The Right to Transparency in Public Governance

Freedom of Information and the Use of Artificial Intelligence by Public Agencies

Olsen, Henrik Palmer; Hildebrandt, Thomas Troels; Wiesener, Cornelius; Larsen, Matthias Smed; Ammitzbøll Flügge, Asbjørn William

Published in:
Digital Government: Research and Practice

DOI:
10.1145/3632753

Publication date:
2024

Document version
Publisher's PDF, also known as Version of record

Document license:
Unspecified

Citation for published version (APA):
The Right to Transparency in Public Governance: Freedom of Information and the Use of Artificial Intelligence by Public Agencies

HENRIK PALMER OLSEN, University of Copenhagen, Faculty of Law
THOMAS TROELS HILDEBRANDT, University of Copenhagen, Department of Computer Science
CORNELIUS WIESENER and MATTHIAS SMED LARSEN, University of Copenhagen, Faculty of Law
ASBJØRN WILLIAM AMMITZBØLL FLÜGGE, University of Copenhagen, Department of Computer Science

What information should and can be transparent for artificial intelligence (AI) algorithms? This article examines the socio-technical and legal perspectives of transparency in relation to algorithmic decision-making in public administration. We show how transparency in AI can be understood in light of the various technologies and the challenges one may encounter. Despite some first steps in that direction, there exists so far no mature standard for documenting AI models. From a legal perspective, this article examined the applicable freedom of information (FOI) regimes across different jurisdictions, with a particular focus on Denmark and other Scandinavian countries. Despite notable differences, our findings show that the FOI regimes generally only grant access to existing documents, and that access can be denied on the basis of the wide proprietary interests and internal documents exemptions. This is why we ultimately conclude that the European data-protection framework and the proposed EU AI Act — with their far-reaching duties to document the functioning of AI systems — provide promising new avenues for research and insights into transparency in AI.

CCS Concepts: • General and reference → Cross-computing tools and techniques; • Computing methodologies → Artificial intelligence; Machine learning algorithms; • Human-centered computing → Human computer interaction (HCI); • Applied computing → Law;

Additional Key Words and Phrases: Transparency, algorithm, freedom of information, administrative decision-making

ACM Reference format:

Research for this article was conducted as part of the research project 'PACTA — Public Administration and Computational Transparency in Algorithms' (grant no. 8091-00025B), which is headed by Henrik Palmer Olsen and funded by the Independent Research Fund Denmark. Authors’ address: H. P. Olsen, C. Wiesener, and M. S. Larsen, University of Copenhagen, Faculty of Law, Karen Blixens Plads 16, 2300 Copenhagen S, Denmark; e-mails: {henrik, cornelius.wiesener, mathias.s.larsen}@jur.ku.dk; T. T. Hildebrandt and A. W. Ammitzbøll Flügge, University of Copenhagen, Department of Computer Science, Universitetsparken 1, 2100 Copenhagen Ø, Denmark; e-mails: {hilde, awaf}@di.ku.dk. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. 2639-0175/2024/03-ART8 $15.00 https://doi.org/10.1145/3632753

Digital Government: Research and Practice, Vol. 5, No. 1, Article 8. Publication date: March 2024.
1 INTRODUCTION

Up through the 20th century, transparency — understood to refer to government openness — has been a leading normative ideal for public governance in democratic states. Believed to be an efficient instrument in uprooting corruption and misuse of government office, transparency has gained almost universal endorsement as a foundation for democracy and good governance. This ideal has also been written into law. In the United States, the Freedom of Information Act was adopted in 1966 [1]. Soon after, European countries followed the US example and enacted similar legislation on freedom of information (FOI) [2]. In Denmark, the first FOI Act entered into force in 1970 [3], with other countries in Europe introducing similar legislation around the same time [4].

FOI was, and still is, designed to make the apparatus of public governance transparent to the public. But transparency in government is not the only value worth protecting in society. Confidential policy information relating to national security, sensitive personal information, information relative to commercial interests (e.g., trade secrets), and efficiency in public administration are, among others, also important societal interests in need of legal protection. FOI legislation — both in the United States and in Europe — therefore seeks to balance transparency against these other interests.

As evidenced in some recent studies [5, 94], transparency is almost universally endorsed as a basic value for the ethical use of artificial intelligence (AI) both in the private and public sector. Nonetheless, this broad consensus belies the complexity involved in implementing a legal transparency regime for AI. Not only is it unclear how the interest in transparency may be balanced against other interests, what transparency really amounts to and how it can be achieved is also unclear. In line with Larsson and Heintz, we define transparency in AI broadly and thus as distinct from the narrow concept of algorithmic transparency. Rather than focusing exclusively on the algorithm itself, transparency in AI takes a system’s perspective and considers information asymmetries and possible beneficiaries of transparencies while balancing related concepts such as explainability and accountability as well as competing interests [95].

