Algorithm Audit

Algoprudence: Jurisprudence for algorithms

Case-based and decentralized judgements for ethical AI

May 2024

Algoprudence: Jurisprudence for algorithms[1]

By elaborating on case-based approaches to advice on ethical issues that arise in AI systems, the concept of 'algoprudence' is introduced and explained. This new term refers to specific, case-based, and decentralized judgement regarding the responsible use of algorithms. Based on an analysis that open legal norms, for instance in EU non-discrimination law, the GDPR, the AI Act and Dutch Public Administrative Law, are insufficient to accurately regulate these algorithms, it is argued that algoprudence can complement and concretize existing legal frameworks in a practical manner.

1. Introduction

In The Netherlands, currently a lot is being done to ensure the responsible use of algorithms in the public domain. Although risk profiling and blacklists are still imprinting their mark on the image of the Dutch government to this day, steps into the right direction have been taken. These include the nationwide algorithm register for public sector organizations, the development of Fundamental Rights and Algorithms Impact Assessment (FRAIA), the government-wide Framework for Algorithms, and the new Directorate for the Coordination of Algorithms (DCA) at the Dutch Data Protection Authority. However, the development of these kinds of policy instruments, like organizational safeguards, does not automatically result in the deployment of responsible AI.[2] Nor do abstract legal standards from legislation and jurisprudence[3] that apply to the use of algorithms by public sector organizations produce the desired effect. This can partly be explained by the gap between the general frameworks and the concrete issues that play a role in algorithmic practice.

In this article, we argue that specific, case-based, and decentralized judgement on the responsible use of algorithms can contribute to the further interpretation of the applicable legal frameworks. This is an emerging *praxis* that takes place outside the scope of positive law. It can, however, indirectly contribute to the development of a clearer normative framework for algorithmic selection. We call this practice 'algoprudence'.

While algoprudence can also be applied outside the public sector, we illustrate the necessity and added value of algoprudence in the context of machine learning-driven (ML) risk profiling by executive public bodies. In section 2, we first discuss a selection of legal framework that apply to this practice, e.g., EU non-discrimination law, the GDPR, the AI Act and Dutch Public Administrative Law. Based on our evaluation that these frameworks are relevant but also insufficient to formulate flexible and concrete standards for algorithms, we argue the need for contextualization to formulate specific standards, and introduce 'algoprudence' as a mechanism to realize this. One of the first normative judgements that we consider to be algoprudence concerns the use of ML-driven risk profiling for social welfare re-examinations, which until recently was applied by the municipality of Rotterdam.[4] This case, together with the related and now stopped 'Smart

[1] This article follows at a high-level the original article (in Dutch) *Hoe 'algoprudentie' kan bijdragen aan een verantwoorde inzet van machine learning-algoritmes* as published in the journal for Dutch legal scholars NJB https://algorithmaudit.eu/nl/knowledge-platform/knowledge-base/white_paper_algoprudence/

[2] For example, on visas application procedures, see: https://www.nrc.nl/nieuws/2023/04/23/beslisambtenarenblijven-profileren-met-risicoscores-a4162837; fraud control and social welfare: https://www.platform-investico.nl/artikel/advocaten-fraudecontrole-duo-treft-vrijwel-uitsluitend-studenten-met-migratieachtergrond/ and https://nos.nl/artikel/2482915-uwv-verzamelde-illegaal-gegevens-van-uitkeringsgerechtigden.

[3] District court of The Hague, 5 February 2020, ECLI:NL:RBDHA:2020:865 (SyRI).

[4] Risk profiling for social welfare re-examination (ALGP:AA:2023:02), consisting of an advice report (ALGP:AA:2023:02:A) and a problem statement (ALGP:AA:2023:02:P). https://algorithmaudit.eu/algoprudence/cases/aa202302_risk-profiling-for-social-welfa-

check sustenance' algorithm of the municipality of Amsterdam, is discussed in paragraph 3 to demonstrate that novel questions arising from algorithmic practice are insufficiently addressed by abstract and open norms in existing legal frameworks. Then, in section 4, by means of the case studies we illustrate how algoprudence can contribute to the responsible use of algorithms. We conclude with a more general reflection on algoprudence as a new concept in the legal landscape (section 5) and a conclusion (section 6).

2. Evaluation of open legal standards for ML-driven risk profiling

In this section, wet introduce the practice of machine learning-driven (ML) risk profiling by Dutch administrative public sector bodies and the applicable normative framework of four pieces of legislation: EU non-discrimination law, the GDPR, the AI Act and Dutch Public Administrative Law. We argue that these frameworks are relevant to the algorithmic practice, but require contextualization, for which algoprudence is a promising mechanism.

