# Random Sample as a Pre-pilot Evaluation of Benefits and Risks for AI in Public Sector

Steven Vethman[1(✉)], Marianne Schaaphok[1], Marissa Hoekstra[2], and Cor Veenman[1,3]

[1] Netherlands Organisation for Applied Scientific Research (TNO) - Data Science, The Hague, The Netherlands
steven.vethman@tno.nl

[2] Netherlands Organisation for Applied Scientific Research (TNO) - Vector, The Hague, The Netherlands

[3] Leiden University - Leiden Institute of Advanced Computer Science (LIACS), Leiden, The Netherlands

**Abstract.** Public organisations have adopted AI into their public service aiming to tap into the promised potential for society, such as increasing efficiency and effectiveness of current processes. Recent studies from the European Commission share, however, that critical issues of AI use only tended to surface when they were already in operation and thus had already affected citizens. To prevent negative impact to citizens, we propose public organisations to use random sampling as a safe, yet valuable practical evaluation step before considering a pilot. This safe pre-pilot evaluation step enables evaluation of the AI system without applying it in any decisions or actions that already affect citizens. We pose six arguments on the added value of random sampling in the evaluation step of AI systems: 1) it provides high quality data for evaluation and validation of assumptions; 2) it supports gathering input for fairness evaluation; 3) it creates a benchmark to compare AI to alternatives; 4) it enables challenging assumptions in the organisation and the AI development; 5) it supports a discussion on the limitations of AI 6) and it provides a safe space to evaluate and reflect. In addition, we discuss limitations and challenges for random sampling in the evaluation, such as temporary loss of efficiency, class and representation imbalances, organizational hesitancy and societal experiences. We invite the participants of this workshop to reflect with us on the potential benefits and challenges, and in turn distill the practical requirements where using a random sample for evaluation is safe and useful.

## 1 Introduction

With the upcoming AI Act, the European Commission (EC) is providing an EU regulatory framework for the responsible development and use of AI. Groundwork for this regulation started in April 2019, when the High-Level Expert Group


S. Nowaczyk et al. (Eds.): ECAI 2023 Workshops, CCIS 1948, pp. 115–126, 2024.
https://doi.org/10.1007/978-3-031-50485-3_10

on AI presented their Ethical Guidelines for Trustworthy AI [7]. These guidelines set forward seven key principles in terms of requirements that Trustworthy Applied AI should meet. Agreeing with these principles in abstract terms is easy. However, turning them into practice has proven to be challenging, especially for public organisations using AI to aid their public services. A recent report from the Joint Research Centre (JRC) of the EC on AI use in the public sector showed that many projects reached "the adoption phase before finding some unexpected, yet critical, issues" [11]. Typical issues reported were: "legal issues, biased recommendations and staff resistance". In other words, the AI application was already adopted in the way of working of the public service, albeit in a pilot, whilst the legal embedding was still uncertain, the risks and harms related to bias were unknown or unaccounted for, and the identified benefits and possible drawbacks of the AI application did not find a sufficiently large support base.

The term *pilot* is often used to describe adoption in a small controlled setting for experimentation purposes, i.e. a pilot may be a way to find out about these critical issues. However, especially in high risk settings such as essential public services, adoption of AI in terms of a pilot, when the AI system is too premature, can already have too much direct impact on citizens. Often, public organisations start the development of an AI solution with an available data set, which has been compiled to report on the current way of working. This data set is therefore based on the experiences and policies of current operations, in which unconscious societal biases of the people involved are embedded. It occurs too often that public services are not equally accessible to all demographics. Accordingly, the collected data sets have an incomplete or skewed representation of the underlying data distribution that represents all relevant citizens. Moreover, information about these demographics, also known as sensitive attributes, such as nationality or gender are often left out of data collection for good reasons. That is, such attributes are left out from privacy and fairness considerations. It creates so-called fairness through unawareness [10], such that public servants cannot directly differentiate action or treatment based on the collected sensitive data. However, discrimination can also be indirect, e.g. through proxies of the sensitive attributes, such as geographic regions that are linked to ethnic groups. Without information about the sensitive attributes, evaluation of the AI system in terms of fairness is impeded and still leads to the risk of negative impact due to biased selections of the AI. Evaluation methodologies are needed, that are safe to citizens, such that they are not selected by an opaque and insufficiently validated AI system. Especially, when risks and harms are unknown and the added value of the AI application is still uncertain in the organizational embedding.

