed

6 The International Organization for Standardization (ISO) is preparing a taxonomy of information that should be disclosed to help stakeholders identify and address the transparency needs of AI systems. For more information, see ISO, "ISO/IEC AWI 12792: Information technology — Artificial intelligence — Transparency taxonomy of AI systems," (n.d.) https://www.iso.org/standard/84111.html.

standard requires disclosure of information on the "rationale and effects of decisions taken by the automated decision system," the data used, means, costs of implementation and/or development of the system, and contact information of the operator, as well as setting out procedures or methods of challenge in the event of a complaint.

It is important to note that most of these transparency mechanisms based on the disclosure of information about the models are not mandatory, and if they are, they are not properly enforced (Ada Lovelace Institute, 2020). It is also key to ask who the actual audiences are that can interact with and form critical evaluations of the ADM based on the information available in these registries.

EXPLANATIONS

Another proposed mechanism to achieve greater algorithmic transparency is to increase the explainability of systems. With criticism of the biases and opacities of algorithms and legal discussions on the obligation of ADM controllers to provide explanations under the GDPR regime,[7] the machine learning community has developed a whole new line of literature around the concept of explainability, or XAI (explainable artificial intelligence). The idea behind this branch of research is often to build trust in AI systems by providing end-users with different explanations of how the algorithmic system arrived at a particular decision or outcome.

In a review of the methods of explainability, Guidotti et al. (2019) show multiple ways of approaching “black box problems” that depend on what is to be explained (i.e. specific decisions to the overall system) and how it is to be explained. Hence, they classify the methods into the following categories: model explanation problem, outcome explanation problem, model inspection problem, and transparent box design problem. Throughout these methods, one encounters different outputs or ways to visualize the explanations from a set of rules, scores, or decision trees, which shows that there is no clear agreement on what constitutes an explanation. In addition, most of these methods assume that the inputs or features are known, so they have problems with models that identify latent and unobserved features within their processing.

Following Schmidt, one important line of research in explainability focuses on providing importance scores for each feature used in a predictive model. An example of this approach can be what Datta et al. (2017) defined as quantitative input influence. Thus, one can understand that the algorithmic system made a decision fundamentally because of a set of variables or inputs more relevant or influential than others or because of certain pixels in image classification (Schmidt et al., 2020). These methods focused on feature scoring may vary in whether they are specific to certain types of models (e.g., neural networks) or rather are model-agnostic and can be used for any type of model. An example mentioned in this case is LIME, or local interpretable model-agnostic explanations (Ribeiro et al., 2016). Another approach related to the previous one is based on offering explanations with counterfactuals, i.e., to show how inputting different values for certain variables would affect the type of risk or classification determination by ADM (Sokol & Flach, 2019; Wachter et al., 2018).

7 The EU's General Data Protection Regulation (GDPR) not only sets out regulations to protect individual privacy, but also contains articles that protect individuals from automated decision-making systems that may be opaque or discriminatory. The Recital 71 of GDPR states that the data subject has the “the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.” However, it is still unclear how such explanations should look in practice.

All of these methods, in one way or another, have been celebrated for not needing to open the black box to gain an understanding of the ADMs (Wachter et al., 2017), thus avoiding problems of trade secret disclosure and allowing ways to game the system.[8] However, while these explanations contribute to the transparency and understanding of the ADMs, they can turn out to be a new "technological fix" to a much more complex socio-technical problem, so this mechanism cannot be thought of in isolation. Furthermore, many of the explainable methods in use tend to be local (i.e. they focus on explaining how specific decisions affect individuals), which can make it difficult to understand the impacts on communities or groups affected by ADMs.

EVALUATIONS

A third group of mechanisms that achieve algorithmic transparency are algorithm evaluations. While these tend to be included within the concept of accountability, we incorporate them here to go beyond just the disclosure of information about algorithms and their workings to address the examination of their impacts on potentially affected individuals and groups. We already anticipated some of this when we talked about model cards. Ultimately, as we saw in the discussion on algorithmic transparency, transparency is not only about providing information to a public forum, but also making it possible for that forum to monitor and/or make critical evaluations of the algorithmic system.

AUDITS

The first type of evaluation can be algorithmic audits. Following Raji and Buolamwini (2019), an audit of algorithms can be understood in general terms as "the collection and analysis of the outcomes from a fixed algorithm or defined model within a system" (p. 429) to evaluate patterns of performance that negatively affect certain groups or individuals. Making algorithmic audits public can provide a mechanism to incentivize companies and governments to address the algorithmic bias present in their systems.

In a report by the Ada Lovelace Institute with DataKind UK (2020), two types of algorithmic audits are distinguished. First is the ***bias audit***, aimed at assessing algorithmic systems for specific biases by testing the systems' outputs to certain inputs. These are typically conducted by actors independent of or external to the development of the algorithmic system, like independent researchers, investigative journalists, or civil society organizations. These audits can be carried out in different ways, either by focusing only on the code or testing the outputs for different inputs using real people's accounts in crowdsourced project platforms or using bots (so-called sock puppet audits), which is done by automatically extracting information published on the web or by requesting data via the application programming interfaces (APIs) of the platforms or developers of the algorithmic systems in question. (For a review of multiple ways of conducting algorithm audits, see Sandvig et al., 2014). A repeated case mentioned in the literature of bias-focused auditing is

[8] In the works reviewed, there are no reported cases of gaming the algorithms of governments. A hypothetical example could be that releasing information about a detection system of tax evasion could lead to third parties learning how to evade detection as well. Another example are studies on the manipulation of search engine algorithms or social media platforms. It is well known, for example, the industry around Search Engine Optimisation (SEO), in which different practices are defined to improve the position of certain websites in search results by considering the variables that the search engine employs for the ranking. This example shows that gaming algorithms do not necessarily pursue negative ends, but can even be a tactic from the margins in the face of oppressive algorithmic systems, so it is always a contested and ambiguous practice (Ziewitz, 2019).

the "Gender Shades" project in which Buolamwini and Gebru (2018) audited three commercial face recognition APIs to evaluate their performance in classifying faces by gender and race, and to determine whether there were accuracy disparities. They found that darker-skinned women were systematically misclassified.

Considering that algorithmic systems mutate or experience new functionalities, an audit focused on biases can quickly become outdated. Thus, the Ada Lovelace Institute with DataKind UK (2020) proposes a second form of audit: ***regulatory inspection***, typically performed by regulators or audit professionals, with a more general or holistic approach that aims to assess whether algorithmic systems comply with regulations or standards. However, this audit format is still in its conception, and it is necessary to provide a legal framework to allow external inspectors to access algorithms, data, and outputs on a regular basis.

A similar differentiation can be found in Metcalf et al. (2021), who include under the notion of auditing what they call "critical third-party audits" or investigations external to the development of an AI system by academics, journalists, or independent researchers that have revealed negative impacts of such systems already in use. The authors also highlight internal auditing initiatives by technology companies and efforts to disclose information about their models. “Nonetheless, internal governance will always run the risk of legal endogeneity and lack external fora that can demand accountability for harms. Similarly, critical third-party audits, wherein the auditor has no formal access to the internal workings of the system, lack the ability to render impacts as changes to the system” (Metcalf et al., 2021, p. 740).

In their framework of ethical algorithm audits, Brown, Davidovic, and Hasan (2021) define these audits as “assessments of the algorithm’s negative impact on the rights and interests of stakeholders, with a corresponding identification of situations and/or features of the algorithm that give rise to these negative impacts” (Brown et al., 2021, p. 2). The authors emphasize prior consideration of the purpose of the audit (legal compliance, risk management, general ethical assessment), as well as the context in which the algorithm is deployed. Then, the audit should include a list of the relevant stakeholder interests (from AI vendors to affected individuals) and key metrics on ethically relevant attributes (listed in the following table). Only then can the algorithm's performance on each of these metrics be evaluated and analyzed for how relevant each metric is to the interests of each stakeholder, in what the authors refer to as the relevancy matrix. In this way, one not only focuses on performance metrics but also on how such performance affects each stakeholder differently.

Category	Metrics
Bias	Societal bias
	Statistical bias
Effectiveness	Accuracy
	Stability and repeatability
	Efficiency of data use
Transparency	Transparency of architecture
	Explainability
	Transparency of use
	Transparency of data use & collection
Direct Impacts	Potential for misuse and abuse
	Infringement of legal rights
Security & Access	Security & access in use of the algorithm
	Data security & access

Source: Brown et al. 2021, p. 4.