2.1 How does ML-driven risk profiling work?

It is widely known that Dutch governmental organizations use ML for risk profiling. For instance, municipalities can use ML to select social welfare recipients for re-examination[5]. In the analogue (non-algorithmic) variant, civil servants manually define criteria for risk profiles each year, such as 'single men with a roommate', complemented by random sampling and event-driven selection. In addition to these analogous profiling methods, ML can be used to select criteria for a risk profile. If correctly embedded in risk management measures (such as organizational check and balances, ML validation, and documentation requirements), this algorithmic-driven selection process increases effectiveness and reduces the risk of discrimination and arbitrariness – so the promise suggests. In practice, these risk management measures prove difficult to implement. It is usually considered a challenge to thoroughly document and validate used ML methods. And even if the documentation is more or less correctly maintained, a list of difficult questions remains. Which variables are fed to the variable selection algorithm and why? What type of ML method is used and why? And according to which performance metrics is the algorithm monitored and evaluated (including metrics to measure fairness and effectiveness)? These are urgent matters with a strong normative connotation to which no silver bullet answers exist. As we illustrate in the case discussions, this concerns value trade-offs that are inherently linked to technical aspects of data modelling. In the following subsections, we analyze the role that (a selection of) legal frameworks play in shaping standards for ML-driven risk profiling.

2.2 EU non-discrimination law

EU non-discrimination law assesses risk profiling on discriminatory practices in a structured manner.[6]

1. **Does unequal treatment occur in comparison to others in a similar situation?**
   Yes, this is the case for ML-driven risk profiling

2. **Is unequal treatment based on the basis of etnicity (e.g., skin color, origin, national origin) or nationality?**
   No, in most cases this is not the case for ML-driven risk profiling

re-reexamination/

[5] Sections 53a and 64 of the Participation Act provide a legal basis for re-examination.

[6] See also Annex IV of Discrimination through risk profiles, Netherlands Institute of Human Rights https://open.overheid.nl/documenten/ronl-c409ea31-2c00-4318-9a45-d47ad8a2ca7f/pdf

3. **Does direct differentiation occur on the basis of etnicity?**
   No, in most cases this is not the case for ML-driven risk profiling

So, the open system of justification applies. Next it questions whether an objective justification exist for differentiation?

4. **Is etnicity or nationality the only selection criterium for the risk profile?**
   No, in most cases this is not the case for ML-driven risk profiling

5. **Is the risk profile targeted to only one specific etnicity or nationality?**
   No, in most cases this is not the case for ML-driven risk profiling

6. **Does the risk profile contain a selection criterium through which direct differentiation on the basis of etnicity or nationality occurs?**
   No, in most cases this is not the case for ML-driven risk profiling

So, we deal with indirect differentiation.

7. **Does the risk profile pursues a legitimate aim?**
   Yes, there is a legal basis for applying risk profiling by Dutch public sector organisations. See also section 2.3 GDPR and the Dutch Participation Law Art. 53, 64.

8. **Is this specific risk profile well-suited for the pursued aim? (selection criteria are relevant and objective)**
   Open question. It depends.

9. **Is this specific risk profile necessary and proportional? (Reasonable balance between the involved interests, does not exceed necessary involvement, no reasonable alternatives)**
   Open question. It depends.

No silver bullet answers for question 8 and 9 exists. It is unclear how to weigh inclusion or exclusion of specific selection criteria in (ML-driven) risk profiling. For many cases no relevant jurisprudence exists.

2.3 General Data Protection Regulation

The General Data Protection Regulation (GDPR) regulates the collection and processing of Dutch social welfare recipients. In the context of ML-driven risk profiling for social welfare re-examination article 6 paragraph 1 sub e of the GDPR provides the legal basis for data processing, based on the necessity for the performance of a task carried out in the public interest. This task arises from the Dutch Participation Act (Article 53a, 64) and the Act on the Structuring of the Implementation Organization for Work and Income (in Dutch: Suwi, Article 62).

Given the lawfulness of processing, two open formulated GDPR provisions relevant for ML-driven risk profiling are stated below:

> **Article 13(2)f, 14(2)g** and **15(1)h** state that the data subject has the right to obtain “meaningful information about the logic involved” pertaining to profiling. How the logic involved in, for instance boosting-based ML ensemble methods, should be explained in natural language to data subjects remains unclear;

> **Article 22** relates to automated individual decision-making, including profiling. Civil servants take the final decision whether social welfare allowances are (un)duly granted based on an in-person re-examination interview. ML-driven risk profiling methods therefore

serve as a sampling method and are therefore not considered as fully automated decision-making and are therefore not regulated by this article. Considering the recent Schufa[7] ruling it is however an open question whether this reasoning is still valid. Besides, it has become an open question whether the Participation Law provides sufficient clear specification of criteria that can be used for risk profiling.

2.4 AI Act

The AI Act imposes broad new responsibilities to control risks from AI systems without at the same time laying down specific standards they are expected to meet. For instance:

> **Conformity assessment (Art. 43)** – The proposed route for internal control relies too much on the self-reflective capacities of producers to assess ML-based risk profiling's quality management, risk management and bias tests. Resulting in subjective self-assessment;

> **Risk- and quality management systems (Art. 9 and 17)** – Requirements set out for risk management systems and quality management systems remain too generic. For example, it does not provide precise guidelines how to perform sensitivity analysis to balance FP/FNs for ML-based risk profiling;

> **Normative standards** – Technical standards alone, as requested the European Commission to standardization bodies CEN-CENELEC, are not enough to realize AI harmonization across the EU. Publicly available technical and normative jugements about fair AI at code-level are urgently needed.