In this paper, we propose the use of a random sample in a safe pre-pilot phase to evaluate the effectiveness and risks of the use of AI before adoption in a pilot. That is, the risks and possibly undesired selection properties of the AI system are in this phase replaced with a random selection procedure. This enables to test the selection properties of the AI on the data set from the random sample, for which biases from policies and experiences of the current way of working have deliberately been mitigated. The random sample can provide a

benchmark as well as help challenge the assumptions made in the data and during translation from organization goal to model objective. We emphasize that we focus on random sampling in the context of the evaluation of AI algorithms and not in the data collection and development phase, which would require larger sample sizes. We draw our arguments from public use cases, experiences with use cases in among others our AI Oversight Lab[1], talks with governmental bodies and peers, as well as the growing body of literature.

In the next section, we discuss the position of our methodology compared to developed impact assessments and widely-used data science methodologies. In the third section, we introduce a fictional high risk use case of AI adoption to concretely illustrate each argument in this paper. The fictional use case makes the discussion more concrete by putting focus on the lessons learned, without directly referring to wrongful practices of specific organisations. The fourth section discusses our five arguments for using a random sample for evaluation purposes before running a pilot. The fifth section presents challenges and critical remarks concerning our proposed additional pre-pilot development phase. The sixth section provides the conclusion. We hope that the insights of this paper ignite a discussion at this workshop on under which conditions random sampling can be a safe and valuable evaluation step in AI development for the public sector.

## 2 Background

A lot of work has been done on guidelines [7] and impact assessment frameworks [6,14] to help organizations with responsible development and use of AI. These frameworks aim to identify benefits and risks of AI systems in an early phase. However, recent work [18] and experiences by public organizations show that the practical application of these frameworks is not evident. There are challenges such as contextualization (how do these generic guidelines fit within the context of ones organization and application?), subjectivity (evaluations and impact assessments can be affected by individual beliefs and experiences of the evaluator) and knowledge deficits (does the organization have the required knowledge on technical, societal and legal issues?). [18] states that many impact assessments are based on subjective answers from people within the organization which introduces the risk of human biases. They argue that concrete, contextualized and more objective solutions or procedures should be developed and/or incorporated into the risk assessment process to improve the utility and reliability of AI risk assessments.

Furthermore, we would like to set current AI development practices in the context of a widely-used data science methodology called Cross Industry Standard Process for Data Mining (CRISP-DM) [9,13,17]. In CRISP-DM, one starts with a business or societal goal in the real world. By making a set of assumptions, said goal is translated into a modelling goal within the limits of available data. After the data are collected and prepared, the model is optimized and

[1] https://appl-ai-tno.nl/projects/ai-oversight-lab/.

tested on the available data. That is, the data that represents the historical way of working, including assumptions on missing information. After a satisfactory performance is measured in the historical data set, CRISP-DM suggests an evaluation step to gauge the effectiveness of the application in the real world. In most cases this results in doing a pilot. Referring back to the JRC report, we state that this evaluation step is often taken too early as discovering critical issues while impacting citizens is undesirable. We state that the evaluation step in CRISP-DM needs a prior effort to safely establish a realistic benchmark as well as to facilitate a critical reflection on the data and the goal translation.

Our proposed random sample adds a pre-pilot evaluation phase that makes the assessments of effectiveness and risks more concrete and objective. It allows to evaluate and compare alternatives on high quality data. Only after such diligent considerations can public organisations decide whether the potential impact of running a pilot is desirable.