Also in the rich body of literature about digital government and the increased use of AI in the public sector [6–9], there is a growing focus on questions of transparency [10–12]. The topic features also prominently among contributions from Scandinavian countries, where a large part of administrative decision-making concerning taxes, student loans, and housing benefits is fully automated, while other areas are increasingly affected by AI-based decision-support systems [13, 14]. This explains also our focus on that specific region of Europe for the present article, in which we consider to what extent legal transparency procedures, including FOI, provide a useful avenue for securing AI transparency in the public sector.

In Section 2 of this article, we discuss how various AI models may lend themselves to possible transparency requirements. We distinguish between different kinds of AI systems, but we focus mainly on explaining how transparency in AI can be understood in light of the various technologies and outline the most predominant forms of AI systems. In keeping with the broad notion of transparency and since AI systems are always embedded in a larger bureaucratic apparatus, we add a human–computer interaction (HCI) perspective to our analysis, which demonstrates that transparency in AI also depends on how such systems are used within public administrative agencies themselves.

In Section 3, we analyse how current legal regimes regulate AI transparency in public administration. We delimit our focus to the regulation of public administration and examine those legal rules and principles that provide access points for citizens to demand information from public authorities about the AI systems at use in the administrative practice of the relevant authorities. The starting point for our analysis will be the Danish FOI Act, which is broadly representative of other Scandinavian countries. We will also cover the European data-protection framework and the proposed European Union (EU) AI Act.

In Section 4, we discuss how our observations may affect policy and practice as well as future research into AI transparency. In particular, it considers the perspectives of software developers and civil society activists in the interplay between the different legal regimes and avenues.
2 TRANSPARENCY IN AI MODELS

2.1 The Many Faces of AI

In the last decade, AI has increasingly been used interchangeably with the term *algorithm*. In its classical meaning, the term *algorithm* refers to a procedure providing a solution to a general and well-defined problem. An example of such a problem is to find the lowest number in a set of numbers or the shortest path between two points on a map. This concept of algorithms is rather different from the meaning of AI given in the pioneering article from 1959 by Turing Award winner John McCarthy, which refers to computer programs solving problems requiring a high degree of human intelligence [15]. Originally, this technology was used to solve problems related to the recognition of machine-printed characters in an image, referred to as *optical character recognition (OCR)*, or playing chess at the level of a professional player. More modern examples of AI would be the recognition of handwritten characters, referred to as *intelligent character recognition (ICR)* [16], and the recognition of plant and animal species [17] or cancer cells [18] in images. Most recently, this technology is being used in facial recognition and natural language processing.

Just like the implementation of classical algorithms, the engineering of AI software systems results in a computer program that computes a function from some set of given inputs to a set of possible outputs. However, the computer programs resulting from an AI approach to algorithm engineering are rather different from classical algorithms, not only when it comes to which kinds of problems they solve but also in how the program is developed and how it is structured. These differences in problems, structure, and development influence what is required to provide transparency about the implementation of an AI algorithm.

2.2 Transparency of AI Algorithms — What Should Be Visible?

A first immediate response to the demand for transparency of AI algorithms is to provide the code of the program implementing the algorithm [19], i.e., via code-sharing platforms such as GitHub, and to provide access to the frameworks, such as TensorFlow, that are used in executing the code [20]. However, even for classical algorithms, it can be very difficult to understand what problem a piece of code solves just from reading the code. Therefore, some documentation is often provided along with the code, explaining the code in more high-level terms and documenting what the code does, i.e., how it processes input. Such documentation could be a combination of pseudocode and models, such as finite-state machines [21] and decision models [22]. The conformity of the code with the documentation can either be shown by providing some kind of proof of correctness, e.g., using model checking [23] or documenting a comprehensive test, showing that the program provides the correct output on a representative set of inputs for which the correct output is known.

However, for many of the approaches to engineering AI algorithms described above, there is no standard way of documenting and verifying the implementation. One of the difficulties stems from the kind of problems they are used to solve. While a classical algorithm solves a problem for which one can precisely define the sets of valid inputs and possible outputs as well as the correct relation between input and output, an AI system usually provides a solution to a problem for which neither the set of inputs nor the correct relation between input and output can be defined. A couple of examples can illustrate how AI systems can be put to use in public governance. In the management of resources attributed to helping unemployed people back into the labour market, a municipality might use an AI system to prioritize scarce resources. For instance, the input may be the unemployed person’s health data, tax records and curriculum vitae and the output a prediction of the individual risk of long-term unemployment and a recommendation about what kind of assistance might be needed to help get that person back into work. Another example would be a public institution performing a control function in the fight against money laundering. Such an institution might use an AI system to detect suspicious transactions in the market of financial derivatives. An input here would be a series of transactions performed over a period of time and the output a recommendation to further investigate specific transactions. A third example would be a municipality that needs to prioritise notifications on children with a possible need for special social support.
measures to remedy the risk of abuse or neglect at home. An input could be missed meetings with the school, missed doctor’s appointments, missed payments, as well as the parents’ social benefits records, and an output could be a risk score. In the absence of a precise definition of both the possible inputs and the problem solved, complete proof that the AI system implements a correct solution to the problem obviously cannot be given. Instead, the correctness of the algorithm is usually stated as a test of how well the program solves the problem on a set of input–output pairs for which the solution is known, which is similar to the test of a classical algorithm. For classical algorithms, a covering set of input–output pairs can often be selected from the structure of the set of inputs and the program. It is less clear how to define a covering test for problems for which neither the possible set of inputs nor the correct solution to the problem can be precisely defined, as in the examples outlined above. For instance, there is no precise definition of when a prediction of future employment, money laundering, or child abuse or neglect is correctly identified. This has also been pointed out for the problem of classifying plants and animals in images [17].