IMPACT ASSESSMENTS

A second type of evaluation distinguished by the Ada Lovelace Institute and DataKind UK (2020) is the Algorithmic Impact Assessment (AIA) that comes from the same lineage as environmental, human rights, and fiscal-impact assessments and more recently the Data Protection Impact Assessments of the GDPR. (For an excellent review of impact assessments and algorithmic impact assessments, see Metcalf et al., 2021). In all of these impact assessments, the fundamental principle is to analyze and measure how a situation is affected by a given action or project, compared to if the action or project were not undertaken or the baseline. In the case of AIA, the objective is to measure the impacts of a given algorithmic system on vulnerable groups and to demonstrate such impacts to a forum that allows the developers or controllers of such a system to be held accountable.

As shown by Oduro et al. (2022), recent regulatory proposals and bills are beginning to integrate algorithmic impact assessments as a requirement in the public sector in the United States and the European Union. For these authors, the AIA proposals do not only facilitate the documentation of societal harms of ADMs and reduce potential forms of discrimination or negative disparities for certain groups (elements commonly associated with accountability or fairness), but also advance public transparency, especially when the evaluations of algorithmic systems are set to be made public as a requirement.

Proposals for AIAs differ in the accountability relationships between the actor (who), the fora (to whom, when, and where), and the content (what) that would be used to create effective algorithmic accountability regimes. A concrete example repeated in the literature is the case of Canada's "**Directive on Automated Decision-Making**," which requires public policy program managers using "automated decision systems" to perform an algorithmic impact assessment (Ada Lovelace Institute and DataKind UK, 2020).[9] In this case, the AIA consists of an "electronic survey" or checklist answered by the organization that developed and uses the algorithmic system. From this survey, numerical scores are assigned in a rubric format to identify levels of risk, although it has been criticized as a shallow form of accountability (McKelvey and MacDonald, 2019; Metcalf et al., 2021). This impact assessment model has been adapted elsewhere—for example, it has been translated into Spanish and taken as a basis for the Guide for Algorithmic Impact Study of Uruguay's e-government agency, AGESIC.[10]

9 The Directive also establishes transparency rules, such as the requirement to notify affected individuals that an automated decision-making system will be implemented prior to decisions and to provide explanations after the decision is taken. For more information, see "Algorithmic Impact Assessment Tool," Government of Canada, date modified April 25, 2023, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

10 The Agency for the Development of Electronic Government and Information Society and Knowledge, (Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento, or AGESIC) is an executing unit under the President of the Republic of Uruguay. For more information, see AGESIC, Preguntas para la evaluación del Estudio de Impacto Algorítmico (EIA): Proyectos de sistemas automatizados para la toma de decisiones (October 2020), https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/sites/agencia-gobierno-electronico-sociedad-informacion-conocimiento/files/documentos/publicaciones/Gu%C3%ADa%20para%20el%20estudio%20de%20Impacto%20Algor%C3%ADtmico%20(EIA)_0.pdf

The Ada Lovelace Institute and DataKind UK (2020) describe a difference between two AIAs: ***algorithmic risk assessment*** and ***algorithmic impact evaluation***. These two types of assessments could be differentiated as ex-ante and post-hoc impact assessments. While the former seeks to make a prospective analysis and assess algorithmic systems for their possible future societal impacts or harms before the system comes into actual use, the latter assesses such impacts after it is already in use. For Metcalf et al. (2021), in contrast, impacts are measurable constructs or proxies, always limited, of both actual harms and potential harms or risks. Whether assessing future impacts of a system to be implemented or current impacts of a system in operation, impact assessments require looking deeply into the question of what impacts can be generated and combining a range of methods and expertise. In this sense, the authors propose that the very idea of impact is something to be discussed within AIAs. "There is no 'impact' without the accountability practices that define, detect, and act upon it. Likewise, there is no accountability without defining, by ongoing institutional consensus, what an impact is" (Metcalf et al., 2021, p. 743).

For Metcalf et al. (2021), it is key within these impact assessments to distinguish between the actual or potential harms and unintended consequences of algorithmic systems and the measurable impacts or evaluative constructs to describe and account for them. In this way, the "impacts" always "emerge from and through (i.e., co-constructed) relationships of accountability" (Metcalf et al., 2021, p. 744).

> The distance between "impacts" as measures of the difference in probabilities of a classificatory outcome between demographic categories, and the tangible risks and harms that arise from that difference is at stake here. While computational methods have demonstrated the ability to describe the former (disparate classificatory probabilities) in any number of important ways [1, 12, 44, 110], computational methods are less well-suited to measure the ways algorithmic systems produce and distribute the risk that people and groups might experience as a result of these classificatory processes. (Metcalf et al., 2021, p. 740)

The authors emphasize that in order to bring impact measurement closer to assessing harm, AIAs must not only be conducted with metrics, instruments, and measurements from computer science, but other knowledge and expertise, including local and indigenous knowledge and that of affected communities. Thus, the authors suggest that AIAs should not only focus on technical questions (how does the system work?) but should also address all the accountability relationships involved, in line with the ethical audit approach mentioned above.

These mechanisms are not yet well established at a practical level, and those that exist are still in the development or testing phase. However, there are several impact assessment formats that can serve as a reference, especially the Data Protection Impact Assessment (DPIA). A document by the Ada LoveLace Institute highlights that "DPIAs constitute the most productive tool for illuminating the function and social dimensions of ADM systems, as their remit covers a wide range of information, including data fields and sources, the system's function within broader administrative processes, the responsible officials, and the effects and legal basis for data processing" (Ada Lovelace Institute, 2020, p. 3). However, these documents are generally not commonly accessible for public debate, and it is difficult to connect information on data management of individuals to

possible impacts or algorithmic biases of models on communities. Furthermore, ambiguity persists in how high-risk data processing is defined and whether this definition of high risk can be equated to the assessment of the high risk of ADMs.

Results of Algorithmic Transparency

To date, little is known about the impacts or outcomes of algorithmic transparency. While positive results can be hypothesized in theoretical terms following Grimmelikhuijsen, few studies actually measure such impacts. The studies closest to analyzing the results of algorithmic transparency are qualitative case studies of specific algorithmic systems or scenario-based online experiments.

SCENARIO-BASED EXPERIMENTS

In linking the studies to the mechanisms reviewed above, we find that most of the studies address the impacts of disclosures and explanations. In this section, we review work that used an experimental design to analyze the outcomes of adding more information or explanations about algorithmic systems of governments.

Algorithmic system scenarios used in child welfare and visa applications in the Netherlands

For instance, Grimmelikhuijsen (2022) developed two scenario-based survey experiments with a representative sample of Dutch individuals (N=897) to find out how algorithmic transparency influences the perceived trustworthiness of automated decision-making. The first scenario consisted of a visa application process in which a person's request was denied because they had visited a "suspect country." The second scenario related to deciding when to search a house based on potential welfare fraud. Rather than simulating a real explanation or providing access to source code, these scenarios usually propose to imagine a case where one has access to and/or explanations of the algorithm.

The results of the experiments suggest that explainability has a positive effect on the perceived trustworthiness of algorithms in both scenarios. In the case of child welfare, explainability would also have a positive effect on trust for street-level bureaucrats who use the algorithmic system. This would be a desirable outcome because it can positively affect trust in the government and the decisions it makes, according to the author. Meanwhile, accessibility has a positive but non-significant effect on trustworthiness. This dimension also affects the perceived trustworthiness of bureaucrats who use algorithmic systems to support their decision making. The context of the decision and its level of discretion and impact are important—in the visa application experiments (low discretion), only explainability had an effect, while in the welfare fraud experiment (high discretion) both accessibility and explainability positively affect perceived trust in the algorithmic system. The author concludes that explainability, then, would be of greater importance to citizens than accessibility, but that both would be positive in increasing trust in government decision making using these algorithmic systems. However, it is important to consider how this effect is moderated by the area and the levels of discretion, and complexity of the decision-making process analyzed.