2.5 Dutch Public Administrative Law

Decision-making processes by Dutch municipalities are subjected to the Dutch Public Administrative Law. This legal framework regulates how governmental bodies, including municipalities, can exercise public power. Three important principles are mentioned:

> **Article 2:4** – Principle of fair play, among others stating that public sector boedies should carry out tasks without bias in ML-based risk profiling;

> **Article 3:2** – The duty of care, among others stating that a situation must be created in which all interest can be weighed and in which a suitable ML method is used;

> **Article 3:46-3:47** – The duty to give reasons, among others stating that it should be explained how ML produced an outcome that contributed to a decision.

Unifying these principles with the ML-based variable selection for risk profiling is a challenge. On itself, algorithmic selection of variables is not a decision as defined in Awb Article 1:3, as an civil servant of the municipality takes the formal decision whether social welfare is (un)duly granted after a re-examination interview is conducted. However, variable selection could be seen as part of the duty of care, i.e., careful preparation of this decision. Difficulties in explaining why certain criteria are included in a risk profile can result in a municipality acting 'lawfully' but not 'appropriately'. Additional organizational and legal requirements on how ML-based profiling methods can align with these principles are an open and context-dependent question. Algoprudence aims to contribute to an answer to the idenftified open questions.[8]

[7] https://curia.europa.eu/juris/document/document.jsf?text=&docid=280426&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=91113

[8] For a detailed article in Dutch on applying principles codified in Dutch Public Administrative Law can be found here: https://algorithmaudit.eu/nl/knowledge-platform/knowledge-base/njb-artikel/

2.6 The need for contextualization and the role of algoprudence

The above analysis shows the need for contextualization to define specific applicable standards for ML-driven risk profiling. The contextualization of the above four legal frameworks, and the possible recalibration that may follow, will not happen by itself. At the same time, it proves to be difficult for courts, the legislator, and the regulator to initiate this process. Courts are certainly interested in defining concrete standards for algorithmic decision-making further, as showed by the well-known case law on SyRI and the Aerius application.[9] However, the court is dependent on the (so far rare) concrete cases that are submitted, and if so, non-discrimination law plays a relatively small role in for instance proxy variable selection. In the case of Dutch administrative courts, there is also the fact that a ruling about an algorithm will always be part of a broader judgment on legality and will therefore not be able to provide answers to all concrete questions that are relevant for ML algorithms that are used in practice. It is not the task of the legislator to design very detailed standards; for example, the AI Act delegates further specification to harmonized standards. The fact that, in the context of the internet consultation on the legislative proposal for the strengthening of Dutch Public Administrative Law, a separate 'Reflection document on algorithmic decision-making' exists shows that the legislator is also struggling with the issue of standardization.[10] In the past, regulatory bodies, such as the Dutch Data Protection Authority, have indicated to explain standards that are common in various legal frameworks, such as usage of sensitive data attributes for bias testing, but so far there has been little evidence of this. The institutional impasse surrounding the concretization of legal standards for algorithmic practice must be overcome. We argue that algoprudence can play a significant role in realizing this.

This contribution introduces the concept of 'algoprudence' for concrete, case-based, and decentralized judgement about the responsible use of algorithms. At its core, algoprudence is an ongoing conversation between various actors in society about the resolution of normative issues that arise in the use of algorithmic applications, based on specific judgments about a case. Algoprudence can contribute to the transparent and effective implementation of open legal standards, and can harmonize judgements about algorithmic applications. Similar to 'legisprudence' (the colloquial jargon for the collection of legislative advice from the Advisory Division of the Dutch Council of State) and 'ombudsprudence' (which provides insights into the principles and working methods of the Ombudspersons and the team of professionals who support them), the concept differs from 'case law' in important respects. In addition to its non-binding character, in the case of algoprudence the decentralized and non-hierarchical characteristics are important examples. Judgments are not left up to formal institutions, but to more or less official societal bodies that may or may not have a formal status and are positioned closer to the algorithmic practice. In order to further develop abstract legal standards for algorithms, such as the principles of good administration, it is essential that judgements can be made on issues even without any legal proceedings having to be initiated.

At the same time, the claim to the 'prudential' nature of this nascent practice is essential. If the organizations and committees that have to decide on the normative aspects of algorithms were to approach them as issues that can be solved pragmatically, which only results in '*best practices*', an essential element would be lost, namely

[9] Supra note 3 and ABRvS 17 May 2017, ECLI:NL:RVS:2017:1259 (Aerius).; See also Wolswinkel 2020.

[10] https://www.internetconsultatie.nl/algoritmischebesluitvormingenawb/b1.

a deliberative and motivated assessment in which the *rightness* of a judgement is explicitly thematized. The concrete application of principles, such as the duty of care and fairness, requires an interpretive community. In the case of ML-driven risk profiling, it is primarily highly interdisciplinary and, secondly, still in full development. At first glance, the introduction of 'algoprudence' may therefore seem somewhat premature. However, we believe that this framing can actually make a positive contribution to professionalizing and streamlining the efforts of this community.

Before we conceptualize and legally embed the nascent practice of algoprudence, we aim to make the above analysis of the above legal frameworks more concrete through two ML-driven risk profiling case studies, after which the potential of algoprudence is demonstrated.

