Article

Framework for Assessing Ethical Aspects of Algorithms and Their Encompassing Socio-Technical System

Xadya van Bruxvoort¹ and Maurice van Keulen^{2,*} ¹ M&I/Partners, 3708 Zeist, The Netherlands; xadya.van.bruxvoort@mxi.nl² Faculty of EEMCS, University of Twente, 7522 Enschede, The Netherlands

* Correspondence: m.vankeulen@utwente.nl; Tel.: +31-534893688

Abstract: In the transition to a data-driven society, organizations have introduced data-driven algorithms that often apply artificial intelligence. In this research, an ethical framework was developed to ensure robustness and completeness and to avoid and mitigate potential public uproar. We take a socio-technical perspective, i.e., view the algorithm embedded in an organization with infrastructure, rules, and procedures as one to-be-designed system. The framework consists of five ethical principles: beneficence, non-maleficence, autonomy, justice, and explicability. It can be used during the design for identification of relevant concerns. The framework has been validated by applying it to real-world fraud detection cases: Systeem Risico Indicatie (SyRI) of the Dutch government and the algorithm of the municipality of Amersfoort. The former is a controversial country-wide algorithm that was ultimately prohibited by court. The latter is an algorithm in development. In both cases, it proved effective in identifying all ethical risks. For SyRI, all concerns found in the media were also identified by the framework, mainly focused on transparency of the entire socio-technical system. For the municipality of Amersfoort, the framework highlighted risks regarding the amount of sensitive data and communication to and with the public, presenting a more thorough overview compared to the risks the media raised.

Citation: van Bruxvoort, X.; van Keulen, M. Framework for Assessing Ethical Aspects of Algorithms and Their Encompassing Socio-Technical System. *Appl. Sci.* **2021**, *11*, 11187. <https://doi.org/10.3390/app112311187>

Academic Editor: Eyad H. Abed

Received: 30 September 2021

Accepted: 22 November 2021

Published: 25 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: ethics; algorithms; artificial intelligence; responsible AI; fairness; ethical framework; fraud detection

1. Introduction

A considerable amount of innovation by private and public sector organizations is driven by the use of data. In 2018, 2.5 exabytes of data were generated per day [1] and it is expected that this amount will only grow. For example, fifty percent of the municipalities in The Netherlands use data to a certain extent and 56% of governmental institutions perform research on algorithms [2].

The issues created by artificial intelligence, or rather algorithms in general, extend far beyond the technical sphere. For a proper implementation and use of algorithms, they are embedded in a socio-technical system consisting of rules and regulations, organizational structure, etc. In our experience, the rules and regulations and the organizational structure are especially important for properly addressing the ethical concerns algorithms may have, hence a well-functioning socio-technical system should be well-designed for this purpose. A major consultancy firm (KPMG) described the AI ethicist as an essential position for a good AI implementation [3].

Illustrative examples can be found among fraud detection algorithms developed by municipalities and other governmental institutions. These algorithms typically scan through individuals' data and create a subset of cases with a high suspicion of fraud. This subset can then be investigated further by a human aiming to confirm and decide whether fraud was indeed committed and to collect evidence to build a case. This is done because municipalities have the duty to monitor the proper and fair use of their societal services, hence to find and punish fraudsters [4]. The damage caused by fraud can be huge;

estimated at more than a yearly EUR 1 billion of benefits fraud in the Netherlands alone [5]. A further reason for the interest in applying data-driven algorithms for fraud detection is that the available resources for human inspection are limited. Municipalities often have only a small team of people responsible for catching fraudsters, so an algorithm allowing them to focus their attention on a small set of suspicious cases is very attractive.

Although attractive and seen as an important future direction, concerns related to the use of data-driven algorithms may arise. For example, the Dutch government came into trouble because of what is now known as the “Dutch childcare benefits scandal” (Dutch: “Toeslagenaffaire”) where, between 2013 and 2019, authorities wrongly accused an estimated 26,000 parents of making fraudulent benefit claims for day-care, requiring them to pay back the allowances they had received in their entirety, which drove many families into severe financial hardship. A parliamentary interrogation committee concluded among other things that “unprecedented injustice” was done and that “affected parents did not receive the protection they deserved” by the system [6]. Another notable example is when the Dutch government introduced the *Systeem Risico Indicatie* (SyRI) system. This was a social security fraud detection algorithm of the Dutch government. It received an ironic privacy prize for the invasion of privacy of people [7]. There were five reasons for this according to Bits for Freedom: citizens were a suspect in advance, it felt like a violation of their privacy, data might have been used without purpose limitation, it might have been discriminating, and it would be the first step towards a control society. Parliamentary discussions were held about the system [8]. It became clear that in the design and realization of the system, too little attention was paid to ethical concerns and that valid points raised by the public were insufficiently addressed.