Contact tracing app in Germany

Even though it moves away from public sector scenarios, another study by Bitzer, Wiener, and Morana (2021) still offers empirical evidence of the positive effects of algorithmic transparency. The researchers studied how algorithmic transparency can affect the adoption of contact tracing apps. The authors review theories that could explain the factors influencing the adoption of these apps, but they emphasize that they focus on individual factors, leaving aside concerns about the opacity of the apps' algorithms. Contact tracing apps are technologies that were used by different governments to monitor and control the spread of COVID-19, enabling rapid detection of infection chains by capturing code exchanges between mobile phones. These technologies involved a high degree of surveillance of people's movements, which made them difficult to trust. For the authors, information disclosure could serve to reduce uncertainty and negative attitudes to apps and promote favorable user reactions, understanding, and trust. By conducting an online survey-based experiment that used vignettes[11] of different app screens of a fictional app on a small sample of German-speaking participants residing in Germany (N=116), the researchers found that increasing the amount of information disclosed significantly increases the likelihood that the app will be selected, comprehension of how it works, and trust in its use.

Algorithmic system scenarios in a child welfare and court system in the United States

A third study with a more robust experimental design was led by Schiff, Schiff, and Pierson (2022), who apply the literature on public value failure to study the implications of the adoption of automated systems within the public sector. From this conceptual framework, the authors argue that when public administrators delegate decision-making authority to technology for the sake of greater efficiency, this may clash with the expression of other public values, leading to a possible failure of those values and thus to lower ratings of government performance. The authors focus on three public values: fairness as "the absence of biases that could harm vulnerable groups" (p. 658); transparency as the understanding of how algorithms work, by "informing the public about the algorithm's existence and explaining how decisions or predictions are made" (p. 658); and human responsiveness in government services, or the importance of having empathetic and trustworthy interactions with government officials rather than cold contact through computer systems.

In order to evaluate the impact of these values of adopting artificial intelligence systems in government, the researchers conducted an online survey experiment in June 2019 of US adults recruited through Amazon's Mechanical Turk (N=1460), measuring people's reactions to hypothetical scenarios in which such values are not achieved. Based on two types of algorithms to predict risk (one to evaluate risk in the child welfare system, and the other to determine a detainee's risk of not showing up for trial to assess eligibility for pretrial release without bail), the authors developed eight vignettes that varied by policy sector (child welfare or court system) and public value failure (fairness, transparency, and responsiveness). The researchers evaluated the effect of the vignettes on the following dependent variables: (1) citizen support for government actions, (2) trust, (3) beliefs about government service quality, and (4) expectations of personal impact, in comparison to control groups that were not exposed to the vignettes.

11 In this type of experiment, vignettes are a set of systematically varied descriptions of news, subjects, objects or situations—usually hypothetical—in order to elicit reactions from participants.

The study showed that in the face of hypothetical public value failures, people respond with lower evaluations of government. The vignette on bias or lack of fairness is the one that presented the largest negative effects in the four outcome measures, although the vignette on lack of transparency also generated statistically significant effects on the dependent variables, except for the variable on expectations of personal impact. Of the three public values, the study finds that lack of responsiveness has the smallest and least statistically significant effects. “This suggests that the public may not be as directly concerned about the loss of human responsiveness itself; yet, the public remains concerned about other public values (i.e., fairness and transparency) that are potentially undermined when human discretion is replaced by technology” (Schiff et al., 2022, p. 664). The researchers found no differences by policy sector, but rather variations by political ideology measured in the survey with a scale from “very liberal” to “very conservative.” The researchers found that conservatives tended to favor the use of ADMs and evaluate the government less negatively compared to liberals, despite the public value failures described in the vignettes.

Comparing transparency measures to measure trust in a machine learning-based assistance system in Europe

Evidence that challenges the thesis of positive results of algorithmic transparency is found in the work of German scholars working with Amazon researchers (Schmidt et al., 2020). Focusing on the question of how receiving information about a machine learning-based decision support tool affects human decision makers' trust in a model’s predictions, researchers conducted an online experiment in which human participants took on the role of decision makers in a series of classification tasks, specifically classifying the sentiment of movie reviews. Two hundred participants, mostly from the UK and other European countries, were recruited via the online platform Prolific.ac. Each participant had to evaluate whether the reviews were positive or negative. In the same interface, participants had at their disposal the predictions of a machine learning-based model on the same text. Along with that, the researchers tested what happened when showing two auxiliary measures of transparency to participants: relevant feature highlighting (in this case, words in the review that were relevant in the prediction for the model) and/or confidence scores. The researchers investigated how the transparency measures affected people's trust in the model's decision in classifying the movie review, finding that in both cases, there is a strong negative effect. And even in experiments with word highlights or confidence scores, humans were found to make more errors in classifying reviews.

> Our results challenge the common and popular narrative of providing highest possible transparency in order to build trust. Quite to the contrary, our results show that providing more insights into how an ML system arrives at its decision can have a negative effect on **trusting behaviour**. Importantly, this effect occurs predominantly for cases in which the ML system’s predictions are correct, showing that improvident use of transparency within assistive AI tools can in fact impair human performance. (Schmidt et al., 2020, p. 2)

Thus, the researchers suggest being cautious in the face of a desire for "maximal algorithmic transparency" and evaluating on a case-by-case basis how explanations are designed and how people interpret them. They write, "It is important to not only provide transparency, but also to make sure users also understand the means of transparency" (Schmidt et al., 2020, p. 14). Although this experiment does not address reactions to vignettes or conditions on an automated decision-making system in a bureaucratic or public service decision, it assesses in more detail the effects of explainability mechanisms than the works reviewed above.

While it is not the aim of this review to assess the generalizability or quality of these studies, their findings offer some insights into the impacts of algorithmic transparency, or the lack of it.[12] Overall, the evidence gathered in this literature review suggests that disclosing information or explaining the outputs of algorithmic systems can increase the use, understanding of, and trust in such algorithmic systems. Conversely, a lack of algorithmic transparency may negatively impact the trust in algorithmically mediated decisions, and thus undermine trust in governments and lower the evaluation of the quality of public services.

QUALITATIVE STUDIES

Thus, the researchers suggest being cautious in the face of a desire for "maximal algorithmic transparency" and evaluating on a case-by-case basis how explanations are designed and how people interpret them. They write, "It is important to not only provide transparency, but also to make sure users also understand the means of transparency" (Schmidt et al., 2020, p. 14). Although this experiment does not address reactions to vignettes or conditions on an automated decision-making system in a bureaucratic or public service decision, it assesses in more detail the effects of explainability mechanisms than the works reviewed above.

While it is not the aim of this review to assess the generalizability or quality of these studies, their findings offer some insights into the impacts of algorithmic transparency, or the lack of it.[12] Overall, the evidence gathered in this literature review suggests that disclosing information or explaining the outputs of algorithmic systems can increase the use, understanding of, and trust in such algorithmic systems. Conversely, a lack of algorithmic transparency may negatively impact the trust in algorithmically mediated decisions, and thus undermine trust in governments and lower the evaluation of the quality of public services.

Algorithmic transparency across three policy areas (UK focus groups)

In particular, studies related to the development of the UK Algorithmic Transparency Standard stand out. One of them was developed by the consultant firm BritainThinks (2021), which organized an online community and focus groups with 36 participants to explore how the government can be meaningfully transparent about algorithmic decision-making. Participants were confronted with three algorithmic decision-making use cases:

- a computer-assisted decision-making system for police officer allocation,
- a computer-assisted decision-making system for the recruitment of candidates, and
- a computer-assisted decision-making system based on the automatic recognition of vehicle license plates in parked cars.

[12] Such critiques outside the scope of this review include that these studies operationalize values like transparency in binary terms (bias or no bias, disclosure or no disclosure), or that they draw a complete separation between human and machine decisions, which is always much more complex and interactive in practice.

In the second phase, they were given categories or models of transparency. In the third phase, researchers worked collaboratively with the participants to develop a prototype of a transparency standard needed in each of the three scenarios.

The researchers found that at the beginning, there was little understanding of the concept of algorithmic decision making and almost no awareness of it in the public sector. Along with this, there was a lack of clarity about what level of transparency is appropriate. Likewise, the researchers found varying degrees of concern with the three scenarios, which is explained by the researchers as the effect of different perceived risks and impacts in each case. “The degree of **perceived potential impact and perceived potential risk** influences how far participants trust an algorithm to make decisions in each use case, what transparency information they want to be provided, and how they want this to be delivered” (BritainThinks, 2021, p. 21).