## 3 Fictional Use Case for Illustration of Arguments

In practice we see many governmental organisations, such as inspectorates and municipalities working on risk-models to identify and prioritize cases (organisations or individuals) for inspection in order to make their inspections more effective [8]. The expected positive impact is to make more effective use of the capacity of inspectors, to reduce the impact on compliant organisations/citizens and to increase compliance in general. Such risk-models can be considered high-risk AI applications, because the outcomes may influence whether or not an individual or organisation will receive an essential public service, which has direct economic and/or social consequences.

Here we propose a hypothetical case as a running example. It considers an organisation that is piloting a risk-model to prioritize inspections to detect fraud by social welfare recipients. Inspectors of the organisation use the outcomes of the AI system to determine which recipients to inspect first. The risk-model is trained on data representing the current way of working (before adoption of AI). Here choices for whom to inspect were based on earlier insights of inspectors or warning signs of fraudulent behavior such as a suspicious neighbour. We would like to emphasize that with the choice of this example we do not argue that AI should or should not be used for the detection of this type of fraud. We choose this setting for relatability, due to the numerous examples of AI adoption for this purpose. We also choose this High Risk setting to underline that such cases in particular need safe experimentation and critical reflection before deciding whether the AI system should be adopted in operation, albeit in a pilot.

In this example, we propose the random sample evaluation in the following way. The organisation uses a random selection procedure to allocate 100 inspections of recipients for whom it is not yet known whether they are fraudulent or not. Inspectors investigate these recipients to acquire information on whether they are fraudulent. Note that only the selection process is adapted to a random sample, the inspection itself and possible follow-up actions remain the same to

the current way of working. In the evaluation step one can ask post-inspection to (other) inspectors: which of these 100 would you have suggested for an inspection? And, similarly to the AI application: which of these 100 would the AI application have given a high risk score. An error-analysis could answer how many fraudulent citizens the AI or inspectors would have correctly suggested for inspection, how many compliant recipients would have been given a high risk score and how many fraudulent recipients would have been missed by either alternative. That is, you measure the effectiveness and risks of each alternative, whilst only the random sample has affected the recipients.

## 4 Arguments

In this section, we present five arguments for a random sample in a pre-pilot evaluation to gain practical insight in the benefits and risks of adoption of AI solutions in the public sector.

### 4.1 Testing Assumptions

First, random sampling is a means to gather data of higher quality for evaluation before running a pilot. In the development of AI algorithms assumptions have to be made on the data distribution. In context of our running example, an important assumption concerns the proportions of fraudulent and non-fraudulent social welfare recipients. For example, this affects discussions on the added value of an AI algorithm (were fraudulent recipients hard to find?), as well as the choice of the type of the AI algorithm (are fraudulent recipients a sizable class with different behaviour or are they uncommon outliers that need to be detected). Additionally, organizations are often limited to the available data from the current way of working. Is data on how the public service is currently performed a sufficiently reliable data set to validate these assumptions? When pursuing a developed algorithm towards operationalization, it is essential to verify the assumptions on the data and the representativeness of the data [2]; does the data set based on the current way of working represent a realistic setting in which the developed algorithm may be applied?

Collecting a random sample in the pre-pilot phase leads to high quality data where certain historical biases in the current way of working are limited. Inspectors choosing who to check for fraud are not free from subconscious human bias, which results to their choices potentially reflecting systemic discriminatory tendencies. Intended systematic preferences in organizational/political policy to give more checks to foreign-born subsidy recipients have also been reported. Moreover, inspectors might do some desk research before choosing who gets a full inspection for fraud, such that those who commit fraud in unforeseen ways are less targeted for full inspections. A data set based on random sampling where these biases are limited constitutes another, arguably cleaner, data set upon which the outcomes of the algorithm can be evaluated.

Moreover, the "cleaner" data set also allows a comparison with the distribution of the data set for development purposes which represents the current way of working. In the context of our example, this could mean that the proportions of fraudulent and non-fraudulent welfare recipients of both data sets can be compared. This comparison can provide a sanity check whether the development data set is a suitable representation of the real (or desired) world for training the AI application. This can empower the organization to decide whether alternatives or additional measurement such as extended data collection are required. For example, in our fictional use case the data set for development could contain significantly more fraudulent recipients from a specific region, because people and institutes in that region were more observant and proactive in issuing warnings. The random sample data set can show that this group is over-represented in the training set.