3. Case study: legal frameworks in the practice of ML-driven risk profiling

In this section, we discuss two recent cases of ML-driven risk profiling by Dutch municipalities, which serve to illustrate both the relevance and the current impotence of the above discussed legal frameworks and institutional actors to rein in ML-driven risk profiling in practice.

3.1 Amsterdam and Rotterdam ML-driven risk profiling

One case concerns to the municipality of Rotterdam between 2017 and 2021, the other has recently been at issue in the City of Amsterdam. They differ in several crucial respects: the type of ML used, the phase of implementation in which they are applied, and the way in which the development process of the algorithm is documented. Because of these differences, the cases offer an adequate impression of the developing practice.

In 2022, the municipality of Amsterdam introduced the pilot 'Smart check sustenance'.[11] This algorithm assigns a research-worthiness score to each new application for social welfare allowances. The score is determined by using an *explainable boosting model* (ebm) algorithm that is trained on past applications and associated data about the living situation of a citizen, previous social welfare applications, income, and assets of the citizen at the time of application. Inquiries that have been assigned a score above a certain threshold are further examined by an employee. With the use of such an algorithm in the application phase, the municipality of Amsterdam wants to shift the emphasis from re-examination to more accurate allocation of allowances. After evaluating the pilot, it was decided in early 2024 to quit using the algorithm.[12]

The special '*explainable*' character of the ebm-algorithm stems from the method in which the research-worthiness score is determined. By applying complex statistics, the model first calculates a score for each characteristic (living situation, income, etc.), after which the scores are summed up to form a final score. This generative additive nature of the ebm method means that the importance of each characteristic on the final score is tracked.

In Rotterdam, an *extreme gradient boosting* (xgb) algorithm was used (and stopped after controversy) to predict risk scores for unduly granted social welfare allowances for citizens who already receive such allowances. This xgb-algorithm has been trained to find patterns between more than 60

[11] Documentation about this algorithm can be found in the Amsterdam Algorithm Register, see: https://algoritmeregister.amsterdam.nl/ai-system/onderzoekswaardigheid-slimme-check-levensonderhoud/

[12] See final evaluation of the pilot 'Smart check sustenance' (TKN8) https://amsterdam.raadsinformatie.nl/vergadering/1203734/Raadscommissie%20Sociaal%2C%20Economische%20zaken%20en%20Democratisering%2014-02-2024

features of social welfare recipients and unduly granted allowances. Based on risk scores, citizens were selected for re-examination. For an xgb-algorithm, it is more complex than for an ebm-algorithm to keep track of feature importance for the predicted risk score. This is because the risk score is not determined by adding scores per feature, but rather is computed at once in a complex statistical calculation in which all characteristics are combined. Only in complex statistical terms it can be traced how a certain characteristic has contributed to the predicted risk score.[13] This *black box* character also applies to the ebm-algorithm, but only relates to how a score is determined per feature. In short, both methods are a *black box*, although the xgb-algorithm has a stronger black box character than the ebm-algorithm, which makes it less explainable.

3.2 Explainable, risk-averse and fair ML-driven risk profiling

How doe the ebm- and xgb-profiling methods fit in the selected legal frameworks? It is important to note that in both cases, ML is not applied directly to decide about whether social welfare allowances are (un)duly granted, but only contribute to preparing such a decision, which is ultimately made by civil servants after an interview or desk research.

3.2.1 Explainability

Our discussion of the above four legal frameworks are particularly relevant to assess the explainability of the ML algorithm that contributed to a certain outcome. As mentioned above, the ebm-algorithm is more explainable than the xgb-algorithm. However, both algorithms are still to a certain extent a black box. Decisions of selecting an individual with help of an ebm- or xgb-algorithm can only be motivated by the municipality in statistical terms, which can already be opaque to experts, let alone to the citizen who wants to appeal a decision. It is therefore highly questionable how explainable the ebm algorithm actually is. The desire to 'express calculations made by an algorithm in natural language'[14] meets the boundaries of the statistical reality.

3.2.2 Risk management

As we have seen, the duty of care as imposed by Dutch Public Administrative Law requires awareness regarding relevant facts and interests and are weighed with help of an appropriate method. For ML applications, this raises questions about the completeness and correctness of the training data, as well as the suitability of the used ML method. First of all, risk management measures are essential here, for which several guidelines have been published.[15]

However, despite the presence of guidelines and risk management measures, classifying an ML application as a 'suitable method' remains a difficult task. Fundamental rights-oriented guidelines, for instance the Fundamental Right Algorithms Impact Asessment (FRAIA), the 'Principles for (semi-) automated decision-making' of The Netherlands Institute for Human Rights and CEN-CENELEC's Risk Management standard (under construction) are quite abstract and procedural in nature.[16] They prescribe in what manner organizations can identify the impact of algorithms on human rights in order

[13] The LIME and SHAP values, which are popular among statisticians, sort insufficient effect. See, for example, D. Vale, A. El-Sharif & M. Ali, 'Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law', AI Ethics 2022 2, p. 815–826.

[14] Supra note 17.