The researchers found tension between full transparency and simplicity. On the one hand, participants wanted the government to make the algorithms as transparent as possible. But at the same time, they wanted information that was simple and easy to understand. In discussions with participants, it was resolved that there should be two tiers of information with different levels of disclosure. “The two-tiered approach balances participants’ expectation that all transparency information is available to access on demand, whilst also ensuring that transparency information shared at the point of interacting with the algorithm is simple, clear, concise and unlikely to overwhelm individuals” (BritainThinks, 2021, p. 19).

Tier 1	Tier 2
• A summary of what the algorithm does, and where to get more information • Provided proactively	• All categories of information available • Easy to find but not shared proactively

Source: BritainThinks, 2021, p. 19

Thus, in the cases of use in which a high risk and impact is perceived (in this study, the recruitment system), participants demanded not only that information be available on a website but that there be active communication of the purpose and use of the algorithm. This two-tier system would eventually be replicated in the **final version of the Standard**. For algorithmic tools included in Tier 1, the institution should provide a basic description of how it works and why it has been introduced into the decision-making process, taking into account that the general public is the primary audience. For tools considered to be Tier 2, the institution should give more detailed information about the algorithmic tool, considering that interested parties such as NGOs, journalists, or other public sector organizations might need such information. In this way, a differentiated approach depending on the audience is privileged, making a difference with EU proposals to focus the explanations according to the criticality of the tools (Kingsman et al., 2022).

The researchers also analyzed how people's opinions changed as the phases of the study progressed, finding that in the case of recruitment, some people remained skeptical. But in the case of police officer allocation, as they gained access to more information, they gained both a better understanding of how the system works and trust in the algorithmic system.

Along with studying these three hypothetical cases of the use of algorithmic systems, the researchers explored how the information should be made public, by asking participants to compare the AI registers of Helsinki and New York mentioned above. The participants evaluated the Helsinki AI registry positively for its simplicity and the possibility to search for more information, and they evaluated the New York registry negatively for being a not very user-friendly database with content that took a very long time to read. Some participants reported feeling that New York's AI registry displayed too much information, and most pointed out that it was not oriented to the general public, but rather to academics or experts. This evidence points to how the information to be disclosed about algorithmic systems should not only consider the different degrees of risk or impact of algorithmic systems, but also the different audiences that the information is intended to reach. Thus, different formats are required for the general public and for experts.

The researchers concluded that the disclosure of information about government algorithm systems in simple and clear terms, as well as the active communication of more detailed information in cases of riskier or high-impact algorithms (e.g., by specifying their description and purpose), can increase both (1) public understanding and (2) public trust in the use of algorithms in the public sector.

Police personnel interviews on the implementation of the UK Algorithmic Transparency Standard

Another related qualitative study is the one by Oswald et al. (2022) in which 16 semi-structured interviews were conducted with police personnel who would be responsible for completing the UK Algorithmic Transparency Standard and with commercial organizations working on police algorithms. The researchers identified six overarching themes that were repeated in the interviews. First, interviewees agreed that the scope of the Standard is unclear, as it does not specify the range of algorithmic tools covered, raising questions as to whether even an Excel macro falls within the Standard.[13] Second, interviewees discussed the benefits of police participation in the Standard. Some emphasized that a key benefit is the "opportunity to demonstrate transparency and improve police legitimacy, crucial in England and Wales where 'policing by consent' is the prevailing model, with public trust and confidence the sine qua non of policing" (Oswald et al., 2022, p. 12). Other benefits mentioned were "addressing public anxieties" about potentially affected individuals and communities, and the improvement of the technical proficiency of policing technologies that can also enhance police legitimacy. However, commercial sector interviewees were more cautious and pointed out that not everything could be made transparent and that the information to be disclosed would be of interest to certain stakeholders. A more reflective and differentiated approach should be adopted according to the research. Third, interviewees claimed that full transparency could affect the perception of risk of the algorithmic system, either by demanding more responses from police officers or by creating a false sense that there are no detrimental impacts. Hence the

[13] In the **Standard**, an algorithmic tool is defined as "a product, application, or device that supports or solves a specific problem, using complex algorithms."

authors, like in Brauneis and Goodman (2018), emphasized the importance of achieving a "meaningful disclosure or legible explanation required for adequate public understanding of the quality and impact of an algorithm" (Oswald et al., 2022, p. 17). Other interviewees pointed to the possibility of gaming the system, although this would depend on the context.

Fourth, interviewees discussed a trade-off between implementing the Standard and adopting useful new tools, arguing that the Standard would raise barriers to innovation for both public sector and private sector partners by making it more costly or complicated to comply with the rules. Along with this, it was pointed out that the responsibility for transparency should also fall on the suppliers of the systems and not only on the government bodies. In this regard, one of the most frequently mentioned concerns or obstacles to algorithmic transparency was trade secrets and the difficulty of obtaining and publishing information from suppliers due to business issues.

A fifth, somewhat blurred, category identified by the researchers is the concerns around explainability, ethical scrutiny, and evaluation. Interviewees were interested in the government preparing comprehensible and succinct information on the technical processes of ADM so that non-experts can easily understand without information overload. In this sense, one of the desired outcomes of the standard is to make the technology explainable for both police and citizens. Along with explainability, there was agreement on the importance of ethical scrutiny as a positive outcome of the Standard. However, researchers point out that such ethical scrutiny may generate a certain fear or self-imposed chilling effect on innovation. This fear was also reflected in the question of how to ethically evaluate ADMs when they are in the testing phase. In some interviews, the possibility of bias and accuracy testing was raised, which according to one interviewee is not done, nor are the ADMs in police work designed to be able to do so. While there was support for publishing accuracy metrics or recall rates, there were concerns about how citizens might interpret these metrics.

The last and sixth category related to the human and financial resources required to comply with the standard. Some interviewees pointed out this may be a constraint to adopting algorithmic tools. Another possible outcome of the standard mentioned in the interviews was the "risk" of increasing Freedom of Information requests discussed above. Some interviewees did not believe that greater transparency would lead to a decrease in requests. On the contrary, the media and academics might increase requests as the use of algorithmic systems becomes more widely known. One interviewee noted on the increased attention on ADMs in police work, “I think that there's a huge risk of giving people sufficient information to get concerned, but not enough to actually satisfy themselves that it's not as bad as they think it may be” (Oswald et al., 2022, p. 24). Overall, interviewees were in favor of algorithmic transparency, as it could help demonstrate the legitimacy of police use of technology and build public trust, promote good practice among police forces and improve the use of technologies.

From these preliminary results from the UK Standard, elements emerge that can apply to other standards and AI registries and give us evidence of the possible public reactions to eventual disclosures of information on algorithmic systems in general. It is key to consider the different audiences of these disclosures, such as by including different levels of detail as well as performance metrics of the algorithmic systems, according to each audience or forum. It is also important to consider that detailed information should be provided by the suppliers of these systems. Moreover, information disclosure should aim to improve the quality of the technologies and government bodies, rather than become an administrative burden (Oswald et al., 2022).

Algorithmic fraud detection system in the Generalitat Valenciana

Another qualitative study is Criado et al. (2020)'s work analyzing SALER, an early warning system implemented in the government of the Generalitat Valenciana. For the authors, it is clear that transparency can not only contribute to better citizen understanding of government algorithmic systems, respect for citizens' rights (non-discrimination), and promotion of the protection of the general interest (legitimacy), but also help public sector decision makers and civil servants themselves make more informed decisions.

As the authors point out, the original purpose of SALER was to enable public service inspectors to analyze data on contracts, grants, subsidies, aid, and so on to detect, prevent, and even anticipate conflicts of interest and corrupt practices using machine learning-based models and descriptive analytics. Conducting six in-depth semi-structured interviews and documentary analyses, the researchers were interested in how this algorithmic system and algorithmic transparency could impact the decisions of public officials. From the research, the authors consider that the transparency of this system has contributed to important outcomes, such as controlling possible cases of discrimination and gaining legitimacy for the system and governance of the government, as well as helping civil servants themselves to make better decisions in their detection of corruption cases.

The authors found that the algorithmic system was considered a “supplementary indicator” in their decisions and therefore did not reduce their discretion or autonomy. The researchers also point out that SALER could easily be audited by actors within the government or by external agents, and that by law, biannual reports are required, although such evaluations are not yet reported given its recent implementation. One interviewee pointed out that the tool would be easy to audit, as it is just an application of rules, nothing too complex. That is, there are no predictive models at the moment, but rather the algorithmic system provides descriptive statistics and social network analysis.

