Understanding the Affordances and Constraints of Explainable AI in Safety-Critical Contexts: A Case Study in Dutch Social Welfare

Aleksander Buszydlik(✉), Patrick Altmeyer, Roel Dobbe, and Cynthia C. S. Liem

Delft University of Technology, Delft, The Netherlands
A.J.Buszydlik@student.tudelft.nl,
{P.Altmeyer,R.I.J.Dobbe,C.C.S.Liem}@tudelft.nl

Abstract. We focus on explainability as a desideratum for automated decision-making systems, rather than only models. Although the explainable artificial intelligence (XAI) paradigm offers an impressive variety of solutions to increase the transparency of automated decisions, XAI contributions rarely account for the complete systems—social and institutional environments—where models operate. Our work focuses on one such system in the domain of social welfare, which increasingly turns to automated decision-making to carry out targeted digital surveillance. Specifically, we present a case study of a black-box machine learning model previously used in a major Dutch city to support its officials in the task of detecting fraud. Employing analyses established in the field of system safety, we identify five types of hazards that could have occurred after the introduction of the model. For each of them, we reason about the potential value of XAI interventions as hazard mitigation strategies. The case study illustrates how the deployment of models may impact processes that exist far upstream and downstream from their decision logic, making explainability and/or interpretability insufficient to guarantee the systems' safe operation. In many cases, XAI techniques may only be able to reasonably address a small fraction of hazards related to the use of algorithms; several major hazards that we identify would have still posed risks if the system had relied on an interpretable model. Thus, we empirically demonstrate that the values, which lie at the heart of XAI research, such as responsibility, safety, or transparency, ultimately necessitate a broader outlook on automated decision-making systems.

Keywords: Explainable artificial intelligence · System safety · Automated decision-making · Social welfare · Technology audits

1 Introduction

Automated decision-making (ADM)—the use of algorithms to make fully automated decisions or support the decisions of humans [6]—is increasingly applied to

S. Hofmann et al. (Eds.): ePart 2025, LNCS 15978, pp. 118–136, 2026.
https://doi.org/10.1007/978-3-032-02515-9_8

improve “efficiency” or “objectivity” of decisions in highly consequential domains such as the justice system [1], public administration [22], or healthcare [36]. Although regulatory frameworks, including the European Union’s *General Data Protection Regulation* (GDPR) and the *Artificial Intelligence Act* (AI Act), or (until recently) the United States’ *Blueprint for an AI Bill of Rights*, propose to ban fully-automated decisions and require mechanisms for human oversight, it is not yet clear what form these should take. Moreover, as observed by [46] in the context of GDPR, policy efforts seek to establish the *“right to explanation”* rather than the *“right to interpretable decisions”*. This distinction is even clearer in newer regulations, such as the AI Act, which does not outlaw black-box decision-making, instead mandating sufficient explanation standards for automated decisions. Therefore, we can reasonably expect post-hoc explainability solutions to become standard, preferred tools in many domains.

However, post-hoc interventions or human oversight procedures were not enough to prevent high-profile failures of ADM solutions. The Netherlands alone has observed the detrimental impacts of several ADM systems, most notably ones developed to detect social welfare fraud. Firstly, an algorithm employed by the Tax and Customs Administration (*Belastingdienst*) resulted in unjustified fraud accusations against tens of thousands of families, driving many households (deeper) into poverty in the childcare benefits scandal (*toeslagenaffaire*) [24]. Secondly, the System Risk Indication (*SyRI*) model established by the Ministry of Social Affairs and Employment to detect potential social welfare fraud at the level of neighborhoods [4] was outlawed by The Hague District Court on the grounds that it violated the provisions of Article 8 of the European Convention on Human Rights [41]. Thirdly, the Dutch Education Executive Service (*DUO*), in charge of student finance grants, used an investigation model that unfairly targeted students from non-European backgrounds [25]. Lastly, the municipality of Rotterdam experimented with a risk profiling model that was found to exhibit biases against vulnerable groups such as migrants or single mothers [19].

These cases share three notable characteristics. First, the common goal of detecting fraud in social welfare settings highlights the keen interest of public authorities in making use of algorithms as a surveillance tool. Second, all people who rely on social assistance are in a vulnerable position; algorithms tend to aggravate these social inequalities by disproportionately targeting certain subpopulations. Third, the models were small components of larger *sociotechnical decision-making systems.* In theory, flawed (e.g., discriminatory) algorithmic predictions could have been filtered out by actors responsible for the operation of these models. However, this did not occur and real-world harms materialized, suggesting the need for a broader lens to describe, understand, and address such *“algorithmic” harms* [14]. Even though methods have been proposed to conceptualize and analyze AI functions in their sociotechnical contexts [15,44], their applications in social welfare and broader public services are still missing.

The abovelisted failures of ADM systems offer valuable lessons to researchers, practitioners, and policy-makers. Thus, we contribute the first in-depth analysis of a social welfare fraud detection system through a sociotechnical and system-theoretic lens. We consider the automated fraud detection tool previously used in Rotterdam, looking at the pertinent technical, social, and institutional processes to arrive at a comprehensive description and understanding of possible harms. We pursue desk research (study of existing documents) to inform an empirical analysis of the case and make use of the tools of system safety to identify core harms and their underlying system-theoretic hazard sources [28]. Using these insights, we perform a critical analysis of the practical value of XAI solutions in this complex system. In the process, we provide a first-of-its-kind evaluation of the affordances and fundamental limitations of XAI solutions, grounded in a sociotechnical and system-theoretic analysis of an ADM system, within the specific context of social welfare. Our approach showcases how relying on XAI tools is fundamentally incomplete and cannot guarantee their responsible operation. We explain how other forms of interventions are necessary to safeguard vulnerable populations from discrimination when algorithms are employed in high-stakes settings, and point to ways in which similar sociotechnical system-theoretic analyses can aid in identifying and designing interventions to comprehensively mitigate and eliminate harms. Our contributions are as follows:

(a) **we provide a first in-depth case study** of a real-world machine learning model for risk profiling in social welfare, conceptualized as a system of social, technical, and institutional components and processes;
(b) within the context of this system, **we evaluate the affordances and fundamental limitations of XAI**, understood as a set of solutions to promote fair, reliable, safe, transparent, and so forth, automated decision-making.