Another challenge in documenting the implementation of AI algorithms stems from the way they are implemented. While a classical algorithm is a codification of a known recipe for solving the problem, AI algorithms are derived from a more complex kind of knowledge. For machine-learning-based approaches, the knowledge typically consists of large sets of examples of inputs for which the correct output is known, also referred to as ‘supervised learning’. Similarly, for statistical approaches, the knowledge is derived from historical pairs of input and output data that validate the statistical model. For logical, rule-based approaches used in expert systems [24], the knowledge consists of expert knowledge, e.g., on the interpretation of laws or the correct use of medicine documented by a combination of medical tests and research literature. It is well known that the selection of the data and knowledge from which the AI algorithms are derived are sources of bias. Obviously, the diversity of the team selecting expert knowledge to be used in an expert system can help reduce bias in the resulting algorithm [25].

For AI algorithms used in public governance, there is the additional problem that the data and knowledge used for creating the algorithm as well as the test data often contain personal information that cannot be made public. Transparency then has to rely on documentation for how the data used for the design and test has been produced instead of access to the underlying test data. This parallels the situation in which access to the algorithm cannot be given, e.g., due to the use of proprietary software from vendors who will not disclose the source code or documentation of algorithms, for which it is necessary to rely on documentation for the engineering process, possibly on whether appropriate standards for the development have been followed. Such documentation should also include information on the development process, including the composition and skills of the development team, which is known to be a source of bias in the selection of data, model features, and logical rules [25–28] and in the annotation of data used for training [29]. Regarding the development process, transparency is needed to be able to evaluate whether the proper risk assessment and ethical considerations have been made [29]. Regarding the team composition and skills, in addition to the issues on selection bias mentioned above, if crowd-workers are used for annotation of data for training, it is relevant to know how they were incentivized and instructed in the task [25].

2.3 Documentation Standards in Support of Transparency in AI
First steps have been taken towards providing standards for documenting machine-learning datasets [31] and AI algorithms [32]. The line of research in Hildebrandt et al. [33] has developed a methodology and tools for transparent modelling of the rules in legal regulations that can form the basis for an expert system supporting case workers. Other researchers have embarked on the challenging task of devising methods for transparency by design [34], transparency and reproducibility [20], and trustworthy AI using formal verification of the correctness of AI algorithms [23]. International standardization bodies, such as the International Standardization Organization (ISO) and the International Electrotechnical Commission (IEC), have established a joint technical committee working on standards for AI [35] and the national Danish Standardization body [36] has published so-called publicly available specifications (PAS) on transparency [37], AI-based decision-support [38], and bias [42]. It is, however, still early days for standardized approaches to transparent documentation, design, and
governance of AI systems, and they all depend on assumptions of the environment of the algorithm, such as the probabilistic distribution of the set of inputs and the risks in making wrong decisions.

Finally, we should point out the intractable problem of providing transparency in ultra-complex pre-trained machine-learning algorithms. The prime example for this is the field of natural language processing (NLP), which sits at the cutting edge of AI. NLP is the technology behind applications such as Google Search, Google Translate, Siri, Alexa, and text message automation implemented in many smartphone apps. NLP algorithms are used to provide a textual (or spoken) output based on a textual (or spoken) input. It often operates as a search-and-find mechanism (Google Search is an example of this) but can also generate text (Google Translate is an example of that). Most recently, NLP has been used to generate new text almost freely. The technology behind this are so-called pre-trained models. The recent Generative Pre-trained Transformer 3 (GPT-3) and 4 (GPT-4) are autoregressive language models that use deep-learning to produce human-like text. A transformer is a deep-learning model that adopts the mechanism of attention, differentially weighting the significance of each part of the input data. The pre-trained models have been developed on very large datasets consisting of human-generated text available in digital format. Using deep-learning techniques, models are automatically constructed to mimic the behaviour of human language. This involves the development of billions of variables in the code. GPT-3’s full version has a capacity of 175 billion machine-learning parameters and is capable of writing text that is seemingly indistinguishable from human text [28]. Since this is a pre-trained model with 175 billion variables (likely to be continuously updated), it is practically inaccessible, thereby making transparency in the classic sense (i.e., access to the model in an understandable way) unachievable. Moreover, OpenAI, the company behind GPT-3 and GPT-4, has in a recent technical report [40] announced that implementation details (including model size), hardware, training compute, dataset construction, training method, or similar are not made open, but the company plans ‘to make further technical details available to additional third parties who can advise … on how to weigh the competitive and safety considerations above against the scientific value of further transparency’.