### 4.2 Fairness Calibration

Second, random sampling supports the gathering of valuable input for quantitative fairness metrics that can help signal undesirable differentiation and negative impact towards certain groups. The way in which the data collection has been performed in the current way of working can lead to the practical issue that relevant quantitative fairness metrics cannot be measured. Since data collection is often designed for administration purposes rather than the purpose of developing AI models, important labels or variables for fairness evaluations may be missing. For example for the fairness metric Equalized Odds, you need an indication per social demographic of the ratio of fraudulent and non-fraudulent recipients to see how the predictions of the algorithm deviate from these ratios [10]; acquiring this indication requires a data collection procedure that is not driven by warning signals or intuition of inspectors on who is worthy of an inspection. If reporting phone calls have mainly come from certain neighbourhoods, the data set may have misleading ratios as the calls disproportionately concern the dominant demographics from those neighbourhoods. Similarly, for the analysis of proxies for protected classes, the value of the protected class is required for the samples. We note that from a nondiscriminatory perspective, organisations have reasons to not store these sensitive attributes in their current way of working [12]. However, the ability and importance to evaluate against sensitive attributes is recognized in the current version of the AI Act as of June 14 2023 [5], which allows for the collection of sensitive attributes with the mere purpose of fairness evaluations. A random sample provides an opportunity to design and reflect on a safe data collection process, whilst not simultaneously dealing with implications of operationalizing AI.

### 4.3 Alternatives Comparison

Third, random sampling facilitates an evaluation of AI in terms of alternatives. Performing a random sample provides a benchmark to which not only the AI algorithm, but also the current way of working and other alternatives can be

compared. This moves the perspective of the evaluation from an isolated evaluation of AI (focusing on the absolute risks and benefits of implementing it) towards a relative evaluation where the downsides and benefits of alternatives are also actively considered. In practice, we see that development and implementation of AI often rely on go/no-go moments which represent the alternatives of either implementing the AI in the public services or maintaining the public service as is. We argue that the benefits and downsides of alternatives such as maintaining the current way of working deserve as diligent of an evaluation such that the consequences of a no-go decision are also clearly understood. In the execution of public service, legal concepts such as proportionality (are the means suitable, necessary and not excessively burdening citizens to achieve the objective?) and subsidiarity (are there no alternative means which impose less burden to attain similar goals?) are key and should be considered carefully [4].

### 4.4 Reflection on Goal Translation

Fourth, random sampling is a means to challenge the assumptions made in the data processing as well as those made to translate the societal goal to modeling criteria. Closely related to the first and third argument, we would like to emphasize that executing a random sample, and therefore partially or temporarily changing the way of working, creates a setting of critical reflection on the current way of working. This is an opportunity to discover blind spots regarding unwanted impact with respect to any public values important to the public organization. For the development of the AI algorithm, this reflection may also help checking the explicit assumptions made in terms of data processing and those made to translate the real world goal to the model goal within the available data. For example, in our fictional example, the real world goal is to use the capacity of the inspectors more effectively to reduce fraud. The current model goal is whether a recipient is likely of committing fraud. It can be discussed whether this focus on fraudulent behavior prediction is the right translation of the real world goal to the model goal. Alternatively a model can also focus on prioritizing cases such that capacity is used most efficiently. For example a model that schedules the cases according to the ability of inspectors or the expected difficulty, duration, and impact of the inspection. Another reflection may pertain to the translation of the societal goal to reduce fraud to a single technical definition of fraud for the model. Is a measure that aggregates different types of fraud desirable? I.e. is it justifiable to consider fraud committed on purpose equivalent to unintentional misunderstanding of the exact duties for receiving welfare [15]?