[15] See, for example, the Algorithm Research Framework of the Dutch Government Audit Agency (2023) https://www.rijksoverheid.nl/documenten/ reports/2023/07/11/onderzoekskader-algoritmes-adr-2023 and the Algorithm Assessment Framework of the Netherlands Court of Audit (2021) https://www.rekenkamer.nl/onderwerpen/algoritmes-digitaal-toetsingskader.

[16] Impact Assessment Human Rights and Algorithms (IAMA), 31 July 2021, https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes; The Netherlands Institute for Human Rights, 'Principles for (semi-)automated decision-making', 9 February 2021, https://publicaties.mensenrechten.nl/publicatie/1980e51e-bb12-4bb1-8a9b-26c7a3aa2b86.

to be able to make an evaluative assessment. The normative resolution of such an assessment is not provided within such soft law frameworks. There is a good reason for this, because judgements are always context-dependent and cannot be determined in material sense by general frameworks. Whether, and if so, which form of ML in casu is the appropriate method for weighing facts and interests remains therefore an open question.

The Rotterdam case demonstrates that there are issues for which there are relatively clear solutions: the training dataset resulted not to be representative, which violated the requirement that facts must be fully known and weighted properly.[17] But the interpretation of risk management is relatively vague with regard to other issues that emerge from the case, such as the question regarding which of the 60 available characteristics are eligible to serve as input for a risk model. Is the number of children or the assertiveness measured by a civil servant eligible as a selection criteria, or not? Should the predictive value of a criterion be taken into account in assessing this eligibility? Or should a qualitative assessment be undertaken of intrinsically (un)suitable criteria independent of the predictive value of a feature? Neither the question about selection of a particular ML method is clear. In the Rotterdam case, the xgb-algorithm was chosen from a handful of alternatives, because this type of algorithm emerged as the most effective method in the test phase.[18] Further justification for the choice of this method is absent.

The Amsterdam case exemplifies a heightened level of risk management in the algorithm's development. The public algorithm register states that the ebm-algorithm was selected due to its explainable nature.[19] In addition, it is explained which variables were (not) fed to the algorithm and why, and a bias test was performed. But why ML-driven risk profiling was chosen in the first place, and why alternatively explainable algorithms (such as rule-based algorithms) were not considered, remains unclear. Here, too, questions relating to the core principle of risk management remain unanswered.

3.2.3 Bias testing

Above, we have noted that from the principle of fair play, if applied to the context of ML-driven decision-making, may result in an obligation to prevent algorithmic bias.[20] At the same time non-discrimination law emphasizes the importance of proportionality, necessity and suitability of ML methods. This issue is relevant to the Rotterdam case, as it has been shown that the training data was not representative regarding young citizens, which allowed the model to develop a bias. The principle of fair play, risk management measures and non-discrimination standards require that such biases in datasets have to be monitored and mitigated. Apart from the quality of the dataset, the model can also develop bias by differentiating upon apparently neutral characteristics, which strongly correlate with protected grounds, known as the problem of indirect discrimination through proxy variables. In both cases, this issue arises regarding features used for the algorithmic risk profiling. The Rotterdam algorithm, considers features, such as literacy rate or ZIP code, that are strongly correlated with migration background.[21] But from

[17] See https://www.lighthousereports.com/suspicion-machines-methodology/.

[18] Freedom of information request from VPRO Argos/Lighthouse Reports, 2017020 Privacy Impact Assessment pilot phase Project Benefit Unlawfulness, https://www.vpro.nl/dam/jcr:c87f2d6c-3f9c-4498-9a9c- f3bc5483a437/Downloads%20Model%20Rotterdam.zip.

[19] Supra note 26.

[20] When determining bias in the algorithmic-driven selection process, the alternative, for example bias in a manual selection process, should also be investigated. See https://www.parool.nl/columns-opinie/opinie-onderzoek-vooringenomenheid-van-zowel-algoritme-als-ambtenaar~bd69aa5e/

[21] See also section 4-5-4 of Coloured Technology, Rotterdam Court of Audit 2021, https://rekenkamer.rotterdam.nl/onderzoeken/

a statistical point of view, all possible features correlate to protected grounds to some extent. So, only gradual differences exist. No silver bullet to resolve the proxy and correlation challenge exists.

In addition to the issue of (proxy) discrimination, non-discrimination principles covered by fundamental rights in various legal frameworks raises a broader complication which we also discussed as a part of risk management: what forms of differentiation are acceptable for an ML-algorithm? Is it fair to profile citizens on the basis of their professional appearance, filled in by a civil servant based on a contact moment? Both the subjective nature of such a finding and the possible unfairness of distinguishing on the basis of personal characteristics raises questions about this practice. But also in this case, an unequivocal standard is absent.

This case discussion of municipal risk profiling shows on the one hand that this type of ML applications results in complications and raises issues that legal frameworks should solve or prevent. On the other hand, it appears that providing concrete interpretation of legal provisions seems to quickly encounter its limits. With the rise of ML algorithms, a spotlight is put on the suitability of a method in a general sense. Simultaneously, the general question about suitability of ML as a method (which includes explainability and fairness) is always context-dependent and cannot be properly prescribed by a general framework. To define open legal standards such as proportionality, suitability and necessity additional case-based normative judgements are needed. In our view, algoprudence is a promising mechanism for addressing these needs.