1.1. Contributions

The main aim of this research is to develop a framework for identifying and assessing ethical concerns regarding the use of data-driven algorithms. We take a socio-technical perspective, i.e., view the whole system of the algorithm as a technical component implemented in a technical infrastructure and an organizational structure accompanied by associated rules and regulations. The framework can be used during the design for identification of relevant concerns so that they can be addressed early in the design of the entire socio-technical system. In this way, it may contribute to a more robust, complete, and effective introduction of such systems, improve the quality of their designs from the ethical perspective, and avoid injustice and associated uproar. The framework has been developed by merging ethical frameworks from the literature and enriching and fine-tuning it on the basis of expert interviews and brainstorm sessions. The resulting framework is based on five ethical principles that algorithms and the surrounding socio-technical system should respect: beneficence, non-maleficence, autonomy, justice, and explicability. The framework has been validated by applying it to two fraud detection case studies: A fraud detection algorithm under development by the municipality of Amersfoort and *Systeem Risico Indicatie* (SyRI) of the Dutch government.

1.2. Outlook

In this research, the background will be provided first in, Section 2. Section 3 mentions and briefly explains related ethical frameworks. The framework is illustrated and validated in Section 4. Conclusions and suggestions for future research can be found in Section 5.

2. Background

2.1. Algorithms

An algorithm can be described as a recipe to solve a certain problem. Algorithms are not a new thing. They have existed for centuries, for example, Euclid’s algorithm was created 300 BC. In modern times, we associate the term ‘algorithm’ usually with computers, in particular for accomplishing intelligent tasks by computers. Examples of well-known algorithms can be found in facial recognition, spam detection, product recommendations

on sites like Bol and Amazon, and fraud detection. The definition of algorithm used by this research will be intelligent computer algorithms that in some way help with the decision-making process about individuals or groups of people, involving personal data.

Fraud detection algorithms typically use data combined from different data sources. For fraud detection, these can be, for example, data on healthcare, social assistance, and property. In the combined and integrated data, one attempts to find certain unusual suspicious patterns. For example, person A receives social assistance, but does not cancel this when getting a new partner. Or person B works ‘under the radar’ not receiving an official salary, while receiving social security assistance at the same time. It is important to note that social security assistance fraud is not always intentional.

In this research, we distinguish two kinds of algorithms: static rule-based algorithms and learning algorithms. Static algorithms are explicitly programmed by humans, typically programmed to execute a number of if-then or other types of rules. An example of such a rule is ‘If person A has most of their transactions done in Amsterdam, but claims to live in Rotterdam, then this is suspicious’. A learning algorithm, on the other hand, applies machine learning, the capability of learning from example data, discovering patterns, which it can subsequently apply to new data producing the intended output even though the exact input data was never seen by the algorithm before. Such learning algorithms are nowadays often also referred to as artificial intelligence although the latter term is much broader than just the algorithm.

2.2. Artificial Intelligence (AI)

According to [9] “Artificial intelligence is the study of ideas to bring into being machines that respond to stimulation consistent with traditional responses from humans, given the human capacity for contemplation, judgment and intention. Each such machine should engage in critical appraisal and selection of differing opinions within itself. Produced by human skill and labour, these machines should conduct themselves in agreement with life, spirit and sensitivity, though in reality, they are imitations”.

As mentioned, AI typically involves machine learning and as such, AI is a special kind of algorithm. Given historic data, machine learning can learn a ‘model’ that is capable of producing an intended output even for unseen case, because it has learned the hidden patterns. In a sense, an AI is able to program itself given only examples. An important class of AI algorithms are classifiers: given input data of a case, the classifier determines in which class it belongs. For example, a fraud detection algorithm is a classifier: given data about the case, it determines whether it belongs to the “high risk for fraud” class rather than the “low or no risk for fraud” class (classification into more than two classes is also possible). Learning can also be integrated into its use, i.e., that it not only derives a model from historic data, but continuously re-learns a model from historic data supplemented with data of new cases. In this way, it allows the AI to adapt itself to changing circumstances.