2 Background

We begin by delineating the landscape of explainability tools in Sect. 2.1. Then, in Sect. 2.2, we discuss AI solutions as systems with not only technical, but also social and institutional components. Finally, in Sect. 2.3, we look at the application of algorithms in broader public administration.

2.1 On Explainability in Artificial Intelligence

"Interpretability" and *"explainability"* tend to be used interchangeably in artificial intelligence research, but we follow the distinction from a literature review of [32]. The authors explain that interpretability is generally understood as an inherent property of a particular model. In contrast, explainability refers to post-hoc solutions that aim to improve the legibility of decision logic across a variety of models. We center our discussions on explainability to remain faithful to the case study, which looks at a specific system where any attempts to explain model decisions would necessarily need to take the form of post-hoc interventions.

Explainability solutions can be applied to models that would otherwise be completely opaque, but they still attract many well-founded objections. We highlight three particularly relevant critiques. First, they are *"likely to perpetuate* *bad practice"* [46]. Opaque models offer reasonable performance out-of-the-box; coupled with an XAI method, they may give an impression of a well-designed algorithm, detracting from critical design decisions, such as selecting informative features. Second, an explanation may seem superficially reasonable but insufficiently faithful to the logic of the model [5], and thus mislead its recipients and promote unwarranted trust [5,16]. Third, explanations should account for contextual factors, such as normative and legal constraints [47], social environment and cognitive processes [33], or users' technical proficiency [13]. Thus, receiving an explanation, receiving a meaningful explanation, and understanding the meaning of an explanation are three different but intrinsically linked challenges.

Various criteria to classify explainability techniques have been proposed in the literature; an unsystematic review from 2022 found 11 taxonomies developed in the three prior years [49]. Other than the *stage* (ante-hoc interpretability vs post-hoc explainability) distinction, the author identified two further criteria considered in all taxonomies: XAI techniques range in *scope* from local (i.e., explaining individual instances) to global (i.e., explaining the complete decision logic of the model), and in *applicability* from model-specific (i.e., suitable for one algorithm class) to completely model-agnostic (i.e., suitable for any algorithm). We adapt the taxonomy proposed in Sect. 4.1 of [49], which classifies XAI solutions by the type of a problem, the modality of input data, the manner of acting upon a model, the stage, the scope, and the output format. Operating on the level of a taxonomy allows us to comprehensively look at XAI by mixing the characteristics of solutions, rather than discussing individual techniques.

2.2 On System-Oriented Approaches to Artificial Intelligence

The concept of artificial intelligence as a tool inherently entangled with its social and institutional environment is not new, but many lines of XAI research are yet to (fully) embrace the sociotechnical understanding of AI. Works that already do, draw on a variety of disciplines, including science and technology studies, ethics and philosophy of technology, or system safety [27]. A key aspect of this vision of AI is its explicit recognition that the behaviors of any model are the product of the interactions of the components in their environment [44]. As an example, generative AI poses the *risk* of producing misleading or deceptive information, which turns into a *hazard* only if, e.g., the model is accessible publicly, and in turn produces *harms* if, e.g., people act on these incorrect outputs [55].

This paradigm promises a way to overcome the blind spots of model-oriented analyses. For example, in the context of AI fairness, [48] defined five traps of abstracting away the context of a model, such as the *"ripple effect trap"*, where researchers fail to recognize that introducing new technologies into an existing system is bound to influence its dynamics. Additionally, it promotes the operationalization of abstract principles and values such as *safety* [14], which is much more difficult to guarantee with "alternative" approaches such as AI ethics [35].

2.3 On Automated Decisions in Public Administration

We look at explainable AI in social welfare, a service of public administration. Although there is little agreement among scholars on what exactly constitutes public administration, its primary objective is widely recognized as the realization of the goals of a government through the implementation of its policies [23]. The applications of ADM tools in public administration are hardly a new phenomenon, but their burden on society keeps growing along with their pervasiveness [20,30]. For example, the Netherlands has been algorithmizing its national and local governments for decades (e.g., with rule-based systems) but only recent advances in AI have raised concerns about the impacts of automated decision-making tools on fundamental freedoms and human rights [40].

Many challenges for ADM systems have been described in the literature, including safety (unexpected dynamics in operational environments), security (risks of attacks and abuse), privacy (retrieval of information stored in the training data or parameters of a model), fairness (equal treatment of individuals), or explainability [9]. As public administration entities operate on sensitive data to decide on a large scale on consequential problems, failing to sufficiently respect and defend these values when automated decision-making systems are in use may provoke questions about the government legitimacy, e.g., [22,45].

Yet, as recognized by [52], all forms of data-driven modeling may discriminate against marginalized groups. For instance, digital surveillance methods (*cf.* risk profiling) have been observed to predominantly affect vulnerable populations [17]. Thus, ADM systems may aggravate power asymmetries inherent to public administration: they create distance between *"those who shape a system"* and *"those affected by a system"* [31], with few accountability mechanisms allowing the latter to achieve meaningful control over algorithmic outcomes [31,38]. At least in principle, XAI solutions could help address this problem.

3 Methods

Our case study focuses on a machine learning model that was developed to support the municipality of Rotterdam in the implementation of the regulations laid out in the Participation Act (*Participatiewet*), which defines *bijstand*, a form of long-term assistance for residents of the Netherlands who are able to work but cannot find employment, and are no longer covered by the national unemployment insurance. According to the law, all municipalities must periodically, but no more frequently than once every two years, re-examine the recipients of *bijstand* to ensure that benefits remain duly granted. As this form of assistance is implemented at the level of individual municipalities, they may employ various tools to nominate residents for re-examination. Historically, these included random and expert-driven selections [3], as well as tips from other residents. According to *Statistics Netherlands* around five million people in the country receive some form of social support, including *bijstand* [10]; specifically in Rotterdam, 6.1% of residents receive *bijstand*, the highest percentage in the country. It is perhaps unsurprising that the decision-makers of Rotterdam, a municipality that actively invests in practically oriented research on responsible digitization, most notably through its *Creating010* research centre (e.g., [37]), looked towards AI as an approach to enable targeted selection for re-examinations.