4. Algoprudence demonstrated in practice

What is lacking to ensure responsible use of algorithmic risk profiling are standards that are both flexible and precise. We introduce algoprudence as a manner to define such standards when using ML algorithms. Before we further embed this new concept in relevant legal frameworks, we first illustrate the potential of algoprudence through the above discussed Rotterdam case. In particular, we focus on the following aspects for which existing frameworks do not provide appropriate answers:

> The suitability of ML as a risk profiling method in social welfare re-examination, compared to alternatives, such as manual (expert-driven) profiling or random sampling;

> Transparency- and explainability requirements for ML-driven risk profiling;

> Determine which variables are considered (in) eligible to be used for risk profiling, among other with respect to the risk on proxy discrimination.

Algorithm Audit recently issued an advice report on these normative issues surrounding the Rotterdam case.[22] This advice is explicitly intended as a contribution to algoprudence. The input of algoprudence consists of a problem statement, in which the normative issues are described given the relevant institutional, legal, ethical and technical context, and an advice document created in response to a deliberative assessment by an independent committee of experts and stakeholders.

To illustrate, we will provide a brief outline how this algoprudence can contribute to resolving the above issues. First, the report (partially) answers the above questions:

> It states that, under certain strict conditions,

algoritmes/.

[22] Risk profiling social welfare re-examination (AA:2023:02), 2023, https://algorithmaudit.eu/nl/algoprudence/#risk-profiling-social-welfare.

algorithmic risk profiling can be used responsibly in the context of social welfare re-examination. Parallel use of multiple selection methods (algorithmic and manual profiling, as well as random sampling) is considered desirable.

> With respect to explainability requirements, the used xgb-algorithm by the municipality of Rotterdam was deemed to be an unsuitable method. It presents a standards what qualifies as a sufficiently explainable algorithm.

> To help determine the (in)eligibility of profiling criteria, a list is provided of responsible and irresponsible variables, along with corresponding rationales (see Figure 1).

In the advice report, these concrete norms are embedded in an evaluation of the institutional and social context of the ML algorithm and of social welfare re-examination. The judgement and resulting norms are case-specific and context-dependent, which guarantees its normative flexibility. Simultaneously, the judgment can be generalized to similar contexts, which would include, for example, the case of the 'Smart check sustenance' of the municipality of Amsterdam. The algoprudence that has been created can therefore contribute productively to the use of responsible algorithms. If another municipality is considering the use of ML-driven risk profiling, it can learn from the algoprudential assessment of the Rotterdam case, and subsequently apply it to *mutatis mutandis* to its own context; something we can assert based anecdotal examples is already happening. In this way, not every Dutch municipality (there are 340 of them) has to reinvent the wheel, but can orient itself on an existing and well-motivated judgment. In this way, shared yet flexible standards emerge that harmonize the use of algorithms in a given context; *in casu* ML risk profiling in the context of municipal benefits.