Compared to static algorithms, learning algorithms are technically more advanced and have the ability to discover hidden patterns that humans might miss. They tend, however, to be more complex, dynamic and less transparent, hence need more safeguards and are more difficult to implement. To illustrate, a systematic review of image-based learning algorithms for the diagnosis or prognosis of COVID-19 revealed that none of the algorithms were of potential clinical use due to methodological flaws and/or underlying biases [10].

In short, AI, especially learning algorithms, can perform similar (cognitive) tasks as humans, sometimes even better than humans. Machine learning is the technique behind the intelligent capabilities.

2.3. Socio-Technical System

When an algorithm is developed to make a sensitive classification such as “high risk for fraud” or not, it is obviously not supposed to be implemented by directly relying on the AI’s decision. The algorithm should be embedded in a socio-technical system with, for

example, procedures for human evaluation, overrides, periodic checks, periodic updates. A socio-technical system is meant to be an effective blend of both the technical and social systems of an organization, viewing both as one integrated system [11].

Since IT has become an integral part of our work and our lives, it is important to take the entire socio-technical system into consideration including how they interact. For fraud detection cases specifically, the goal is to compensate for the weaknesses of the one with the strengths of the other, so that in the end they enhance each other and can work together smoothly.

3. Method

3.1. Research Method

The research method is illustrated in Figure 1. As a starting point, we take the AI4People ethical framework (see Section 3.2.1). Based on other related work on ethical frameworks, we propose additions to make it more complete. Furthermore, based on considerations for applicability in practice, we add more detailed questions to the main aspects of the framework. The framework is validated by means of expert interviews and by application to two real-world cases. To be transparent, experiences during the validation instigated a few alterations in wording as well as a few additions in the form of detailed specifications. The subsequent sections report on these research steps.

Figure 1. Research method.

3.2. Related Research on Ethical Frameworks

To allow for complete and thorough identification of ethical concerns, ethical frameworks have been developed. Important to note is that not many (relevant) frameworks exist. We only chose six frameworks relevant to the mindset and moral of the Netherlands or the Western part of Europe. Note also that most frameworks we found are not of a scientific origin, but are nevertheless commonly known and used. The ethical frameworks refer to several different ethical issues such as bias and privacy. Although it is beyond the scope of this paper to provide an overview of scientific literature on these ethical issues, we do provide a brief overview of survey papers and other important papers on these ethical issues in Section 3.3.

3.2.1. AI4People

AI4People is a framework on ethical AI developed by Floridi et al. [12]. For its conception, six documents on responsible AI were analyzed to find the common ground. In the complete research, they describe risks and opportunities as well as twenty concrete recommendations in assessment, development, incentivisation and support. Their framework comprises five ethical principles of which the first four were already proposed by others.

Beneficence: This principle states that the humans' and planet's well-being should be put upfront.

Non-maleficence: This principle states that algorithms should 'do no harm', which entails not infringing the privacy of people and avoiding misuse of AI technologies in other ways. Important for this principle is to prevent harm both by intent of humans and by unforeseen behavior of machines.

Autonomy: This principle states that people should have the right to decide. That this is not always the case is illustrated with a medical example: a patient may not have the mental capacity to decide on receiving a certain treatment and will be given the treatment decided by others. Here their autonomy is infringed, which is, as illustrated here, not necessarily a bad thing. Besides right to decide, there should also be a possibility to turn off the AI to regain control.

Justice: This principle is a rather classic one stating that no discrimination should occur. Regarding AI, it effectively means that the interests of all parties that may be affected by AI advances should be respected.

Explicability: This principle proposed by Floridi et al. is a combination of transparency and accountability and includes questions like ‘how does it work?’ and ‘who is responsible for the way it works?’.