### 4.5 Understanding Limitations of AI

Fifth, evaluating with a random sample facilitates the conversation on the limitations of an algorithm. The application of algorithms can lead to automation bias and can give people a false sense of objectivity. This can lead to insufficient validation of the algorithm outcomes and decrease the incentive for a human

touch in exceptional cases. Comparing the current way of working and the algorithm on a random sample allows to challenge the false sense of objectivity in the perception of the algorithm's outcomes. AI's association of objectivity often comes from the inhuman/ emotionless characteristic of computers as well as the fact that the AI generalizes, i.e. it provides suggestions based on pattern found on a large number of examples. We argue however, that a generalized pattern based on historical practice, may be void of an individual subjectivity from a particular inspector, but is not void of a shared (undesirable) subjectivity. Think of systemic racism, or misogyny, xenophobia, which are forms of societal oppression not individual isolated phenomena. These may for example systemically alter which groups in society receive inspections. Random samples can show the differences between inspections based on a random sample, the current way of working and the algorithm. This can contribute to the conversation about the possibilities and limitations of the algorithm, such that the trust that is placed in the system is more adequate and responsible.

### 4.6 Safe Space

Sixth, our overarching argument for a random sample is that it provides a relatively safe environment for the evaluation of the AI algorithm and alternatives. Safe meaning here that critical issues such as the biased recommendations mentioned in the JRC report, do not yet impact the citizens during evaluation of the AI algorithm. As described in Sect. 3 only the random sample decides which recipients are inspected and the AI-based recommendations are evaluated on the results of these inspections. This ensures that citizens are not yet impacted by the recommendations of the AI system. A fundamental assumption here is that it is more safe for citizens and society when inspections are conducted at random rather than steered by human inspectors or by AI. That is, the harm experienced by a social welfare recipient to undergo a full inspection, completely by chance, is less than when the inspection is based on an inspector's intuition, warning signals, or an AI application, which are often considered opaque and inexplicable.

This additional evaluation phase also provides space in the organization to start the safe discussion on the social embedding of quantitative metrics. For example, consider the question of human accuracy versus computer accuracy; is 80% accuracy of a human valued similarly to 80% accuracy of the algorithm or do we require a higher standard for systematic evaluations? Or think of the previously mentioned discussion on whether unintentional fraud and intentional fraud should be aggregated when partially automating your public service. These discussions are essential to be able to translate evaluations to decisions on operationalization and recognize their capabilities and limitations. Conscious, transparent and documented decisions on these topics support the accountability and hence the responsible use of AI.

## 5 Discussion

The notion that random sampling will provide clean data is of course not new and there are multiple arguments why random sampling on a large scale for the training of AI models is often not possible. Hence we do not propose that random sampling should be the basis for AI development, but should merely be used in the pre-pilot phase to test assumptions, evaluate effectiveness and risks and compare alternatives. In this setting, we expect that the required sample size for evaluation purposes can be much smaller than for development of AI algorithms [16]. However, also for this application there are limitations and challenges that need to be considered.

### 5.1 Temporary Loss of Efficiency

Often the aim of using AI is to make a certain process faster and more efficient; in case of governmental institutions this results in helping more people. Especially since many organisations are looking at AI solutions to help dealing with their increasing workload. Performing a random sample will take up space and time from the current employees, which might result in less efficiency for a specific period of time. Assuming that an alternative finds more fraud, during the time of random sampling less cases of fraud are detected. From a societal perspective this means that taxpayer's money is lost. One should keep in mind here that the loss of efficiency is based on the assumption that the current way of working is more effective than the random sample, which is often unknown. In the context of fraud in social welfare in the Netherlands, most cases consist of unintentional fraud linked to the complexity of eligibility rules [3]. Because of this, inspectors may therefore have the unfounded notion that almost every visit was useful due to their skillful intuition who is committing fraud, whilst in practice almost all social welfare recipients have a difficult time to have an overview of their financial situation and the rules pertaining social welfare.