To conduct the case study, we employ desk research methods to distill information from publicly accessible sources, and carry out system-theoretic analyses to reason about the hazards—unsafe system states—related to the model. To our knowledge, the fraud detection model employed by the Work and Income (W&I) department was first disclosed in the "Colored Technology" (*Gekleurde Technologie*) report from 2021 by the Rotterdam Court of Audit, which examined the ethical use of algorithms by the city authorities, following principles such as responsibility, transparency, or fairness [42]. The report concluded that on the organizational side, the W&I department failed to satisfy multiple ethical standards, including the lack of clearly assigned responsibility for the complete system, unsatisfactory transparency towards benefit recipients affected by the model decisions, and inadequate motivation for decisions related to ethics [42].

Later, in 2023, the model became the centerpiece of a journalistic investigation coordinated by Lighthouse Reports that focused on *"suspicion machines"*, or fraud detection tools employed across Europe for social welfare purposes [19]. As acknowledged by the journalists, *"out of dozens of cities we contacted, [Rotterdam] was the only one willing to share the code behind its algorithm"*. Their work not only revealed that the model was characterized by poor technical performance, but also that it was prone to discrimination against people from vulnerable backgrounds, recommending them for re-examination significantly more often than expected [7]. This behavior likely stemmed from the fact that expert-driven selection was used to collect the training data. Hence, some groups of residents may have been overrepresented in the dataset. Although the model was used to make predictions outside of the lab—in 2019 it was used to launch 22% of fraud investigations [42]—it never officially left the pilot stage; Rotterdam decommissioned the model in 2021, two years prior to the investigation.

We look at this model for three reasons. First, as informed by The Algorithm Register [34], the Dutch national database of algorithms, there is notable interest among public administration organizations to use automated data processing for social welfare. We find that 80 out of 674 algorithms (11.8%) registered by the end of 2024 fulfill this criterion and many of them—28, or (35.0%)—are used for fraud detection purposes. Such algorithms impact millions of people, also outside of the Netherlands [19], highlighting the need to improve their safe operations and motivating our study. Second, the Rotterdam model was part of a complex decision-making system, allowing us to discuss the value of XAI solutions in the context of real-world social and institutional processes. Third, the available documents enable us to map the complete system, thus enabling comprehensive system-theoretic analyses. In particular, Lighthouse Reports published *"the holy trinity of algorithmic accountability: the training data, the model file and the code for [the] system"* in their GitHub repository [7]. We can also learn from the entry about the model in The Algorithm Register, reports of (1) the Rotterdam

Court of Audit that discussed the development and operations of the system [42], and (2) Algorithm Audit, a normative advice organization, that looked at the quality of features used by the model [3], as well as national and local laws. What distinguishes our contribution is that we evaluate the complete system as it could have been if various explainable AI interventions (in their most optimistic, strongest interpretation) had served as risk mitigation strategies.

We rely on the tools of the Systems-Theoretic Accident Model and Processes (STAMP) framework established in the field of system safety, as proposed by [28] and operationalized for AI settings by [44]. STAMP aims to mitigate accidents by enforcing (behavioral) constraints on the system. Specifically, we apply a form of Safety-Theoretic Process Analysis (STPA) [29], one of the STAMP methods that aims to collect *"information about how behavioral safety constraints, which are derived from system hazards, can be violated"*, see p. 212 of [28].

4 Case Study of Automated Risk Profiling

We draw the boundaries of the system, and hence our analysis around the operating process: the set of components and activities immediately relevant to the daily operations. We identify them in Sect. 4.1 and map their relationships in Sect. 4.2. This operational lens allows us to reason in Sect. 4.3 about causal scenarios that could potentially drive the system to an unsafe state, based on four types of inadequate control actions distinguished by [28] and established in the field of system safety: (1) ones that are never provided or never followed, (2) ones that are provided at the incorrect time, (3) ones that are executed for an incorrect amount of time, and (4) ones that are unsafe. We analyze in Sect. 4.4 if any XAI interventions could eliminate or mitigate these hazardous states.

4.1 Stakeholder Analysis

We start by introducing key stakeholders with influence on the daily operations of the system, playing a role in its design, development, and use (i.e., "actors"):

(a) **Benefit recipients** in Rotterdam would expect a re-examination roughly once every six years, if the selection was fully random.
(b) **Consultants** are the client-facing professionals in the W&I department. They are responsible for interviewing benefit recipients and storing information about them [42], and thus co-created the dataset used by the model.
(c) **Team Testing and Monitoring** (*T&T*) provided domain expertise during the development of the model. They were also co-responsible for its operation, focusing on the selection of people for re-examination [42].
(d) **Team Reinvestigations** (*THO*) is responsible for carrying out the re-examinations of benefit recipients [42] based on the inputs from Team T&T.
(e) **Team Research and Business Intelligence** (*OBI*) is a research team of the municipality that joined the Accenture team for model development. Later, it took over for the maintenance and operation of the system [42].

(f) **Team Complaints** is a unit in the W&I department responsible for handling complaints about welfare benefits, e.g., from customers who have not been treated fairly or have not received sufficient information.

(g) **Privacy Officers *and* Data Protection Officer** oversee the handling and protection of personal data. While the former are employed in one of the departments, the latter is an independent officer of the municipality that focuses, among others, on compliance with the GDPR.

(h) **Concern Management** of W&I, one of six organizational units in the municipality [39], oversees all matters related to the department, including ADM systems. However, the model was developed with the support of another department, so the assignment of responsibility was not clear [42]. It further includes three independent units; among these, **Concern Auditing** that performs audits of systems, processes, and programs.

(i) **College of Mayor and Aldermen** is the executive board of a municipality selected by the city council. Its tasks include the establishment of the structure of the municipal organization (also the W&I department) and the nomination of the Concern Management. In Rotterdam, the Alderman responsible for W&I coordinated the schedule (i.e., the number of model-driven nominations) of re-examinations with the department [42].

4.2 Operating Process

Our evaluation focuses on the state of the system when the model was in use. The analysis revolves around the *controlled process*, i.e., some activity of a system that is actively and/or passively guided by a set of *controllers* [29], in this case, human employees and an AI subsystem. The controllers are expected to ascertain the state of the system through feedback channels and, when necessary, enforce behavioral constraints over the system through control channels [29], thus realizing or preventing certain outcomes in the controlled process. We present a map of the system in Fig. 1 and discuss it in the next paragraphs.