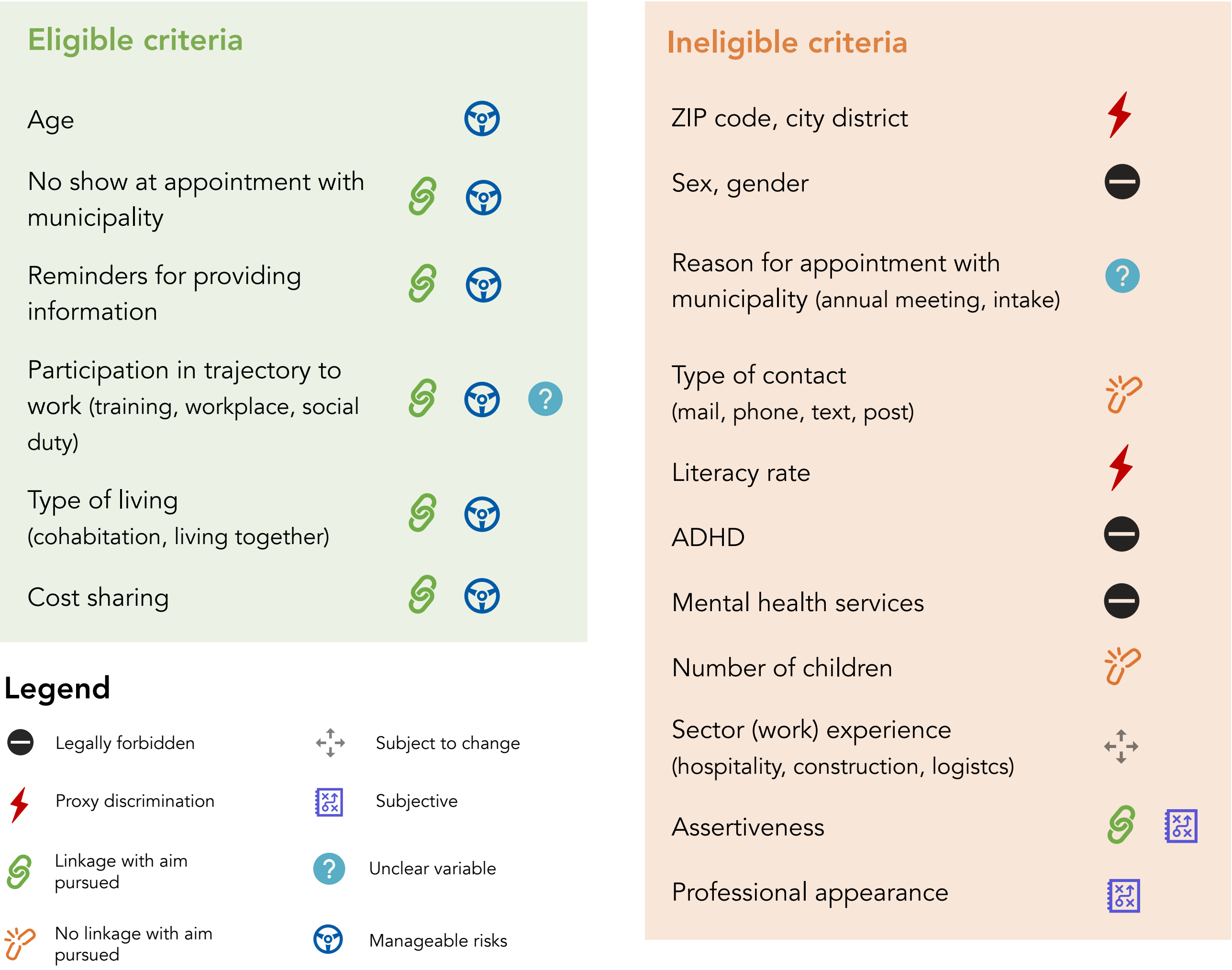
5. Algoprudence legally embedded

In the foregoing, we have demonstrated the modus operandi and added value of algoprudence through practical examples. Lastly, we anticipate on further development of this new concept and elaborate on how it can be embedded in the existing legal landscape as a complementary instrument.

5.1 Algoprudence as a *praxis*

We introduce the concept of 'algoprudence' for the practice of specific, case-based, and decentralized judgement about the responsible use of algorithms. How algoprudence functions as praxis and in what manner exactly the algoprudential corpus is created are too broad questions to be fully discussed here. The answer should certainly not depend solely on the first cases in this area by Algorithm Audit. In essence, it should be an ongoing discussion between various stakeholders in society, based on case studies about the specific resolution of normative issues that arise in the use of algorithmic applications. Algoprudence does not have to be limited to the domain of public law, but could apply to all spheres (private and public) where normative questions emerge about the application of algorithms that are not answered by technical-pragmatic solutions or legal frameworks. The foundation of algoprudence is formed by the transparent publication of case-based judgments, which consist of an explanation of the normative issue, its context along with a motivated judgement. In theory, algoprudence, like jurisprudence, can give rise to annotations of judgments, which further develops collective judgement in a transparent and deliberative manner.

How this form of decentralized judgment takes shape and who is allowed to act as the judging organisation are still to be determined. Full decentralization is a possibility, in which everyone is allowed to contribute, even if decisions of



Eligible criteria	
Age	⎈
No show at appointment with municipality	🔗 ⎈
Reminders for providing information	🔗 ⎈
Participation in trajectory to work (training, workplace, social duty)	🔗 ⎈ ?
Type of living (cohabitation, living together)	🔗 ⎈
Cost sharing	🔗 ⎈

Ineligible criteria	
ZIP code, city district	⚡
Sex, gender	⛔
Reason for appointment with municipality (annual meeting, intake)	?
Type of contact (mail, phone, text, post)	⛓‍💥
Literacy rate	⚡
ADHD	⛔
Mental health services	⛔
Number of children	⛓‍💥
Sector (work) experience (hospitality, construction, logistcs)	✥
Assertiveness	🔗 ▦
Professional appearance	▦

Legend

⛔	Legally forbidden	✥	Subject to change
⚡	Proxy discrimination	▦	Subjective
🔗	Linkage with aim pursued	?	Unclear variable
⛓‍💥	No linkage with aim pursued	⎈	Manageable risks

Figure 1 – Example of algoprudence: overview of variables of which a normative advice commission has judged whether they are (in)eligible as a selection criterion for (ML-driven) risk profiling in the context of social welfare re-examination.

certain bodies will be more important than others. Limited decentralization is another possibility, where predefined and recognized institutes that have the authority to make such rulings (for instance ethical advice boards alongside regulators and other formal institutions). Depending on the form of decentralization it would require coordinating organizations and a certain degree of standardization of what counts as algoprudence, and how it is published. Over time, algoprudence can follow the example of the ECLI numbering and adopt an internationally standardized coding.[23]

The question of which preconditions must apply to algoprudence remains open. Judgements do not necessarily have to be made in the way that Algorithm Audit advocates, namely a deliberative judgement formed by commission consisting of academic experts, stakeholders, and affected groups, although this specific approach does offer advantages. Regardless of how the method is institutionalized, the basis for a judgement will always depend on the diligence and appropriateness of the followed procedure.

[23] For example, Stichting Algorithm Audit uses the coding 2023:AA:02 where the terms refer to the year, the organization and the case number, and the problem statement (2023:AA:02:P) or the advice document (2023:AA:02:A). Thanks to Martijn Staal for the idea. As a prefix ALGP: is suggested.