### 5.2 Imbalance in Class and Representation

Secondly, major class imbalances can prove a challenge for the random sample. Considering a scenario where only 10% of social welfare recipients are fraudulent. In that case, a smaller random sample is required to find the correct distribution than in cases where the classes are balanced (50% is fraudulent) [1]. To illustrate, assume that we have a population of N = 10 000 for which we assume a 10–90% distribution, where 10% is non-compliant. In this case, a random sample of approx. n = 50 is required to get estimate with a margin of 8%. If we assume a 50-50% distribution in the same population a random sample of approx. n = 150 is required to achieve an estimate within a margin of 8%. However, in many cases not only the distribution in the population but also information about the minority classes is required. In this case, the 5 samples (10% of n = 50) of non-compliant recipients are not sufficient. To achieve a suitable representation

of the minority class(es) a larger random sample is required. Especially when fairness evaluation requires substantial representation of multiple demographics.

Related to the class imbalance there is also a challenge regarding small sensitive groups. In order to measure whether the AI application functions desirable for all relevant demographics, the individuals from these sensitive groups should contain samples both positive and negative. Since sensitive groups can be minorities, performing a random sample can be challenging in cases with large class imbalances and small minorities. Alternatively, stratified random sampling could prevent this problem, whilst this would also mean that the inspections are allocated based on the demographic membership of social welfare recipients.

### 5.3 Organizational Hesitancy

From an organizational perspective, inspectors might be reluctant to perform random samples instead of following their intuition or warning signs, as they consider it a waste of time. This drawback is even more pertinent when an inspection is very costly in terms of time spent by inspectors or time spent by welfare recipients. There is a related challenge as inspectors might execute an inspection less elaborately, if they know it is based on random selection rather than insight. This could affect the reliability and quality of the random sample.

On the higher organizational level of program manager, team lead or department head, hesitancy can occur due to the fact that the advantage and necessity of better evaluation and monitoring is not always properly understood. A meaningful size of a random sample takes often a substantial time to execute, which is in contrast to the entrepreneurship and innovation mindset of "move fast and break things". Random sampling has value in being diligent, avoiding errors and investing in sustainable innovation, whilst managers are often rewarded for short-term gains in efficiency or effectiveness. Especially with the sensitive nature of leading a team that experiments with AI for public service (due to increased scrutiny from society), the turnover in these positions is fast. Managers may therefore be disincentivized to do initiatives which are perceived to only have long-term benefits and prefer to run a pilot in operations to show the rewards from the investment in AI development. Hence from their perspective this does not directly weigh up against the short-term downside of loss of efficiency.

### 5.4 Experience of a Random Inspection

Lastly, as flip side of Sect. 4.6, the use of a random sample can still lead to citizens experiencing increased stress as they feel they are under suspicion and, in our example, marked as (potentially) fraudulent by the public organisation. Even though it is a random selection, who undergoes the inspection is still selected from the larger population. To experience that you are selected whilst others are not, may feel unfair. Especially those, who have seen public organisations making mistakes, are critical about the government and therefore may doubt the randomness of the selection. Based on the current way of working the assumption can live within society that if you follow the rules you will not be inspected. In

this case an inspection may still feel as an invasion to the citizen. On the other hand a random sample can also be experienced by individuals or society as a just way to inspect and validate the way of working and to maintain societal support for public services.

## 6 Conclusion

Public organisations want to tap into the potential societal benefits of adopting AI into their public service. Unfortunately, we are all familiar with too many instances where critical issues of adoption of AI only surfaced when they were already in operation and thus had affected citizens. We propose public organisations to use random sampling as a safe, yet valuable practical evaluation step without possible negative impact on citizens. Random sampling is a means to gather higher quality data for evaluation, including input for common fairness metrics. Additionally, a random sample provides a benchmark to compare performance of alternatives to. Relating the current way of working to alternatives also sets the scene where assumptions of the model, data processing and goal translation can be challenged. This comparison also facilitates discussions that lead to understanding of the limitations of AI and a more adequate level of trust. Most importantly, it provides a safe environment to evaluate AI systems without negatively impacting citizens.

Critical reflection on random sampling indicates that class and demographic imbalances provide challenges for desired evaluation of effectiveness and risks such as fairness. Moreover, the random sampling can be met with organisational hesitancy due to expected loss of efficiency, which in turn affects the reliability of the inspections. Lastly, the aspect of safety of a random sample is only as valid as the assumption that equal unconditional chance for inspection is considered fair and experienced less burdensome than being selected for an inspection based on an inspector's intuition or an AI's suggestion. We invite the participants of this workshop to reflect with us on the potential benefit and challenges, and in turn distill the practical requirements where using a random sample for evaluation is safe and useful.