The system works by answering questions from officials, and the interviewees were unanimous that the tool produces humanly understandable answers. This is interpreted as transparent and free of bias, and if there is bias, one developer pointed out that it is only “transferred” by the data. Thus, the researchers conclude that SALER's algorithmic transparency has a positive impact on the efficient work of civil servants in detecting corruption, as its results are easy to understand.

While this case does not provide evidence of the impacts of algorithmic transparency in predictive models, it does provide insight into how analytics and algorithmic systems are being introduced within governments and the importance of making them easily understandable and auditable.

Initiatives of information disclosure and evaluation of algorithmic systems in US cities

Another relevant qualitative study is the work of Baykurt (2022). She reviewed two examples of algorithmic transparency and accountability initiatives in US cities: New York City’s Algorithmic Accountability Task Force and Seattle’s surveillance ordinance in 2017. As the author shows, the first example sought to achieve greater transparency by disclosing information about the algorithmic systems used by the municipal government. However, it followed a “technology-centric view” that focused on providing transparency to the general public. In 2021, the task force resulted in the creation of an Algorithms Policy and Management Officer who published a report or an algorithmic tool directory in a PDF format—mentioned above—on the number of automated decision-making tools used within each municipal agency, with brief descriptions of what they do

with little citizen participation in the process. As Baykurt (2022) shows, this kind of transparency through disclosure resulted in a low legitimacy outcome, did not allow for citizen engagement, quickly became outdated, and would not be sufficient to address the impacts of such algorithmic systems.

In contrast, the approach of the municipal government of Seattle focused on assessing the impacts of its algorithmic systems on marginalized groups. The Seattle ordinance stipulates, among other things, keeping a "publicly available list of technologies in use or in any stage of procurement, inviting public comment and city council approval before acquisition, and delivering routine equity/impact reports for public review" (Baykurt, 2022, p. 5). In this way, the author highlights that Seattle's approach complements the emphasis on transparency and allows moving toward the use of algorithmic systems that mitigate negative impacts on citizens.

However, for the author, both the dissemination of information to the general public and the assessment of the impact on specific marginalized groups would not be sufficient without a critical analysis of the institutions involved in the design and use of these algorithmic systems. Taking as an example the case of a public controversy around San Diego's smart streetlights, the author argues that many of the automated systems implemented by weak municipal governments end up relying on work-in-progress and over-hyped technologies of powerful private companies and have no legal tools to enforce accountability for failures or misuse of such technologies. That is why the author suggests adopting a political-economic approach to algorithmic accountability that identifies the relationships between public agencies and private organizations in both the design and use of algorithmic systems in the public sector. "By this, I mean a model of algorithmic accountability that starts from the assumption that automated decision systems are designed and used by entities whose practices reflect particular economic and political interests" (Baykurt, 2022, p. 6). In this way, to understand the impacts of these systems, a broader analysis of how governments relate to tech companies in a process of increasing commodification of data must be included. This approach, according to the author, would go beyond transparency or forms of evaluation such as impact assessments, because it would seek to consider power relations, lack of enforcement, or even incompetence between municipal governments and the tech industry.

Ethnography of a French housing tax algorithm through Etalab

Another qualitative study is the ethnographic work of Loup Cellard (2022a, 2022b) at the French Etalab. For this author, the concept of algorithmic transparency is paradoxical because transparency is a fixed state that one tries to apply to algorithms that are inherently dynamic systems.[14] Moreover, Cellard problematizes the belief that there is a "utopian state" of algorithms in which they are fully understandable and transparent—commonly assigned to release their source code: "Algorithms cannot be made transparent because they are distributed systems implemented through the movements of numerous entities and practices" (Cellard, 2022, p. 7). Based on critical algorithm studies and inventive methods, Cellard argues that instead of calling for open algorithm codes or the backend computational workings, we should enable the transformation or reformatting of algorithms in what he calls a "surfacing algorithms method." Rather than documenting algorithms and accessing their depths, Cellard argues that we should design and focus on "the ability of citizens to produce meaningful accounts about such an algorithm: a localized, indexical, personalized, and fleeting ability-to-account necessary to produce a more collective, durable, and normative algorithmic accountability" (Cellard, 2022, p. 799).

[14] This, of course, varies from algorithms based on fixed rules to algorithms that continuously identify new patterns or dynamically change the main variables used for predictive models.

To this end, he proposes to design mediation devices that allow interaction with the "surfaces" of algorithms: "A surface is intended as a mediation device, an interaction between citizens and operators in charge of providing clarification and ideal justifications for their algorithmic decision-making systems" (Cellard, 2022, p. 799). This method was developed in a series of workshops concerning the housing tax algorithm at Etalab. Precisely these workshops gained momentum because of a controversy in which the French Ministry of Public Finance refused to respond to Freedom of Information requests directed at the housing tax algorithm. In the workshops, the algorithm was presented as a cooking recipe, and attendees were invited to speculate on the ingredients, to try to replicate the steps of the tax calculation, or to compare differences and similarities between different tax letters. With this, the attendees made sense of how the algorithm works, problematized the selection of its variables, or recomposed the relationships between the tax letter and an ecology of entities (fiscal tax rates, national legislation, citizens, fiscal administrators, e.g.), without necessarily having to see its source code. This type of intervention, inspired by the reverse engineering discussed above, would open the algorithms not to reveal their deep secrets but to generate collective instances of an “ability-to-account” between citizens and algorithm operators.

> If we try to escape the epistemology of a frontier between the internal workings and visible manifestation of an algorithm, if we forget the idea of transparency-as-openness, we can leave aside optical metaphors of access in favor of the relationality needed for accountability—hence connections between a distant technology and a knowing public can be woven anew. What then needs to be made accountable and recomposed in a meaningful way are relationships (between data and their sources, type of algorithmic treatment and their effects, some metrics and their weights, etc.) more than the algorithm understood as a monolith. It is through the haptic manipulation of surfaces in everyday settings and their experimental redesign that our ability to account for algorithms will be repaired. (Cellard, 2022, p. 812)

Both the ethnographic work of Cellard and the comparative examples of Baykurt challenge the evidence reviewed above and open new critical flanks to problematize the assumptions of algorithmic transparency. This kind of work suggests broadening transparency mechanisms toward more critical, collaborative, and participatory methods than just the publication of information. And they also contest the idea that algorithmic transparency should aim for greater legitimacy or trust in these technologies. Citizen coalitions, such as the one discussed by Baykurt (2022), can aim to stop the use of oppressive or invasive algorithmic systems, which may improve the evaluation of governments in the long run.

EVALUATIONS

The repercussions of audits or impact assessments of algorithmic systems remain understudied, or at least no studies have been found that address the implications of such evaluations in the public sector. This may be due to the recentness of the discussion and the fact that audit and impact assessment formats are still under development.

Impact of audits on facial recognition companies

Although they do not evaluate algorithmic systems of governments, a study by Raji and Buolamwini (2019) provides evidence of the impacts of auditability sought through algorithmic transparency. The authors examined the effects of the aforementioned "Gender Shades" audit one year after the first study. Along with including the same three companies evaluated in the first study (Microsoft, IBM, and Face++), they included two more (Amazon and Kairos). When they conducted the audit again, they discovered that all target systems had released updated API versions with less overall error by 5.7 to 7.7 percent. Disparities by subgroups were also reduced. Depending on the company, the classification error rate was reduced between 17.7 and 30.4 percent in the darker females' subgroup. In contrast, the companies that were not audited in the first study presented much larger overall errors and disparities by groups and subgroups.

While the results of the study do not establish that the companies audited in the first study changed their APIs because of the audit, they do offer clues that external audits can put pressure on companies to improve their classification models, such as by including an assessment of disparities by intersectional subgroups. Also, these audits can contribute to actions by civil society organizations and government entities and to greater awareness by users of these systems of their shortcomings and problems, which again puts pressure on companies to reduce the biases of their algorithmic systems. Another aspect highlighted by the authors is that the impact of bias audits can be restricted to audited companies only, as they found a high overall error rate and significant subgroup performance disparities in non-audited companies in the first study. Although this evidence corresponds to positive impacts on private companies, its scope can be extended to governments. Independent bias audits can benefit greater fairness and transparency, either by putting pressure on companies that sell ADM to governments or by auditing public algorithms of government entities.