2.4 AI Transparency in Human–Computer Interaction

The discussion has so far focused on the algorithms and their development. Prior work on transparency and algorithms has often taken a techno-centric approach, focusing on the technical or mathematical aspects of algorithms to support transparency [41, 42]. Researchers in Human–Computer Interaction (HCI) have recently stressed that a purely technical approach does not capture the role of the social context in which the algorithm is operated or targeted. Transparency can be both enhanced and weakened through the design of system interfaces. This serves as a reminder that ensuring case workers’ ability to provide transparency for citizens as well as their colleagues depends on the development and design of the information technology (IT) systems (e.g., case management systems) they use. This is the reason why they should play a crucial role in the development of such systems, especially in the case of legally trained personnel, whose involvement will ensure a consistent application of relevant rules and established practice. This adds an important perspective to the composition of teams mentioned above [13].

Within HCI, algorithmic systems that directly interact with and influence decisions about people are often conceptualized as ‘street-level algorithms’ [43], drawing on Lipsky’s seminal work of street-level bureaucrats [93], who are the human intermediates between the state and its people. Street-level algorithms, like their human counterpart, make decisions with a direct impact on their users or other stakeholders [44]. Within public administration, such algorithms have been used for a long time for automated procedures, for example, in operating payments related to welfare services (e.g., pensions, child benefits). This has not been seen as problematic because these systems simply execute decisions that have been taken elsewhere in the bureaucratic apparatus or which are suitable for automation because the eligibility criteria are simple and easily verifiable against existing data (e.g., age, nationality, registered place of residence). In recent years, however, algorithms have also found their way into more sensitive areas of public administration, albeit not as fully automated administration but rather in the form of algorithmic decision-support. Algorithmic decision-support has been used for assessing
eligibility to public welfare [45], risk-profiling in employment services [46, 47], and efforts to predict child harm [48, 49]. These are examples of algorithmic systems that affect the work of case workers in public administration and influence citizens’ interaction and experience with the public administration.

From a user perspective, transparency can be improved by providing explanations of the outcomes of the algorithm through visualizations (e.g., line charts for time series and the step-by-step process) and by providing more tailored training for domain users [50, 51]. Displaying the reasoning behind the algorithm’s actions is particularly important when the decisions or recommendations for decisions are on-the-edge cases or otherwise surprising [52]. One way to imagine this in public administration could be to provide an on-demand explanation if, for example, an algorithmic recommendation is to reject or only partially accept citizens’ applications. Explanations of outcomes are not enough to provide transparency to algorithms, but explanations should support downstream user actions such as acting on the decisions [53–55]. Hence, a transparent algorithmic system would not only explain its recommendation or decision but also advise or make it clear how case workers should act upon it, for example, overrule it, dive more into the data underlying the decision, or request support from a colleague.

3 LEGAL REGULATION: TRANSLATING TRANSPARENCY INTO LAW

In this section, we analyse how current legal regimes regulate AI transparency in public administration. We delimit our focus to those legal rules and principles that provide access points for citizens to demand information from public authorities about the AI systems at use in the administrative practice of the relevant authorities. The starting point for our analysis will be on the FOI regime. Subsequently, we will cover the data-protection framework and the proposed EU AI Act.

3.1 Freedom of Information Law: Access Rights for the Public

There is a distinction in administrative law between two different sets of rules on access to information. There is, on the one hand, a general right for everyone to access information from public authorities through FOI requests. On the other hand, there is a specific right for individuals who are party to a case decided by a public authority to gain access to case files relevant to that decision and to be provided with an explanation of why the case was decided the way it was. Here, we only deal with the first set of rules on FOI.

One of the central goals of FOI laws is to further transparency in public governance by making the apparatus of public administration visible to the public [56, 57]. Hence, it is immediately relevant to consider the role that FOI laws can play in the area of transparency in AI when such technology is used by administrative agencies. Bloch-Wehba has pointed out that the FOI model, whereby anyone can request access to government documentation [58], allows the public at large, including non-governmental organizations (NGOs) and similar actors, to act as controls on government use of AI, thus lifting the burden of challenging opacity from individuals faced with algorithmic decisions to interest groups, critical journalists, etc. [59]. FOI legislation, therefore, is a cornerstone in the relationship between ‘the people’ and their ‘government’. This is also clearly reflected in §1 of the Danish FOI Act [60]. Danish law is largely representative of a broader Scandinavian approach to FOI legislation. We will document this with frequent references to Swedish and Norwegian law throughout this section.