5.2 Algoprudence versus alternative instruments

A prominent question regarding algoprudence is the status and legitimacy of the resulting judgements. It should be clear that algoprudence, as a decentralized process for contextualizing standards, cannot have a directly binding character. A logical objection is whether it would not be better to codify additional standards for the usage of ML algorithms within existing enforceable laws and regulations. While the regulatory framework could certainly be more specific, codification of many of the algoprudential judgements would not be appropriate. To give an example, a legal ban on the use of xgb-algorithms would not fit well into the system of administrative law, nor into the risk-based approach of the AI Act, and is over-rigid. After all, there are imaginable scenarios where the trade-off favors this technique, or in which further technical development overcomes the problems of explainability.

Even though algoprudential judgements are not binding, they can still sort effect in various ways. First, a self-regulatory effect will emerge if AI professionals are aware of consensus in the algoprudence in a certain field. In that case, they must have good reasons to deviate from the judgements. Secondly, this effect can be reinforced by legal frameworks, which can, for example, through technical documentation requirements, mandate an organization to justify their ML validation, in which the *state-of-the-art* must be taken into account. Thirdly, algoprudential jugements can serve as input for positive legal interpretation of legal frameworks. If algoprudence is available for a particular issue, a judge will be more inclined to attach legal consequences to it, rather letting public sector bodies touch in the dark. We distinguish a fourth way in which algoprudence can sort effect: through political decision-making. Algoprudence can be used in the political arena to question and critically reflect upon public and private sector organisations on the basis of an independent jugement.[24]

Standards for (high-risk) ML algorithms will also result from the AI Act. The AI Act is based on the regulatory model of product safety. The conformity assessment of high-risk AI systems will therefore, as with other EU product legislation, be carried out on the basis of harmonized standards (such as CEN and ISO standards). However, the AI Act will not eliminate the need for algoprudence. Primarily, it follows from the fact that the AI Act is about 'product safety' that the harmonized standards will be predominantly technical in nature. Insofar standards touch upon fundamental rights, it will be rather procedural. Second, as we have shown in the cases above, the concrete evaluative trade-offs that must be made are highly context-dependent and thus cannot be settled with a generic harmonized standard. Thirdly, harmonized standards are not public: they are often made available by private organizations and are beyond a pay wall.[25] In contrast to algoprudence, harmonized standards, can therefore never fulfil the role of building public knowledge and transparent collective judgement.

6. Conclusion

We argue that current legal frameworks, such as EU non-discrimination law, the GDPR, AI Act and Dutch Public Administrative Law lack standards that are specific enough to adequately address particular issues related to the application of ML-based risk profiling. As our discussion of the Rotterdam and Amsterdam cases shows, there is a need for further specification to comply with legal requirements, such as explainability, risk management and bias testing. With the introduction of the concept of algoprudence, we propose an additional instrument to fill in the

[24] See, for example, questions from the Amsterdam city council about the ebm-algorithm, https://amsterdam.raadsinformatie.nl/document/13573898/1/236+sv+Aslami%2C+IJmker+en+Garmy+inzake+toegepaste+profileringscriteria+gemeentelijke+algoritmes.

[25] Even though a recent ruling by the Court of Justice of the European Union could potentially change this: CJEU March 5, 2024, ECLI:EU:C:2024:201.

normative gap for algorithmic applications through specific, case-based, and decentralized judgement. Algoprudence holds the promise that organizations that use algorithmic applications will be provided with concrete standards on how to deal with specific issues for which there is no technical-pragmatic, nor a univocal legal solution. In this article, we argue that algoprudential judgements have its own independent position alongside other tools and resources. The development of algoprudence obviously needs to mature. By introducing algoprudence as a concept, we hope to give an impulse to the actual development of algoprudence, as well as an initial impulse to its legal embedding.

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. The goals of the NGO are three-fold:

Normative advice commissions

Forming diverse, independent normative advice commissions that advise on ethical issues emerging in real world use cases, resulting over time in algoprudence

Technical tools

Implementing and testing technical tools for bias detection and mitigation, e.g, bias detection tool, synthetic data generation

Knowledge platform

Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, see for instance our AI Policy Observatory and position papers

Structural partners of Algorithm Audit

SIDN Fund

The SIDN Fund stands for a strong internet for all. The Fund invests in bold projects with added societal value that contribute to a strong internet, strong internet users, or that focus on the internet's significance for public values and society.

European AI&Society Fund

The European AI&Society Fund supports organisations from entire Europe that shape human and society centered AI policy. The Fund is a collaboration of 14 European and American philantropic organisations.

Dutch Ministy of the Interior and Kingdom Relations

The Dutch Ministry of the Interior is committed to a solid democratic constitutional state, supported by decisive public management. The ministry promotes modern and tech-savvy digital public administrations and govermental organization that citizens can trust.

Building *public knowledge*

on responsible AI *not-for-profit*

🌐 www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

✉ info@algorithmaudit.eu

Algorithm Audit

Stichting Algorithm Audit is registered as a non-profit organisation at
the Dutch Chambre of Commerce under license number 83979212