Conclusions

Algorithmic systems are permeating different domains of society, crucially including the public sector. This implies a need to consider how the opacity of these algorithmic systems can affect citizens' expectations of transparency and fairness. In this literature review, we review a number of mechanisms that have been proposed to increase transparency in the use of algorithmic systems within governments. It can be noted that there is evidence to support the idea that algorithmic transparency can increase the trust and legitimacy of algorithmic systems in the public sector. However, it is important to take into account the evidence that complicates this relationship, considering for example the possible negative effects of algorithmic transparency on the performance of humans assisted by algorithmic systems.

The evidence reviewed here shows increasing efforts to make more transparent the use of ADMs in the public sector, mainly through disclosure mechanisms, explanations, and evaluations of algorithmic systems. Rather than reducing algorithmic transparency to a simple dichotomy between a system being “transparent” or “opaque,” we need to understand these systems dynamically and on a spectrum of information closures and disclosures (Garrido et al., 2021). As Diakopoulos claims, “there are many flavors and gradations of transparency that are possible, which may be driven by particular ethical concerns that warrant monitoring of specific aspects of system behavior” (Diakopoulos, 2020, p. 199). This spectrum, then, can be characterized from a low achievement of transparency to a maximum achievement of transparency when all stakeholders recognize and accept the account given by the ADMs' controllers. But as several researchers suggest, one should be wary of the desire for “maximal algorithmic transparency.” Instead, the different contexts in which these algorithmic systems are embedded must be addressed and a mix of mechanisms employed, while providing differentiated and meaningful transparency for each stakeholder. This implies, on the one hand, considering the mechanisms reviewed here not in isolation or as mutually exclusive, but as elements or dimensions of algorithmic transparency, understood as a “multidimensional goal” (Levy et al. 2021, p. 320) or social achievement. On the other hand, the multiple actors who may be interested in learning more about algorithmic systems implies the need to analyze which mechanisms allow algorithmic transparency to be guaranteed for different audiences. In this sense, within the spectrum of information disclosure, information can be shared differently across stakeholder forums—as in the two-tier model of the UK’s Algorithmic Transparency Standard, based on different audiences like the general public, journalists, or experts —to avoid both an excess of information and a lack of information.

In terms of the scope of the literature reviewed, there is a predominance of publications from the Global North, with a clear emphasis on cases from European Union countries, the UK, and the United States. This can be explained by the limitations of our methodology but also by the predominance of countries that are leading the debate on algorithmic transparency. Along with this, one thing to note from the scope of the literature found is the plurality of disciplines that discuss algorithmic transparency. Future work could greatly benefit from combining literature on the public sector with healthcare and media studies.

One important finding we discovered is that many of the studies analyzed address the question of how entities (news media, governments, companies, and so on) can make transparent the algorithms they are using or starting to use in their processes, as well as assessing how transparent the algorithms are. However, the question of the impacts or results of such efforts remains largely unaddressed. In a word, the emphasis is on the “how” and not so much on the consequences of algorithmic transparency. When looking for evidence on the outcomes of algorithmic transparency initiatives in the public sector, we mainly find scenario-based experiments and qualitative case studies. This may be due to the recentness of the topic, as well as the difficulty of measuring the impacts of these mechanisms in concrete metrics. But it can also be explained by the opacities of these algorithmic systems themselves. As they are difficult to access, it becomes more complex to study real situations of use of these algorithmic systems.

Another point that several of the studies seem to take for granted is that trust and reliability are positive aspects that should always be aspired to when achieving algorithmic transparency. It would seem that algorithmic transparency would only be desirable to achieve greater trust in the adoption of new technologies, leaving aside the possibility that greater transparency could positively increase scrutiny of these systems and perhaps demonstrate that, in certain cases, it is untrustworthy to use certain algorithmic systems in specific, critical areas.

Finally, it is important to consider the reactions that algorithmic transparency can generate once the mechanisms discussed here have been implemented. That is, all the mechanisms reviewed here are not neutral instruments but can have important interactive effects on the people affected by the ADM. This can be especially relevant with public audits or impact assessments, in that the entities under evaluation, when they become aware of being measured or monitored, may change their behavior in different ways (Metcalf et al., 2021). This implies considering the achievement of algorithmic transparency not as a finished product, but rather as a process of analyzing how different stakeholders react to, understand, and repurpose the information available about algorithmic systems for different ends.

References

Study	Location	Method	Transparency Mechanisms Discussed
Ada Lovelace Institute & DataKind UK. (2020). *Examining the Black Box: Tools for assessing algorithmic systems*. https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/.	UK	Review and synthesis of existing research and policy documents related to algorithm assessment tools	• Evaluations (Algorithmic impact assessments and audits)
Ada Lovelace Institute, AI Now Institute, & Open Government Partnership. (2021). *Algorithmic accountability for the public sector.* https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.	Various	Several methods: (1) a database of more than 40 examples of algorithmic accountability policies (2) semi-structured interviews with decision-makers and members of civil society (3) Feedback received at a workshop (4) Feedback from participants of a private roundtable at RightsCon 2021 (5) a review of existing empirical studies	• Disclosures • Evaluations (Algorithmic impact assessments and audits) • Explanations • Other reviewed mechanisms (principles and guidelines, prohibitions and moratoria, external/independent oversight bodies, rights to hearings and appeal, procurement conditions)

Study	Location	Method	Transparency Mechanisms Discussed
Ada Lovelace Institute. (2020). *Transparency mechanisms for UK public-sector algorithmic decision-making systems.* https://www.adalovelaceinstitute.org/wp-content/uploads/2020/10/Transparency-mechanisms-explainer-1.pdf.	UK	Policy review of existing UK mechanisms for transparency	• Evaluations • Disclosures (Source code, information requests, and standardized disclosure of data used or produced in the deployment of ADM systems)
Arksey, H., & O'Malley, L. (2005). "Scoping studies: Towards a methodological framework." *International Journal of Social Research Methodology*, 8(1), 19–32. https://doi.org/10.1080/1364557032000119616.	N/A	N/A	N/A
Baykurt, B. (2022). "Algorithmic accountability in US cities: Transparency, impact, and political economy." *Big Data & Society*, 9(2), 20539517221115426.	US	Comparative review of: • New York City's task force for regulating automated decision systems • Seattle's surveillance oversight ordinance • San Diego's coalition of activists and researchers organized against the city's smart streetlights initiative, launched in 2016 and shut down in 2020.	• Disclosures (AI or algorithm registers) • Evaluations (Algorithmic impact assessments)

Study	Location	Method	Transparency Mechanisms Discussed
Bitzer, T., Wiener, M., & Morana, S. (2021). "Algorithmic Transparency and Contact- tracing Apps–An Empirical Investigation." Twenty-Seventh Americas Conference on Information Systems, Montreal.	Germany	Survey-based online experiment (N= 116 completed and valid responses) using real-life scenarios in an adapted vignette technique using the crowdsourcing platform Prolific.	• Disclosures
Bovens, M. (2007). "Analysing and Assessing Accountability: A Conceptual Framework." *European Law Journal*, 13(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x	N/A	N/A	N/A
Brandão, R., Oliveira, C., Peres, S. M., da Silva Junior, L., Papp, M., Veiga, J. P. C., Beçak, R., & Camargo, L. (2022). "Artificial Intelligence, Algorithmic Transparency and Public Policies: The Case of Facial Recognition Technologies in the Public Transportation System of Large Brazilian Municipalities." In J. C. Xavier-Junior & R. A. Rios (Eds.), *Intelligent Systems* (Vol. 13653, pp. 565–579). Springer International Publishing.	Brazil	Questionnaire through public information requests to 30 municipalities about the use of Facial Recognition systems in the public transportation system of Brazil to prevent frauds in discounts and gratuities guaranteed by law to specific audiences.	• Disclosures (Information requests)

Study	Location	Method	Transparency Mechanisms Discussed
Brauneis, R., & Goodman, E. P. (2018). Algorithmic Transparency for the Smart City. Yale Journal of Law & Technology, 20, 103–176.	US	FOIA requests of 42 open records requests for different algorithms developed by foundations, corporations and government entities that would be in use by the government. These requests focused on six programmes: Public Safety Assessment; Eckerd Rapid Safety Feedback; Allegheny Family Screening Tool; PredPol; HunchLab; and New York City Value-Added Measures. Of these, only one programme from Allegheny County was able to provide both the prediction algorithms and substantial details on how they were developed	• Disclosures (Information requests)
BritainThinks. (2021). *Complete transparency, complete simplicity: How can the public sector be meaningfully transparent about algorithmic decision making?* https://www.gov.uk/government/publications/cdei-publishes-commissioned-research-on-algorithmic-transparency-in-the-public-sector/britainthinks-complete-transparency-complete-simplicity.	UK	Qualitative analysis of an online community and focus groups (N=36) to explore how the government can be meaningfully transparent about algorithmic decision-making.	• Disclosures (Algorithmic transparency standard)