3.1.1 Existing Documents. FOI laws in general differ in their definitions of what exactly the public can gain access to. Some grant a right to access ‘documents’, while others provide for a right of access to ‘information’ [56, 61]. Regardless of this difference in formulation, however, the general rule is that this right of access only refers to information that the public administration is actually in possession of, i.e., in the form of pre-existing documents. Thus, in Danish law, for example, one cannot demand that a public authority create new documents as part of an FOI request [62–64]. This is also the case in most other jurisdictions [56]. A limited exception can be found in rules mandating that public bodies extract data from databases, but even this presupposes that the government has information in a database to compile [65]. This reliance on existing documentation means that, unless there are laws mandating documentation of the development of machine-learning systems or IT
systems in general [66], a public body can decline a request for information on a machine-learning system if no documents exist.

Since there is at present no standard way of documenting machine-learning algorithms (see Section 2.3), there will rarely be relevant documentation available for disclosure under the ‘only existing documents’ rule. Furthermore, even if such documentation existed, it could very well be exempted from disclosure, e.g., provided it contained proprietary information belonging to a private developer or if it had the character of an internal document.

3.1.2 Transparency Exceptions Due to Proprietary Information. Another common feature of FOI regimes is some sort of exception to access due to protection of proprietary information. As an example, §30 (2) of the Danish FOI Act exempts information on ‘technical configurations or processes or operating conditions or commercial activities, insofar as it would be of major economic importance to the company concerned that the request is not acceded to’ [67]. While not completely equivalent, similar rules can be found in, inter alia, the Norwegian and Swedish FOI laws [68].

These exemptions are obviously of relevance for determining to which extent AI transparency can be achieved through FOI laws. Private companies often provide machine-learning systems to public authorities. These companies will have a legitimate interest in protecting information on how their systems are constructed in order to make sure that such information does not reach their competitors [58]. It follows from the wording of the Danish proprietary information exemption rule, read alongside the rest of the Danish FOI law, that a public authority cannot exclude information from public access through a contract with a private partner unless the law provides for such an exclusion [62, 63]. We have not found any examples of this legal basis in Danish law [69].

Generally, then, administrative agencies are obliged to make a concrete assessment about whether to disclose a specific document or exempt it from disclosure on the grounds that it would damage the property owner if the document is made public. The determining factor in this assessment is whether it would objectively be of major economic importance to the private company delivering the AI technology if the information were revealed, a determination which is made by the public authority [70]. The same is true in Norway [71] and, given the formulation of its FOI law, Sweden as well. In other words, it is not sufficient that the company itself claims that the disclosure of the document in question would be damaging; the administrative agency must make its own independent assessment and may only withhold a document from disclosure if it considers the release of that document to cause major economic damage to the company in question. However, not all algorithmic systems used by public agencies are developed by private companies. Sometimes, public agencies themselves develop AI technology to support various case handling tasks. In those cases, it is not private ownership, but instead the so-called internal document exemption rule, that may prevent access to transparency in AI.

3.1.3 Internal Document Exemptions. All of the FOI regimes we have looked into for the purpose of this article contain an exception meant to protect the ability of a public body to discuss cases and policies internally without their deliberations being subject to public access. This exemption from FOI rights serves the important function of reserving a space for confidential deliberation within an agency, where proposals can be made without the possibility of them later being released to the public [58, 62].

In Danish, Norwegian, and Swedish law, an internal document is generally a document that has been created exclusively within and has not been circulated outside a public agency [71, 72]. If a document does not leave the agency, it is exempt from FOI requests [73]. As soon as it is sent outside the agency, either to another agency or a third party, it generally loses its privileged status [74]. This exemption might prove an issue when a technological solution is developed in-house, as any documentation about the system created in the process as a starting point would be exempt. However, the internal document exemption rule is not without exceptions. In Danish and Norwegian law, internal documents are not exempted if they contain guidelines for deciding cases, i.e., if they have the status of internal rules that impact on the determination of the rights and duties of citizens [75].

Access to information about an AI system developed internally in a public agency then depends on whether the use of such a system amounts to the implementation of an internal rule or whether the logic of the
system itself can be considered an internal rule used for decision-making purposes. To be covered by the Danish internal rule provision, a document must contain general guidelines for the handling of particular case types. These guidelines must be binding for case workers in the public agency in question [62, 76]. For instance, the Parliamentary Ombudsman has found that exam marking guides, which are binding, are covered by the exemption [77]. Other examples include guidelines for parking control, provisions regarding internal delegation of administrative competencies, and guidelines for handling applications regarding governmental financial capital contributions [78]. In light of this, it is interesting to note that Bloch-Wehba has shown how US courts have not accepted the argument that computer models, etc., can be described as deliberative and thus rejected granting an exemption under this part of the American FOI Act [58].