Controlled Process: Social Assistance Duly Granted. Assistance should be provided to people (1) who need it, (2) who are eligible, (3) at the time when they are eligible, (4) and at the amounts for which they are eligible. The goals are two-fold. Firstly, people who are not (or are no longer) eligible should be re-examined without delay. Secondly, people who remain eligible should not have their privileges unnecessarily suspended during the re-examination. Naturally, the latter depends on the administrative procedures of re-examinations themselves, rather than the procedures of nomination, but we note that minimizing the number of false positive nominations would address this issue.

Human Controller: Employees. Three teams in Rotterdam W&I jointly fulfill the role of the human controller: Consultants, Team T&T, and Team THO. We do not consider Team OBI to control the process—they act as an extension of the W&I teams—but their activities still have an impact on the system.

Automated Controller: AI Subsystem. In principle, the automated controller should impact the controlled process only indirectly because its predic-

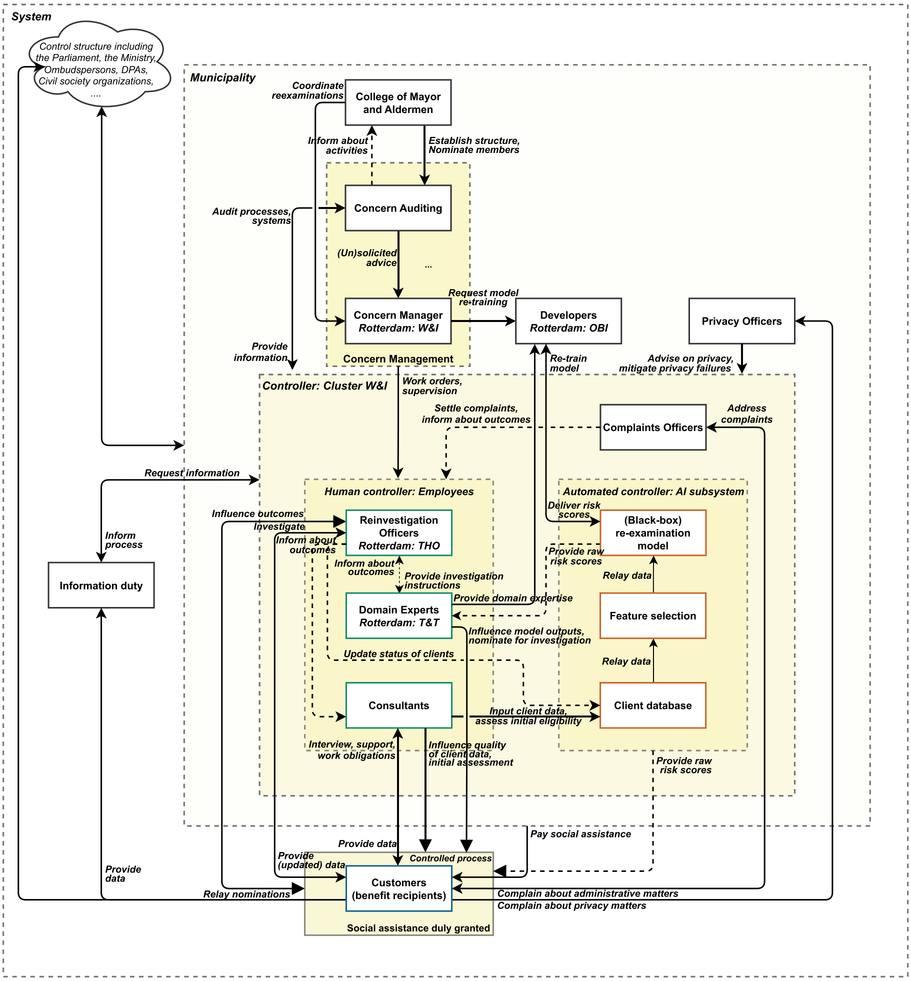
System
Control structure including the Parliament, the Ministry, Ombudspersons, DPAs, Civil society organizations, ...
Municipality
Coordinate reexaminations
College of Mayor and Aldermen
Inform about activities
Establish structure, Nominate members
Audit processes, systems
Concern Auditing
(Un)solicited advice
...
Request model re-training
Concern Manager Rotterdam: W&I
Developers Rotterdam: OBI
Privacy Officers
Provide information
Concern Management
Re-train model
Advise on privacy, mitigate privacy failures
Controller: Cluster W&I
Work orders, supervision
Settle complaints, inform about outcomes
Complaints Officers
Address complaints
Request information
Human controller: Employees
Automated controller: AI subsystem
Influence outcomes
Investigate
Reinvestigation Officers Rotterdam: THO
Deliver risk scores
(Black-box) re-examination model
Inform process
Inform about outcomes
Inform about outcomes
Provide raw risk scores
Information duty
Provide investigation instructions
Relay data
Domain Experts Rotterdam: T&T
Provide domain expertise
Feature selection
Influence model outputs, nominate for investigation
Update status of clients
Relay data
Consultants
Input client data, assess initial eligibility
Client database
Interview, support, work obligations
Influence quality of client data, initial assessment
Provide raw risk scores
Pay social assistance
Provide data
Provide data
Provide (updated) data
Controlled process
Customers (benefit recipients)
Relay nominations
Complain about administrative matters
Complain about privacy matters
Social assistance duly granted

Fig. 1. Functional control diagram of the operating process in the Rotterdam case. Relationships marked by solid lines are established based on [7,39,42]; relationships marked by dashed lines are assumed. We focus on the "daily" operations, omitting parts of the hierarchical control structure; these are highlighted in the thought bubble. The responsibility of an actor is stated on their side of a relationship (arrow).

tions are supposed to be filtered by humans. However, the analysis of Lighthouse Reports unveiled that model predictions exhibited biases [7] but the city still carried out a large number of re-examinations based on the risk scores assigned by the model. This suggests that the AI subsystem had non-negligible direct impacts. The W&I department developed a gradient boosting machine model [18] trained on 315 features. Thus, the model predictions could not be readily interpreted, both due to the opaque algorithm and the complexity of the underlying data. The features belong to 14 categories, including appointments with the municipality, barriers to reintegration into the workplace, personal characteristics, or relationships. Rotterdam's data scientists calculated the relative importance of all features compared to the baseline of *age at investigation*. The top 5 features include (1) *the age at investigation*, (2) *the number of existing cost-sharer relationships*, (3) *the number of no-show appointments*, (4) *the number of "applying expertise" competencies*, and (5) *the number of contacts on the topic of income* [7].