Study	Location	Method	Transparency Mechanisms Discussed
Brown, S., Davidovic, J., & Hasan, A. (2021). "The algorithm audit: Scoring the algorithms that score us." *Big Data & Society*, 8(1), 2053951720983865.	US	It does not include a research method but presents a proposed method for conducting what the authors define as "ethical algorithm audits"	• Evaluations (Audits)
Buolamwini, J., & Gebru, T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.* Proceedings of Machine Learning Research. Conference on Fairness, Accountability, and Transparency.	N/A	N/A	N/A
Burrell, J. (2016). "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big Data & Society*, 3(1), 205395171562251. https://doi.org/10.1177/2053951715622512.	N/A	N/A	N/A
Čartolovni, A., Tomičić, A., & Lazić Mosler, E. (2022). "Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review." *International Journal of Medical Informatics*, 161, 104738. https://doi.org/10.1016/j.ijmedinf.2022.104738	Croatia	Scope review of literature	• Disclosures • Explanation • Evaluations

Study	Location	Method	Transparency Mechanisms Discussed
Cellard, L. (2022a). "Algorithmic Transparency: On the Rise of a New Normative Ideal and Its Silenced Performative Implications." In E. Alloa (Ed.), *This Obscure Thing Called Transparency*. Leuven University Press. https://doi.org/10.11116/9789461664464	France	Ethnography at the French open data task force Etalab to propose an algorithmic surfacing method based on a series of workshops on the housing tax algorithm.	• Disclosures (Participatory and inventive methods)
Cellard, L. (2022b). Surfacing Algorithms: An Inventive Method for Accountability. Qualitative Inquiry, 28(7), 798–813. https://doi.org/10.1177/10778004221097055	France	Ethnography at the French open data task force Etalab to propose an algorithmic surfacing method based on a series of workshops on the housing tax algorithm.	• Disclosures (Participatory and inventive methods)
Criado, J. I., Valero, J., & Villodre, J. (2020). "Algorithmic transparency and bureaucratic discretion: The case of SALER early warning system." *Information Polity*, 25(4), 449-470.	Spain	Exploratory case study of SALER, an early warning system implemented at an emerging stage in the government of Valencian region (GVA) in Spain (Interviews, documentary analysis)	• Disclosures (Accessibility, explainability)
Datta, A., Sen, S., & Zick, Y. (2017). "Algorithmic Transparency via Quantitative Input Influence." In T. Cerquitelli, D. Quercia, & F. Pasquale (Eds.), *Transparent Data Mining for Big and Small Data* (Vol. 32, pp. 71–94). Springer International Publishing.	US	Quantitative Input Influence method, for generating explanations of the outcomes of algorithmic systems is presented and discussed.	• Explanations

Study	Location	Method	Transparency Mechanisms Discussed
Diakopoulos, N. (2014). *Algorithmic accountability reporting: On the investigation of black boxes*. Tow Center for Digital Journalism.	US	Case study of journalistic investigations into algorithms that used reverse-engineering	• Disclosures
Diakopoulos, N. (2015). "Algorithmic accountability: Journalistic investigation of computational power structures." *Digital Journalism*, 3(3), 398-415.	US	Case study of journalistic investigations into algorithms that used reverse engineering	• Disclosures
Diakopoulos, N. (2016). "Accountability in algorithmic decision making." *Communications of the ACM*, 59(2), 56-62.	US	Workshop on algorithmic transparency in the media	• Disclosures
Diakopoulos, N. (2017). "Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens." In T. Cerquitelli, D. Quercia, & F. Pasquale (Eds.), *Transparent Data Mining for Big and Small Data* (Vol. 32, pp. 25–43). Springer International Publishing. https://doi.org/10.1007/978-3-319-54024-5_2	US	Application of the framework in Diakopoulos and Koliska (2017) to three qualitative case studies focusing on editorial information about a news bot, news product tied to Google search rankings, and investigative journalism piece on Buzzfed News.	• Disclosures

Study	Location	Method	Transparency Mechanisms Discussed
Diakopoulos, N. (2020). "Transparency." In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press.	US	N/A	• Disclosures
Diakopoulos, N., & Koliska, M. (2017). "Algorithmic transparency in the news media." *Digital Journalism*, 5(7), 809-828.	US	Focus groups (9) with participants (50) from national news outlets and universities. Focus on news production, curation, and dissemination, using specific cases: NLP to write content (creation), Facebook News Feed ranking (curation), and simulation in news stories	• Disclosures
Diakopoulos, N. (2017). "Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens." In T. Cerquitelli, D. Quercia, & F. Pasquale (Eds.), *Transparent Data Mining for Big and Small Data* (Vol. 32, pp. 25–43). Springer International Publishing. https://doi.org/10.1007/978-3-319-54024-5_2	US	Application of the framework in Diakopoulos and Koliska (2017) to three qualitative case studies focusing on editorial information about a news bot, news product tied to Google search rankings, and investigative journalism piece on Buzzfed News.	• Disclosures
Diakopoulos, N. (2020). "Transparency." In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press.	US	N/A	• Disclosures

Study	Location and Year	Method	Transparency Mechanisms Discussed
Diakopoulos, N., & Koliska, M. (2017). "Algorithmic transparency in the news media." *Digital Journalism*, 5(7), 809-828.	US	Focus groups (9) with participants (50) from national news outlets and universities. Focus on news production, curation, and dissemination, using specific cases: NLP to write content (creation), Facebook News Feed ranking (curation), and simulation in news stories	• Disclosures
Fink, K. (2018). "Opening the government's black boxes: Freedom of information and algorithmic accountability. Information." *Communication & Society*, 21(10), 1453–1471. https://doi.org/10.1080/1369118X.2017.1330418.	US	Qualitative analysis of 73 FOIA requests obtained in four ways: (1) Using FOIA to request prior requests that agencies had received. Specifically, this study sought FOIA requests from fiscal years 2010–2014 that included the terms 'algorithm' or 'source code.' (2) A search of FOIA logs that agencies had made available online. (3) A search of the FOIAonline database for the same keywords: 'algorithm' or 'source code.' (4) MuckRock, an investigative news site that helps people file and track FOIA requests.	• Disclosures (Information requests)

Study	Location	Method	Transparency Mechanisms Discussed
Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Berkman Klein Center for Internet & Society. http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420.	N/A	N/A	N/A
Gaffney, D., & Puschmann, C. (2014). *Algorithmic transparency and the Klout score*. https://www.researchgate.net/publication/276974372_Game_or_measurement_Algorithmic_transparency_and_the_Klout_score.	N/A	N/A	N/A
Garrido, R., Lapostol, J. P., & Hermosilla, M. P. (2021). *Transparencia Algorítmica en el Sector Público*. Consejo para la Transparencia y Gob Lab UAI.	Chile	Questionnaire through public information requests, Cadastre of ADM in public sector and case studies	• Disclosures
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). *Datasheets for Datasets* (arXiv:1803.09010). arXiv. http://arxiv.org/abs/1803.09010.	N/A	N/A	N/A

Study	Location	Method	Transparency Mechanisms Discussed
Giest, S., & Grimmelikhuijsen, S. (2020). "Introduction to special issue algorithmic transparency in government: Towards a multi-level perspective." *Information Polity*, 25(4), 409–417. https://doi.org/10.3233/IP-200010.	N/A	N/A	N/A
GobLab Universidad Adolfo Ibáñez. (2022). *Repositorio de Algoritmos Públicos de Chile. Primer informe de estado de uso de algoritmos en el sector público*.	Chile	Repository of public algorithms managed by GobLab UAI built with information obtained through desk research.	• Disclosures (AI or algorithm registers)
Grimmelikhuijsen, S. (2022). "Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making." *Public Administration Review*, puar.13483. https://doi.org/10.1111/puar.13483.	The Netherlands	Two scenario-based survey experiments: automated decision of a visa application and a bureaucrat who used an algorithm to predict welfare fraud	• Disclosure • Explainability
Grimmelmann, J. (2010). "Some Skepticism About Search Neutrality." In *The Next Digital Decade: Essays on the Future of the Internet.*	N/A	N/A	N/A