**Acknowledgement.** We would like to thank all our colleagues in the AI Oversight lab[2], our external partners, as well as all other public and private organisations that have facilitated transparency on this urgent yet sensitive topic such that the lessons described in this paper could be learned.[2]https://appl-ai-tno.nl/projects/ai-oversight-lab/

## References

1. Bethlehem, J.: Applied Survey Methods, a statistical perspective. John Wiley and Sons Inc (2009)
2. Clemmensen, L., Kjærsgaard, R.: Data representativity for machine learning and AI systems (2022)

3. Dannenberg, E.: Factsheet overtredingen van de inlichtingenplicht - meer maatwerk en eenvoudigere regels (2021). https://www.divosa.nl/publicaties/factsheet-overtredingen-van-de-inlichtingenplicht/factsheet-overtredingen-van-de
4. EUR-Lex Access to European Union Law: glossary proportionality. https://eur-lex.europa.eu/EN/legal-content/glossary/principle-of-proportionality.html
5. European Parliament: amendments adopted by the European parliament on 14 June 2023 on the proposal for a regulation of the European parliament and of the council on laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2023). https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html
6. Gerards, J., Schäfer, M., Muis, I., Vankan, A.: Fundamental rights and algorithms impact assessment (FRAIA). Utrecht University, Tech. rep. (2021)
7. High-Level Expert Group on Artificial Intelligence: ethics guidelines for trustworthy AI. Tech. rep, European Commission (2019)
8. Hoekstra, M., Chideock, C., Veenstra, A.: Quick scan AI in de publieke dienstverlening ii. Tech. rep. (2021)
9. Martínez-Plumed, F., et al.: CRISP-DM twenty years later: from data mining processes to data science trajectories. IEEE Trans. Knowl. Data Eng. **33**(8), 3048–3061 (2021). https://doi.org/10.1109/TKDE.2019.2962680
10. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
11. Molinari, F., Van Noordt, C., Vaccari, L., Pignatelli, F., Tangi, L.: AI watch beyond pilots: sustainable implementation of AI in public services (KJ-NA-30868-EN-N (online)), 14 (2021). https://doi.org/10.2760/440212(online)
12. Reventlow, N.J.: Data collection is not the solution for Europe's racism problem (2020). https://www.aljazeera.com/opinions/2020/7/29/data-collection-is-not-the-solution-for-europes-racism-problem
13. Schröer, C., Kruse, F., Gómez, J.M.: A systematic literature review on applying CRISP-DM process model. Proc. Comput. Sci. **181**, 526–534 (2021). https://doi.org/10.1016/j.procs.2021.01.199, https://www.sciencedirect.com/science/article/pii/S1877050921002416
14. Stahl, B., et al.: A systematic review of artificial intelligence impact assessments. Artif. Intell. Rev. (2023)
15. Steen, M., Timan, T., Vethman, S.: Using an extended error matrix to promote transdisciplinary collaboration and jointly work towards social justice (2022). https://marcsteen.nl/docs/ESDiT_2022__Error_Matrix.pdf
16. Valizadegan H, Amizadeh S, H.M.: Sampling strategies to evaluate the performance of unknown predictors. In: Proceedings SIAM International Conference Data Mining (2014)
17. Wirth, R., Hipp, J.: CRISP-DM: towards a standard process model for data mining (2000)
18. Xia, B., Lu, Q., Perera, H., Zhu, L., Xing, Z., Liu, Y., Whittle, J.: Towards concrete and connected AI risk assessment (C2AIRA): a systematic mapping study. In: 2023 IEEE/ACM 2nd International Conference on AI Engineering - Software Engineering for AI (CAIN), pp. 104–116. IEEE Computer Society, Los Alamitos, CA, USA (2023). https://doi.org/10.1109/CAIN58948.2023.00027