4.3 Scenarios of Inadequate Control

We use the functional control diagram in Fig. 1 to identify actions that may influence the controlled process and trace their potential causal scenarios, which allows us to diagnose five hypothetical scenarios of inadequate control. For each of them, we describe what types of hazards it could produce, how this influences the system, and derive requirements for XAI that ought to be satisfied to (ideally) eliminate or at least mitigate a hazard. As a methodological contribution, we propose to translate between these requirements and the technical promises of XAI solutions by formulating the former using categories from XAI taxonomies (we follow [49]). In our case study, certain requirements must hold for all scenarios: we require *post-hoc* algorithms applicable to *tree-based models* trained on *tabular inputs* for a *classification* task. Other requirements are scenario-specific.

A. Insufficient data quality (*Unsafe control actions provided*)

Description: Two types of unsafe control actions can be identified: (1) the use of uninformative or harmful features, and (2) the collection of substandard or biased data. The "garbage-in, garbage-out" principle is at play here: non-meaningful inputs are likely to produce non-meaningful outputs because the model learns spurious correlations rather than constructive patterns in the data.

Impacts: A normative advice commission established by [2] proposed a set of eligible and non-eligible profiling criteria in the social welfare setting; the framework includes eight criteria for variables, such as clear *"linkage with aim pursued"*, *"subjective"* evaluation, or propensity to produce *"proxy discrimination"*.

Requirements for XAI: Improve (e.g., address bias) *features* on a *global* scale.

B. Erroneous interpretation of outcomes. (*Control actions never provided*)
Description: (1) domain experts are expected to intervene on the outcomes of the model, but it can be reasonably expected that some interventions cannot be enforced due to the model's complexity; (2) the model does not provide any explanation of its decisions, which may make objections impossible.

Impacts: Employees may be able to apply simple interventions, e.g., exempting (groups of) people from the re-examination. More complex actions, such as identifying and acting on model biases, are difficult without an understanding of the decision logic and model statistics. Additionally, a non-interpretable model does not contribute to organizational knowledge, e.g., by suggesting fraud patterns.

Requirements for XAI: Elucidate the *decision logic of the model* on both *local* (as justification for benefit recipients) and *global* (as explanation for the decision-makers) scales. The former constrains the output format to explanations that are *legible for non-expert users* (e.g., natural language rules or descriptions).

C. Weak transparency standards. (*Control actions never provided*)

Description: Benefit recipients were not informed if an investigation was started following a model-driven nomination, see p. 42 in [42]. It is unclear if the reinvestigation officers were aware of the applied selection tools.

Impacts: Transparency wrt. the selection is necessary for the reinvestigation officers to identify wrongful nominations and evaluate the model, and for the benefit recipients to exercise their rights. Sufficient transparency standards towards data subjects are required by GDPR Art. 13(2)f, 14(2)g, and 15(1)h.

Requirements for XAI: Inform recipients (on a *global* scale, with *individual* explanations) about the reason why a fraud investigation was launched.

D. Unreliable handling of complaints. (*Control actions at incorrect time*)

Description: Benefit recipients may be reluctant to file complaints. Introducing a model further increases complexity in this step: it is technically challenging (for non-experts) to argue about the unfairness of model-driven nominations.

Impacts: Social welfare benefits are necessary for the fulfillment of recipients' basic needs, which discourages complaints. Therefore, control signals may be delayed relative to the moment when harms begin to occur.

Requirements for XAI: Ensure that individuals (*local* scale) unhappy with an automated decision have access to meaningful interventions to achieve more favorable outcomes. For XAI solutions, this would entail (1) output formats that alleviate the burden of complaint procedures (e.g., *rules*), or (2) result types that enable the recipients to control the process (e.g., *feature relevance*). Especially in case (2), these insights should be *actionable* and *provably improve outcomes*. Additionally, citizens will need (human) assistance, in particular to comprehend complex ADM-driven decisions, and understand and exercise their rights.

E. Imprecise model feedback. (*Unsafe control actions are provided*)

Description: A negative outcome of an investigation—"no fraud"—should increase trust in a benefit recipient. Repeated and excessive re-examinations not only can be considered prejudiced, but also entail ineffective use of resources. Yet, some benefit recipients were examined multiple times while the model was in use [8]. Signals on the trustworthiness of individuals (both "fraud" and "no fraud" outcomes) may also follow from random and expert-driven nominations.

Impacts: Without external interventions, a sufficiently "well-trained" model can be expected to target the same individuals multiple times when their features remain relatively stable. Thus, some benefit recipients may be unfairly targeted even after an earlier investigation established no wrongdoing.

Requirements for XAI: Safeguard all individuals (i.e., *global* scale) from unnecessary repetitious investigations, especially when the features and the decision logic do not change substantially between nomination periods.

4.4 Affordances and Constraints of XAI in Social Welfare

Finally, we reason about the ability of various XAI solutions to mitigate or eliminate the identified hazards by satisfying derived requirements. In our analysis, we consider the strongest, most optimistic form of XAI solutions to identify what types of hazards could *never* be addressed with these types of interventions.

A. Insufficient data quality

Given that the model already exists, we are looking for post-hoc interventions, but the requirements must be satisfied for all input data, which directly contradicts our goals. Still, the W&I department periodically retrained its model so we can also consider ex ante interventions. Feature selection has been a consideration in XAI research (e.g., [57]), but the quality of input data in the case study is primarily an administrative problem. For example, the top-50 most relevant features include a number of criteria that have been argued by [2] to be inappropriate for profiling; problems such as *"subjectivity"* or *"linkage with the aim pursued"* are domain-specific and cannot be resolved in a fully automated manner. Similarly, expert advice is generally needed to decide whether a feature could be considered a legitimate proxy for protected attributes. Even though Rotterdam developed tools to evaluate its model on disparate impact metrics, these were never actively applied by the operators [7]. Moreover, the model exhibited biases against groups described by multiple features. The "single mother" archetype analyzed by [7] integrates ∼5 features; even simple 2-feature archetypes in a dataset of 315 features already lead to at least 49455 possible combinations of features. Thus, we believe that XAI tools cannot sufficiently address the challenges of data quality, but they may encourage decision-makers to reflect on the inputs.