3.1.4 International Regulation at the European Level. A similar FOI framework can be found at the European level, which is reflective of the fact that access rights are not limited to the Nordic region alone but have evolved across Europe. For instance, Article 42 of the Charter of Fundamental Rights of the European Union provides for a general right of access to documents [79]. It is particularly relevant where the EU operates its own public agencies. However, through case-law and EU legislation, a corresponding right applies where public authorities of the member states implement EU law [80].

Beyond the scope of EU law, the Council of Europe (CoE) provides another relevant legal framework, albeit with a much broader membership of states, including the United Kingdom and Turkey. The European Court of Human Rights recognised a right to access publicly held information for journalists, watchdogs, NGOs, academics, and bloggers to exercise their freedom of expression rights under Article 10 of the European Convention on Human Rights [81]. By contrast, the Convention on Access to Official Documents of 2009, another CoE convention, which is also known as the Tromsø Convention, provides a general right to access for everyone [82].

Nonetheless, both under EU law and the CoE framework, such FOI rights are subject to similar limitations and restrictions as under national FOI laws, including in relation to existing documents as well as exemptions for propriety information and internal documents.

3.2 AI-Specific Regulation

In the previous section, we considered the general FOI framework applicable to public agencies regardless of their use of technology. The following section will specifically examine the rules and legal regimes that require transparency in relation to the use of AI.

3.2.1 Data-Protection Framework. The EU’s General Data Protection Regulation (GDPR) is perhaps the clearest example of an explicit requirement for transparency in AI. Articles 13 to 15 provide a right for data subjects affected by automated decision-making to obtain ‘meaningful information about the logic involved, as well as the significance and the envisaged consequences’ [83]. A corresponding right is provided by the CoE Convention for the Protection of Individuals with Regard to the Processing of Personal Data of 2018, also known as Convention 108+ [84], which entitles data subjects to obtain, upon request, ‘knowledge of the reasoning underlying data processing’ in case of fully automated decision-making [85]. The Article 29 Working Party on Data Protection (WP29) — the predecessor of the newly established European Data Protection Board (EDPB) — has held that ‘meaningful information about the logic involved’ in the GDPR does not necessarily mean a ‘complex explanation of the algorithms used or disclosure of the full algorithm’:

‘Instead of providing a complex mathematical explanation about how algorithms or machine-learning work, the controller should consider using clear and comprehensive ways to deliver the information to the data subject, for example: the categories of data that have been or will be used in the profiling or decision-making process; why these categories are considered pertinent; how any profile used in the automated decision-making process is built, including any statistics used in the analysis; why this profile is relevant to the automated decision-making process; and how it is used for a decision
concerning the data subject. Such information will generally be more relevant to the data subject and contribute to the transparency of the processing. Controllers may wish to consider visualisation and interactive techniques to aid algorithmic transparency.’ [86]

This interpretation implies a significant duty to document. This view was supported in a recent opinion of an EU advocate general. Even though the case concerned private credit scoring institutions, it may also have important consequences for the documentation and disclosure duties of public authorities. Interestingly, the advocate general acknowledged the importance of protecting proprietary information but stressed that at least a minimum of information must be provided, including on the factors considered in the decision-making process and their weighing at the aggregated level. It is to be seen how the Court of Justice of the EU will decide this case and whether it will follow the opinion of the advocate general [87].

3.2.2 EU Proposal on Artificial Intelligence. As is evident from the above sections, there is no comprehensive and exhaustive access to AI transparency under existing legal frameworks. This might change in the near future as a result of the upcoming adoption of the EU Act on AI. After an initial proposal from the European Commission in 2021 and amendments from the Council of Ministers in late 2022, the most recent draft regulation reflects the amendments adopted by the EU Parliament on 14 June 2023 [88]. The proposal seeks to set up a certification and registration procedure that will provide enhanced transparency into AI systems used by public institutions. In this subsection, we take a closer look at the proposed regulation in order to examine what impact it might have in terms of advancing transparency in AI.

A number of provisions are worth noticing. First, Article 52 on Transparency obligations for certain AI systems might be relevant for some public agencies. The provision obliges providers of AI systems to ensure that systems intended to interact with natural persons are designed and developed in such a way that natural persons are informed that they are interacting with an AI system unless this is obvious from the circumstances and the context of use. This would be relevant for public agencies that would — for example — use chatbots to interact with citizens. Article 52 also obliges users of emotion recognition systems or a biometric categorisation system to inform natural persons of the operation of such system when they are exposed thereto. This could be relevant — for example — for immigration services that want to verify information about asylum seekers or when operating travel restrictions under a pandemic. However, it should be noted that the article exempts certain uses of such systems, notably crime detection and prevention systems, from the transparency requirement.

Another provision in the proposed regulation is related to the use of AI systems that are categorised as high-risk AI (see Article 6 of the proposal for a definition of high-risk AI). For such systems, it is stipulated in Article 13 on transparency and provision of information that: ‘High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable providers and users to reasonably understand the system’s functioning. Appropriate transparency shall be ensured in accordance with the intended purpose of the AI system, with a view to achieving compliance with the relevant obligations of the provider and user set out in Chapter 3 of this Title [i.e, Articles 16–29].’. It is important to note here that this transparency requirement relates to the system’s ‘operation’ and that the level of transparency that this system must meet is described in relation to users. In other words, Article 13 is not aimed at those who are subject to algorithmic decisions but rather at those who operate such systems, e.g., case workers in a public agency [89].