Study	Location	Method	Transparency Mechanisms Discussed
Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009.	N/A	N/A	N/A
Haataja, M., Fliert, L. van de, & Rautio, P. (2020). *Public AI Registers: Realising AI transparency and civic participation in government use of AI*. Saidot.	Europe	Offers proposals and recommendations on how to set up AI registers based on the experiences in the development of AI registers in Amsterdam and Helsinki.	• Disclosures (AI or algorithm registers)
Introna, L. D., & Nissenbaum, H. (2000). "Shaping the Web: Why the politics of search engines matters." *The Information Society*, 16(3), 169–185.	N/A	N/A	N/A
Jobin, A., Ienca, M., & Vayena, E. (2019). "The global landscape of AI ethics guidelines." *Nature Machine Intelligence*, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2	N/A	N/A	N/A

Study	Location	Method	Transparency Mechanisms Discussed
Kingsman, N., Kazim, E., Chaudhry, A., Hilliard, A., Koshiyama, A., Polle, R., Pavey, G., & Mohammed, U. (2022). "Public sector AI transparency standard: UK Government seeks to lead by example." *Discover Artificial Intelligence*, 2(1), 2. https://doi.org/10.1007/s44163-022-00018-4	UK	Critical evaluation of the policy of the UK's Algorithmic Transparency Standard.	• Disclosures (Algorithmic transparency standard)
Koliska, M., & Diakopoulos, N. (2018). "Disclose, Decode, and Demystify: An empirical guide to algorithmic transparency." In *The Routledge handbook of developments in digital journalism studies* (pp. 251-264). Routledge.	US	Focus groups (9) with participants (50) from national news outlets and universities. Focus on news production, curation, and dissemination, using specific cases: NLP to write content (creation), Facebook News Feed ranking (curation), and simulation in news stories	• Disclosures
Kroll, J.A., Joanna Huey , Solon Barocas , Edward W. Felten , Joel R. Reidenberg , David G. Robinson & Harlan Yu (2017). "Accountable Algorithms," *University of Pennsylvania Law Review,* 165, no. 633 (2017). https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.	US	Literature review	• Disclosures, • Evaluations (Audits)

Study	Location	Method	Transparency Mechanisms Discussed
Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. Philosophy & Technology, 31(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x.	N/A	N/A	N/A
Levy, K., Chasalow, K. E., & Riley, S. (2021). "Algorithms and Decision-Making in the Public Sector." *Annual Review of Law and Social Science*, 17(1), 309–334. https://doi.org/10.1146/annurev-lawsocsci-041221-023808	US	Literature review	• Disclosures (Information requests) • Evaluations (Algorithmic impact assessments) • Evaluations (Audits, among others mechanisms)
McKelvey, F., & MacDonald, M. (2019). "Artificial Intelligence Policy Innovations at the Canadian Federal Government." *Canadian Journal of Communication*, 44(2), 43-50. https://doi.org/10.22230/cjc.2019v44n2a3509.	N/A	N/A	N/A
Medina, E. (2015). Rethinking algorithmic regulation. Kybernetes, 44(6/7), 1005–1019. https://doi.org/10.1108/K-02-2015-0052	US and Chile	Historical review of the development of cybernetics and the Chilean Cybersyn project (1970s), for contemporary issues in big data and algorithmic regulation	N/A

Study	Location	Method	Transparency Mechanisms Discussed
Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). *Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts*. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 735–746. https://doi.org/10.1145/3442188.3445935	N/A	Literature review	• Evaluations (Algorithmic impact assessments)
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). *Model Cards for Model Reporting*. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596	US	Development of a proposal for machine learning model reporting based on a brief documentation that they called "model cards" to increase transparency about how AI technologies work.	• Disclosures (Model cards)
Oduro, S., Moss, E., & Metcalf, J. (2022). "Obligations to assess: Recent trends in AI accountability regulations." *Patterns*, 3(11), 100608.	Various	Review of four recently proposed and/or enacted bills and regulations in terms of their potential consequences for three major themes of algorithmic accountability: identifying and documenting possible harms, public transparency, and anti-discrimination and disparate impact.	• Evaluations (Algorithmic impact assessments)

Study	Location	Method	Transparency Mechanisms Discussed
Oswald, M., Chambers, L., Goodman, E. P., Ugwudike, P., & Zilka, M. (2022). "The UK Algorithmic Transparency Standard: A Qualitative Analysis of Police Perspectives." https://doi.org/10.2139/ssrn.4155549	UK	Qualitative study based on semi-structured interviews (N=16) with police personnel who would be responsible for completing the UK Algorithmic Transparency Standard, and with commercial organizations working on police algorithms.	• Disclosures (Algorithmic transparency standard) • Explanations
Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information* (1 edition). Harvard University Press.	N/A	N/A	N/A
Pénicaud, S. (2021, May 12). "Building Public Algorithm Registers: Lessons Learned from the French Approach." Open Government Partnership. https://www.opengovpartnership.org/stories/building-public-algorithm-registers-lessons-learned-from-the-french-approach/	N/A	N/A	N/A

Study	Location	Method	Transparency Mechanisms Discussed
Raji, I. D., & Buolamwini, J. (2019). *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 429–435. https://doi.org/10.1145/3306618.3314244	US	Bias audit of five API of Facial Recognition and structured disclosure procedure	• Evaluations
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778	N/A	N/A	N/A
Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms.* "Data and Discrimination: Converting Critical Concerns into Productive Inquiry," A preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA.	N/A	N/A	N/A

Study	Location	Method	Transparency Mechanisms Discussed
Schiff, D. S., Schiff, K. J., & Pierson, P. (2022). "Assessing public value failure in government adoption of artificial intelligence." *Public Administration*, 100(3), 653–673. https://doi.org/10.1111/padm.12742.	US	Survey-based online experiment with US adults (N=1460) recruited through the Amazon Mechanical Turk. Two scenarios were used based on two real-world cases of ADS in child welfare and in the court system	• Disclosures • Evaluations
Schmidt, P., Biessmann, F., & Teubner, T. (2020). "Transparency and trust in artificial intelligence systems." *Journal of Decision Systems*, 29(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094.	Global	Online experiment with participants recruited via the online platform Prolific.ac (N=200, majority from the UK (35%), other European (48%), or English-speaking countries such as the US, Canada, or Australia (12%)	• Explanations
Sloan, R. H., & Warner, R. (2018). "When Is an Algorithm Transparent? Predictive Analytics, Privacy, and Public Policy." *IEEE Security & Privacy*, 16(3), 18–25. https://doi.org/10.1109/MSP.2018.2701166.	N/A	N/A	N/A
Sokol, K., & Flach, P. (2019). *Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety*. Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019.	UK, 2019	Generating explanations of the outcomes of algorithmic systems is presented and discussed through counterfactuals	• Explanations (Counterfactual explanations)

Study	Location	Method	Transparency Mechanisms Discussed
Springer, A., & Whittaker, S. (2019). *Making Transparency Clear*. Joint Proceedings of the ACM IUI 2019 Workshops. ACM IUI Workshops, Los Angeles, USA.	US	It does not present a methodology but establishes a conceptual framework of algorithmic transparency that emphasises explainability and auditability and which is employed in other works	• Explanations • Evaluations
Tironi, M., & Valderrama, M. (2022). "Worth-making in a datafied world: Urban cycling, smart urbanism, and technologies of justification in Santiago de Chile." *The Information Society*, 38(2), 100–116. https://doi.org/10.1080/01972243.2022.2027587	N/A	N/A	N/A
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. https://doi.org/10.2139/ssrn.3063289	UK	Generating explanations of the outcomes of algorithmic systems is presented and discussed through counterfactuals	• Explanations (Counterfactual explanations)

Study	Location	Method	Transparency Mechanisms Discussed
Wieringa, M. (2020). *What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability*. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1–18. https://doi.org/10.1145/3351095.3372833	The Netherlands	Systematic literature review on algorithmic accountability (N=242), following the PRISMA statement and using Web of Science and SCOPUS with a recursive query design and computational methods. 93 'core articles' were identified as the most important. To structure the material, accountability theory was used as a focal point.	• Evaluations
Ziewitz, M. (2019). "Rethinking gaming: The ethical work of optimization in web search engines." *Social studies of science*, 49(5), 707-731.	N/A	N/A	N/A