B. Erroneous interpretation of outcomes

This problem is arguably most directly relevant to the interests of XAI research as it relates to the explanation of the decision logic. Several solutions could meet the needs of decision makers. First, they could simplify the model with global surrogates, but this approach is counterproductive in that an explanation can never be completely faithful to the original model [46], bringing into question the utility of a non-explainable model in the first place. Second, they could evaluate the interactions of features with techniques such as partial dependence plots [18]; however, with 315 features, they are again only able to consider a small subset of all potential interactions. Third, they could focus on the instances with the strongest influence of the model (e.g., prototypes and criticisms [26]). Although such approaches do not explain the decision logic, they may uncover instances that require special attention from model operators, such as vulnerable groups. These three options cover all forms of results described in the taxonomy of [49].

We also look at explainability towards benefit recipients. XAI research typically points to counterfactual explanations (CEs) as a way to help non-experts understand the grounds of decisions and act upon them [53,54]. However, when we evaluate the utility of CEs for the Rotterdam model, we observe, e.g., that the algorithm of [54] produces highly impractical CEs that impact up to 70 features.

C. Weak transparency standards

This scenario relates to an organizational process, so XAI tools are not relevant. However, transparency about selection methods is a *necessary but not sufficient* prerequisite to explainability, e.g., for people to request information about the involved logic. A sociotechnical understanding of the broader process is required to ensure observability [28,43] in order to analyze this task and its outcomes.

D. Unreliable handling of complaints

Although the handling of complaints is again an organizational process, XAI interventions may contribute to mitigating the hazards. In particular, we look to algorithmic recourse (AR), the provision of actionable recommendations [50,53], as an idea that could satisfy the requirements. AR could encourage and empower benefit recipients to control the process because certain behaviors that affect risk scores can be reasonably expected from them (such as attending appointments or fulfilling information duty). It could also serve as an indirect encouragement to file complaints. Notably, these qualities would remain true if the decisions of an interpretable model were provided along with recourse recommendations.

E. Imprecise model feedback

The requirements for this setting do not fit neatly into the taxonomy of [49] because they generally require an intervention on the model rather than an explanation of its decisions, which already suggests that XAI solutions may be unhelpful here. Still, we reason about this scenario more broadly to decide on the value of some related approaches. First, the municipality could further decouple predictions from decisions [51] by introducing an alternative, stricter classification threshold for individuals who have been previously re-evaluated, which would discourage the repeated nomination of benefit recipients. Still, moving thresholds does not safeguard individuals whose features spuriously lead to high risk scores, and a more transparent administrative solution would be simpler and preferred for accountability reasons. Second, the municipality could introduce small random perturbations into the dataset to counteract the problem of stable input features leading to the same outputs. However, unless explicitly framed as adversarial examples [21], random noise would not predictably impact the predictions. The adversarial examples approach is counterproductive as it entails a situation where decision-makers are attempting to fool their own model. Furthermore, both solutions increase complexity in this scenario, defeating the purpose of an explainable approach. Thus, ultimately, XAI solutions cannot meaningfully address this challenge and the interventions must occur at different stages of the process. For example, decision-makers could evaluate the frequency of investigations as part of human-in-the-loop filtering (which was in place in Rotterdam). Alternatively, as repeated investigation may remain meaningful if different factors contribute to the two decisions, the municipality could retrain its model on different subsets of features every year. Although feature selection was part of the pipeline, we were unable to establish to what extent the decision-makers modified the features year-to-year.

5 Discussion

We begin by discussing the implications of the Rotterdam case study for research on explainable artificial intelligence as a safety mechanism in Sect. 5.1. Next, we address the generalizability of our findings to other decision-making contexts in Sect. 5.2. Finally, we address the shortcomings of our work in Sect. 5.3.

5.1 Explainable AI as a Safety Mechanism

In light of the "right to explanation" legislation, we believe that research on XAI solutions remains important. However, our work highlights that explainability may have far-reaching *fundamental* limitations when considered to address reasonable issues understood as safety hazards. We explicitly discourage reading this work as a discussion of the "current shortcomings of explainable artificial intelligence" or as an overview of the "challenges for the future". Rather, our goal was to demonstrate that certain hazards related to the deployment of automated decision-making tools in safety-critical domains can *never* be mitigated with XAI because hazards can emerge within stages of decision-making processes that are not directly influenced by the logic of a (black-box) model.

We identified five scenarios related to the use of a risk profiling model that could produce hazards. Even in the most optimistic interpretation—with faithful, meaningful, and so forth explanations—XAI could reasonably mitigate only two of them. The other hazards call for (non-technical) interventions elsewhere in the process. Inherently interpretable models cannot be expected to fully eliminate the identified hazards either. Once more, this issue arises because Rotterdam relied on 315 features, making the predictions of even the simplest logistic regression models inscrutable, especially to non-expert benefit recipients.

Our case study empirically demonstrates the need to shift *"from explainable algorithms to explainable processes"* and *"from an instrumental to an institutional approach"* previously recognized by [13]. The authors formulated seven general strategies for policy-makers to improve the acceptance of XAI methods; we postulate that these extend to practitioners who are ultimately responsible for the operations of any system and to researchers who should be aware of the affordances and limitations of their contributions. This work is a step towards operationalizing the strategies of [13] through an effective method to reason about the value of XAI interventions in (and tailor them to) specific domains.

5.2 Explainable AI in Other Contexts

Naturally, our findings are context-specific: potential hazards are functions of the actors, technologies, and processes involved in a system. We explain why many hazards related to ADM systems can *never* be reasonably mitigated with technical XAI interventions, but reliably generalizing these findings requires, e.g., repeated analyses in different decision-making systems. Still, our work showcases the applicability of system-theoretic process analysis (STPA) to reason about the potential hazards, the value of XAI solutions, and their limitations in a granular and systematic manner. Promoting these values makes our approach an effective complement to and extension of simpler (in application) auditing tools such as the Oracle Test [11]. Future work may take up these implications to inform sociotechnical and system-theoretic approaches for integrating XAI techniques.