The transparency requirement vis-à-vis users is formulated as a requirement that the system is ‘designed and developed’ in such a way that its operation is understandable to its users in light of the obligations that the user is under when using the system. Deciding whether a specific system meets the transparency requirement, then, becomes a matter of deciding whether the user can sufficiently make sense of how the system operates to use it in a responsible way. This, in turn, must be ensured through a quality management system, which providers of high-risk AI systems are required to put in place according to Article 17.
It seems that this requirement to produce documentation and instructions should make it easier for individuals to gain access to information about these systems from public agencies through the use of their general FOI rights. Yet, the proposed regulation does not alter the status of the FOI exemptions mentioned above. Consequently, even where such documentation exists, the internal document exemption as well as the proprietary information exemption still apply. Thus, depending on what information is provided, they could be invoked to refuse access to the algorithm documentation database [90].

A number of additional requirements will apply to AI systems that are labelled as high risk, and the proposed regulation introduces a certification system that obliges producers and users to carry out a compulsory conformity review. It is worth noting in this connection that this certification system includes a compulsory registration of some high-risk AI systems in an EU database [91]. It follows directly from the proposed Act that information contained in the EU database shall be ‘freely available to the public’, according to Article 60(3). Among others, the database will hold a copy of the EU declaration of conformity and information about the intended purpose of the AI system and its ‘Electronic instructions for use’ (Annex VIII, part II). The certificate and declaration mentioned above are issued on meeting certain conditions. It is unclear what specific information will be contained in these documents: whether they will simply take the form of a ‘stamp of approval’ or whether they will contain detailed information about the grounds of approval (i.e., technical information about how the system meets the various requirements for high-risk AI). We expect that there will be some delimitation regarding intellectual property and other sensitive information, as Article 70 of the proposed AI Act is intended to protect the confidentiality of such information.

4 IMPLICATIONS AND OUTLOOK ON THE FUTURE OF TRANSPARENCY IN AI

On the basis of the foregoing sections, we can conclude that there is not one sole approach to securing transparency in AI. Rather, there is a range of different avenues that can be pursued. The primary focus of the legal analysis was on whether and how the ordinary FOI framework can be relied upon. Despite notable differences across jurisdictions, it can be concluded that FOI regimes generally only grant access to existing documents. As we noted in Section 2, there exists so far no mature standard for documenting AI models. Despite some first steps in that direction, verifying conformance of the code with the corresponding documentation remains a real challenge. However, even if extensive reliable documentation of algorithmic models existed, access to the general public could be restricted on the basis of the wide proprietary interests and internal documents exemptions. It remains to be seen whether a court adjudicating FOI requests will uphold protection under those exemptions or whether they will grant access. The approach may very well differ from country to country and perhaps even among courts within the same jurisdiction.

From an advocacy and civil society perspective, the possibility to limit FOI access rights is certainly disappointing. A much more promising avenue may be provided by the data-protection framework. As outlined in Section 3.2.1, data subjects are entitled to request meaningful information about the logic involved. This implies a corresponding duty on the part of the data controller to provide information that the data subject can reasonably understand. Such information would appear to be different from the technical documentation standards discussed above and would thus not face the same technical challenges. There are, however, three potential caveats worth considering. First, this data-protection approach only applies to fully automated decision-making.

In other words, access to information about AI used in decision-support systems would have to rely on other approaches, including the EU’s future AI Act. Second, this data-protection approach seems to reverse the burden of challenging the opacity of algorithmic decisions to individual data-subjects. The clear advantage of the FOI framework was that it entitles interest groups, critical journalists and other civil society actors to lift that burden collectively. However, to pursue the same strategy in the data-protection field, they can easily reach out to affected groups and advise them to make individual access requests. This would certainly require a higher coordination effort but would also allow them to obtain information not otherwise available to them. Third, legislation could significantly limit the duty to document where public authorities are the data controllers. However, this
option is much more restricted where such bodies are involved in administrative decision-making, as it would also engage the right of individuals to access their own administrative files. This right is closely linked to the duty to give reasons in administrative decision-making and provides individuals with a much more privileged access than through the FOI framework. This is why future research into transparency in AI should consider the broader aspect of access rights rather than exclusively focusing on the FOI alone.

Also, the EU proposal on AI does not provide a comprehensive approach for all AI systems. Instead, its regulatory scope is largely confined to the use of high-risk AI. To increase, among others, transparency for such systems, the proposal provides for extensive documentation duties on the part of the providers, involving two different sets of databases. As outlined above, developing appropriate documentation and verification of such documentation is technically extremely challenging. The EU Act, if adopted, will certainly lead to further development in that field. Nevertheless, it is difficult to see whether and how this can be achieved for ultra-complex pre-trained machine-learning algorithms. Regardless of the technical challenges, the enhanced documentation requirement will invariably lead to significant costs for the providers of AI systems, which has led the Centre for Data Innovation, among others, to warn of a deteriorating business climate for that sector [92].