Even in the specific context of social welfare, the ability to generalize our findings depends on various characteristics of public administration entities, such as their willingness to automate decision-making, their approaches to responsible development of technologies, or their prior experience with algorithms. Notably, public administration in the Netherlands tends to be efficient, stable, and low-conflict. Municipalities that actively invest in the development of ADM systems (such as Rotterdam or Amsterdam) tend to introduce safeguards on the process and embrace accountability [56]. Thus, the Dutch landscape of public administration offers, in many ways, a highly favorable setting for the introduction of novel algorithmic solutions. Yet, as showcased by our case study and other failures introduced in Sect. 1, even in this optimal "testing ground", algorithms can act to the detriment of vulnerable populations and undermine the authority of public administration entities. Therefore, this evaluation of the fundamental limitations of techno-centric XAI solutions can be considered as a cautionary tale for decision-makers interested in the development and deployment of ADM tools, especially when their solutions would be applied in less "predictable" systems.

Finally, we recognize that tools of system safety are themselves not neutral [14] because any map of a system imposes a subjective view that promotes certain power relationships [12]. Thus, it is necessary to exercise extreme caution already at the stage of system design, e.g., by inviting the perspectives of diverse stakeholders and employing alternative analyses to reduce the blind spots.

5.3 Limitations of the Current Work

We must stress three limitations of this work. First, we evaluate an "external" case without the involvement of its owners. As our request for an interview with the experts from the Rotterdam W&I department was not accepted, our analyses fully rely on evidence from high-quality publicly available sources. This should not impact the validity of our arguments in a significant way, but it means that the decision-making system depicted in this work may be a simplification of reality. Second, relatedly, our analysis is not fully objective. This is a common shortcoming of STPA-like approaches because they necessitate certain design decisions [44]. For example, the analyst must decide on the system boundary, the relevant interactions between components, or the scenarios of inadequate control. We make our best effort to motivate all of these choices, but it is possible that other authors would arrive at slightly different conclusions. Finally, we decide to follow the taxonomy of [49] to reason about XAI interventions because it was developed as a comprehensive reinterpretation of earlier efforts. Nevertheless, as also recognized by the author, different (forms of) analyses may be best served by different taxonomies, and other choices were possible.

6 Conclusions

We carried out a case study of an automated decision-making system used for risk profiling in social welfare. Building on earlier works that integrate system-theoretic analyses into the evaluation of AI systems, we applied taxonomy-level reasoning as a method to decide on the value of XAI interventions in specific contexts, and demonstrated its effectiveness in the discussion of five scenarios of inadequate control derived in the case study. Our analysis highlights important fundamental limitations of XAI in that many hazards related to the deployment and use of a model may emerge at other stages of the decision-making process. As a consequence, such hazards cannot be readily mitigated by enhancing the interpretability of algorithmic decisions. This is not to say that XAI solutions are seen as a panacea; as we discuss, many of their drawbacks are well documented. Rather, we believe that with the growing recognition of the "right to explanation" and without an in-depth understanding of the systemic affordances and limitations of XAI, and practical methods to reason about them, such solutions may all too easily become a way to ethics-wash poorly designed ADM systems.

Acknowledgments. We sincerely thank all reviewers for their thoughtful comments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alder, M.: DOJ seeks public input on AI use in criminal justice system (2024). https://fedscoop.com/doj-seeks-input-on-criminal-justice-ai/. Accessed 20 Jan 2025
2. Algorithm Audit: Risk Profiling for Social Welfare Re-examination. Advice document. Technical report AA:2023:02:A, Algorithm Audit (2023)
3. Algorithm Audit: Risk Profiling for Social Welfare Re-examination. Problem statement. Technical report AA:2023:02:P, Algorithm Audit (2023)
4. AlgorithmWatch: how Dutch activists got an invasive fraud detection algorithm banned (2020). https://algorithmwatch.org/en/syri-netherlands-algorithm/. Accessed 20 Jan 2025
5. Altmeyer, P., Farmanbar, M., van Deursen, A., Liem, C.C.S.: Faithful model explanations through energy-constrained conformal counterfactuals. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 10 (2024). https://doi.org/10.1609/aaai.v38i10.28956

6. Araujo, T., Helberger, N., Kruikemeier, S., de Vreese, C.H.: In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc. **35**(3) (2020). https://doi.org/10.1007/s00146-019-00931-w
7. Braun, J.C., Constantaras, E., Aung, H., Geiger, G., Mehrotra, D., Howden, D.: Suspicion Mach. Methodol. (2023). https://www.lighthousereports.com/methodology/suspicion-machine/. Accessed 20 Jan 2025
8. Burgess, M., Schot, E., Geiger, G.: This Algorithm Could Ruin Your Life (March 2023). https://www.wired.com/story/welfare-algorithms-discrimination/. Accessed 20 Jan 2025
9. Castelluccia, C., Le Métayer, D.: Understanding algorithmic decision-making: opportunities and challenges. Technical report, European Parliament (2019)
10. Centraal Bureau voor de Statistiek: Centraal Bureau voor de Statistiek (2024). subpages therein. https://www.cbs.nl/. Accessed 20 Jan 2025
11. Chouldechova, A., Benavides-Prado, D., Fialko, O., Vaithianathan, R.: A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR, vol. 81. PMLR (2018)
12. Crampton, J.W.: Maps as social constructions: power, communication and visualization. Prog. Hum. Geograph. **25**(2) (2001). https://doi.org/10.1191/030913201678580494
13. De Bruijn, H., Warnier, M., Janssen, M.: The perils and pitfalls of explainable AI: strategies for explaining algorithmic decision-making. Gov. Inf. Quart. **39**(2) (2022). https://doi.org/10.1016/j.giq.2021.101666
14. Dobbe, R.: System Safety and Artificial Intelligence (2022). https://arxiv.org/abs/2202.09292
15. Dobbe, R., Wolters, A.: Toward Sociotechnical AI: mapping vulnerabilities for machine learning in context. Mind. Mach. **34**(12) (2024). https://doi.org/10.1007/s11023-024-09668-y
16. Ehsan, U., Riedl, M.O.: Explainability pitfalls: beyond dark patterns in explainable AI. Patterns **5**(6) (2024). https://doi.org/10.1016/j.patter.2024.100971
17. Eubanks, V.: Want to Predict the Future of Surveillance? Ask Poor Communities (2014). The American Prospect. https://prospect.org/power/want-predict-future-surveillance-ask-poor-communities./. Accessed 20 Jan 2025
18. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Statist. **29**(5) (2001). https://doi.org/10.1214/aos/1013203451
19. Geiger, G., et al.: Suspicion Machines (2023). https://www.lighthousereports.com/investigation/suspicion-machines/. Accessed 20 Jan 2025
20. Gomes de Sousa, W., et al.: How and where is artificial intelligence in the public sector going? A literature review and research agenda. Gov. Inf. Quart. **36**(4) (2019). https://doi.org/10.1016/j.giq.2019.07.004
21. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples (2015). https://arxiv.org/abs/1412.6572
22. Grimmelikhuijsen, S., Meijer, A.: Legitimacy of algorithmic decision-making: six threats and the need for a calibrated institutional response. Perspect. Public Manag. Gov. **5**(3) (2022). https://doi.org/10.1093/ppmgov/gvac008
23. Hasan, S.: Governance and public administration. In: Global Encyclopedia of Public Administration, Public Policy, and Governance. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-31816-5_1820-1
24. Heikkilä, M.: Dutch scandal serves as a warning for Europe over risks of using algorithms (2022). https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/. Accessed 20 Jan 2025