A significant focus of the transparency requirement is on the user of AI systems, e.g., case workers in a public agency. This is to enable them to comply with their own obligations as well as to understand and use the system appropriately. In fact, where the public agency uses a fully automated decision-making system, the information that the developer has to provide may make it easier for the agency (i.e., the data controller) to make reasonably understandable documentation available to the data subjects. Moreover, the emphasis on enhancing the understanding among case workers for the AI system corresponds to our consideration in Section 2.4 on HCI, especially in relation to decision-support systems.

From the perspective of advocacy and civil society groups, the EU proposal on AI in its current form presents a welcome opportunity to test the FOI procedures vis-à-vis the proposed databases held by the EU Commission and national authorities. In view of the widely accepted proprietary interests and internal documents exemptions, however, we do expect a certain limitation of access to documentation data. Nevertheless, the proposal explicitly calls for public access and enhanced transparency. Hence, there is a need for further specification of what information should be made available to the public through the documents kept in the databases and where exactly the balance between transparency and other protected interests should be drawn.

5 CONCLUSION

As AI becomes an integrated part of public decision-making, we need to reconcile technical, legal and socio-technical perspectives of transparency. It is therefore necessary to understand the interplay between the AI systems, the available information, the legal context, and the people affected by and working with the systems. In this article, we showed that there exist different avenues to ensure transparency in AI.

From a legal perspective, we examined the applicable FOI regimes across different jurisdictions, with a particular focus on Denmark and other Scandinavian countries. Despite notable differences, we conclude that these FOI frameworks generally only grant access to existing documents and that access can be denied on the basis of the wide proprietary interests and internal documents exemptions. As a result, we believe that the European data-protection framework and the proposed EU AI Act — with their far-reaching duties to document the functioning of AI systems — provide much more promising avenues for transparency in AI, including through FOI requests.

From a socio-technical perspective, we provided an overview of the challenges of understanding what should and can be made visible in relation to AI algorithms and how emerging documentation standards will play a role in the provision of transparency in AI. Developing appropriate documentation and verification standards is technically extremely challenging. So far, there exists no mature standard for documenting AI models. In fact, despite some first steps in that direction, verifying conformance of the code with the corresponding documentation remains challenging. It is unclear whether this can ever be achieved for ultra-complex pre-trained machine-learning algorithms. Whether the development of explainable black box machine-learning systems provides a
viable solution remains to be seen. As Rudin has argued, those explainable black boxes run the risk of providing inaccurate explanations, which may have particularly harmful effects. This is why high-stakes decision-making, as in the criminal justice and health sectors, should instead rely on interpretable models [96]. Whether developers in the industry will heed this call will largely depend on their business models and their proprietary interests.

REFERENCES


[2] In Europe, the term ‘access to information’ is more frequently used. Yet, for the sake of consistency, we will use ‘freedom of information’ or ‘FOI’ throughout this article.


[4] Sweden is an outlier since it had already recognised the principle of public access to official documents under the Freedom of the Press Act of 1766.


[59] It should be noted, though, that FOI laws in the United States have also been criticized for failing to provide transparency, because they are too abstract, thereby generally allowing too much room for manoeuvring around these rules. See especially: Mark Fenster. 2006. The opacity of transparency. Iowa Law Review 91 (March 2006), 885–949. https://scholarship.law.ufl.edu/facultypub/46


The Right to Transparency in Public Governance • 8:15


[85] Article 9. See also: Explanatory Report. https://rm.coe.int/16808ac91a


[89] In the proposed Article 3, ‘user’ or ‘deployer’ is defined as ‘… any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity’.

[90] It should also be noted that both exemptions equally apply under Article 42 EU Charter of Fundamental Rights.

[91] Article 51(1) sets out that: ‘Before placing on the market or putting into service a high-risk AI system referred to in Article 6(2) the provider or, where applicable, the authorised representative shall register that system in the EU database referred to in Article 60, in accordance with Article 60(2)’. Article 51(2) introduces a similar requirement for public users: ‘Before putting into service or using a high-risk AI system in accordance with Article 6(2), the following categories of deployers shall register the use of that AI system in the EU database referred to in Article 60: a) deployers who are public authorities or Union institutions, bodies, offices or agencies or deploying authorities on their behalf’. Article 60 sets out that: ‘The Commission shall, in collaboration with the Member States, set up and maintain a public EU database containing information referred to in paragraphs 2 and 2a concerning high-risk AI systems referred to in Article 6(2) which are registered in accordance with Article 51’.


Received 30 June 2022; revised 20 July 2023; accepted 12 October 2023