25. Hoger Onderwijs Persbureau: Government apologises for discrimination by DUO in fraud detection (2024). https://www.cursor.tue.nl/en/news/2024/maart/week-1/government-apologises-for-discrimination-by-duo-in-fraud-detection/. Accessed 20 Jan 2025
26. Kim, B., Khanna, R., Koyejo, O.: Examples are not enough, learn to criticize! criticism for interpretability. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA (2016)
27. Kudina, O., van de Poel, I.: A sociotechnical system perspective on AI. Mind. Mach. **34**(3) (2024). https://doi.org/10.1007/s11023-024-09680-2
28. Leveson, N.G.: Engineering a Safer World: Systems Thinking Applied to Safety. The MIT Press, Cambridge, MA, USA (2011). https://doi.org/10.7551/mitpress/8179.001.0001
29. Leveson, N.G., Thomas, J.P.: STPA Handbook (2018). https://psas.scripts.mit.edu/home/books-and-handbooks/. Accessed 20 Jan 2025
30. Levy, K., Chasalow, K.E., Riley, S.: Algorithms and decision-making in the public sector. Annu. Rev. Law Soc. Sci. **17**(1) (2021). https://doi.org/10.1146/annurev-lawsocsci-041221-023808
31. Maas, J.: Machine learning and power relations. AI Soc. **38**(4) (2023). https://doi.org/10.1007/s00146-022-01400-7
32. Marcinkevičs, R., Vogt, J.E.: Interpretable and explainable machine learning: a methods-centric overview with concrete examples. Wiley Interdisc. Rev. Data Min. Knowl. Disc. **13**(3) (2023). https://doi.org/10.1002/widm.1493
33. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. J. Artif. Intell. **267** (2019). https://doi.org/10.1016/j.artint.2018.07.007
34. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties: Handreiking Algoritmeregister. Technical Report 1.0, Ministerie BZK (2023)
35. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11) (2019). https://doi.org/10.1038/s42256-019-0114-4
36. Murgia, M.: Algorithms are deciding who gets organ transplants. Are their decisions fair? (2023). https://www.ft.com/content/5125c83a-b82b-40c5-8b35-99579e087951. Accessed 20 Jan 2025
37. Netten, N., Shoae-Bargh, M., Choenni, S.: Exploiting data analytics for social services: on searching for profiles of unlawful use of social benefits. In: Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance. ICEGOV '18, Association for Computing Machinery, New York, (2018). https://doi.org/10.1145/3209415.3209481
38. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, USA (2016)
39. Overheid.nl: Lokale wet- en regelgeving (2024). , subpages therein. https://lokaleregelgeving.overheid.nl/. Accessed 20 Jan 2025
40. Rathenau Instituut: governing algorithmic decision-making in government. The role of the Senate. Technical report Rathenau Instituut (2021)
41. Rechtspraak.nl: SyRI legislation in breach of European Convention on Human Rights (2020). https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-legislation-in-breach-of-European-Convention-on-Human-Rights.aspx. Accessed 20 Jan 2025
42. Rekenkamer Rotterdam: gekleurde technologie. verkenning ethisch gebruik algoritmes. Technical report, Rekenkamer Rotterdam (2021)
43. Rieder, B., Hofmann, J.: Towards platform observability. Internet Policy Rev. **9**(4) (2020). https://doi.org/10.14763/2020.4.1535

44. Rismani, S., Dobbe, R., Moon, A.: From Silos to Systems: Process-Oriented Hazard Analysis for AI Systems (2024). https://arxiv.org/abs/2410.22526
45. Rizk, A., Lindgren, I.: Automated decision-making in the public sector: a multidisciplinary literature review. In: Electronic Government. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-70274-7_15
46. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5) (2019). https://doi.org/10.1038/s42256-019-0048-x
47. Selbst, A.D., Barocas, S.: The intuitive appeal of explainable machines. Fordham L. Rev. **87** (2018). https://doi.org/10.2139/ssrn.3126971
48. Selbst, A.D., boyd, d., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3287560.3287598
49. Speith, T.: A review of taxonomies of explainable artificial intelligence (XAI) Methods. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22, Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3531146.3534639
50. Ustun, B., Spangher, A., Liu, Y.: Actionable recourse in linear classification. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19, Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3287560.3287566
51. Van den Goorbergh, R., van Smeden, M., Timmerman, D., Van Calster, B.: The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J. Am. Med. Inform. Assoc. **29**(9) (2022). https://doi.org/10.1093/jamia/ocac093
52. Veale, M., Binns, R.: Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. Big Data Soc. **4**(2) (2017). https://doi.org/10.1177/2053951717743530
53. Verma, S., et al.: Counterfactual explanations and algorithmic recourses for machine learning: a review. ACM Comput. Surv. **56**(12) (2024). https://doi.org/10.1145/3677119
54. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard J. Law Technol. **31** (2017). https://doi.org/10.2139/ssrn.3063289
55. Weidinger, L., et al.: Sociotechnical Safety Evaluation of Generative AI Systems (2023). https://arxiv.org/abs/2310.11986
56. Wieringa, M.: Municipalities Enacting Algorithms: A Typology of Dutch Municipal Strategies for Leveraging Algorithmic Systems, chap. 1, pp. 19–41. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-84748-6_2
57. Zacharias, J., von Zahn, M., Chen, J., Hinz, O.: Designing a feature selection method based on explainable artificial intelligence. Electron. Markets **32**(4) (2022). https://doi.org/10.1007/s12525-022-00608-1