3.2.2. Guidelines for Using Data from Online Networking Sites

Van Wynsberghe, Been and van Keulen [13] proposed guidelines to provide a method for early assessment of ethical privacy concerns for research involving social media data. As such, the guidelines are more specific than the general scope of algorithms of this paper, but insightful for the larger scope nonetheless. Five guidelines were proposed:

- Get a good overview of the stakeholders.
One should identify all stakeholders and, for example, consider not only users, but also developers and system owners.
- Evaluate the context and privacy within this context.
To stay within the social media domain, there is a difference in privacy concerns between a message posted in a private, closed Facebook group and a message posted openly, possibly with hashtags, on Twitter. Even if a message is posted in an open public forum, this does not mean that it is public property for which any usage is allowed. The metaphor of a (private) conversation in a pub illustrates this point: even though the pub is a public place and it is evident that everyone can overhear it, it does not mean that it is ethically justified to use the overheard information, or one step further, put a microphone on every table in the pub to record everything, which is what social media analysis actually does.
- Consider the type and method of data collection.
This concerns whether the data is collected with consent and whether it is collected actively (following people on a defined list created beforehand) or passively (follow everyone who comes by).
- Determine the intended use of information and amount of information collected.
- Perform a value analysis.
This guideline advises to go beyond the direct purpose of the use of data and consider all effects, even indirect and/or global ones. With social security fraud detection, as mentioned by van Wynsberghe, Been and van Keulen [13], there is a positive intended purpose of achieving more fairness in the system, more financial security for the Dutch state and the obligation to prevent fraud and maintain fair taxation. However, there is also a negative easily overlooked side which includes non-fraudsters still being checked by an AI leading to a possible infringement of their privacy. This is known as the value trade-off between justice and privacy. It is important to realize that ethics is not absolute and that often a value trade-off should be made causing a use of data to be acceptable for one purpose and context while it is not acceptable in another. For example, using social media data for fraud detection may be justified, while a similar use for marketing purposes is not.

3.2.3. Partnership on AI

Partnership on AI is a community with a mission to benefit people and society with AI by conducting research, organising discussions, sharing insights, and responding to questions from the general public [14]. It is a diverse community with people from different backgrounds, such as academics, researchers and companies. Insightful in the light of AI frameworks are the tenets its members endeavor to uphold:

1. We will seek to ensure that AI technologies benefit and empower as many people as possible.
2. We will educate and listen to the public and actively engage stakeholders to seek their feedback on our focus, inform them of our work, and address their questions.

3. We are committed to open research and dialogue on the ethical, social, economic, and legal implications of AI.
4. We believe that AI research and development efforts need to be actively engaged with and accountable to a broad range of stakeholders.
5. We will engage with and have representation from stakeholders in the business community to help ensure that domain-specific concerns and opportunities are understood and addressed.
6. We will work to maximize the benefits and address the potential challenges of AI technologies.
7. We believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.
8. We strive to create a culture of cooperation, trust, and openness among AI scientists and engineers to help us all better achieve these goals.

3.2.4. Data Ethics Decision Aid (DEDA)

Dataschool Utrecht, in cooperation with data analysts from the city, developed DEDA [15] to help data analysts, project managers and policy makers to recognise ethical issues in data projects, data management and data policies. DEDA is comprised of several aspects.

Collection of data: The algorithm itself should be clear and explainable and the data used should be of good quality.

Usage of data: Attention should be paid to anonymization of the data, and (if applicable) how to visualise the data.

Storage of data: This aspect concerns accessibility and reusability of the data.

After these three aspects, the framework goes on to more general considerations.

Responsibility: This consideration is not only about responsibility of protocols and communication, but also about prevention of discrimination.

Transparency: It should be evaluated how transparent one can be.

Privacy: DEDA takes a more legal approach to privacy by requiring conducting a Privacy Impact Assessment and having contact with the organisation's privacy officer.

Bias: Questions about the representativeness of samples should be addressed.

Informed consent: Questions like "Can people opt-out?" and "How are they informed?" should be addressed.

The DEDA framework is more extensive than the one described above, because it also includes parts that are more related to legal aspects illustrating that legal and ethical concerns are related.

3.2.5. Artificial Intelligence Impact Assessment (AIIA)

AIIA is a relatively well-known framework for the use of AI proposed by the ECP "Platform voor de InformatieSamenleving" [16]. It is similar to a Data Protection Impact Assessment (DPIA), which is sometimes obligated by organisations to perform. AIIA consists of eight steps to decide on whether AI would have a good use in an organisation. Most attention is given to ethical, privacy, and security considerations. First, in step one, it is questioned whether the AIIA should be performed at all. If this is the case, step two asks about general information about the AI application, such as its goals. If these are completed, in step three the benefits should be described, for the organisation, individuals and society. The fourth step is to check whether the goal is ethically and legally justified. Step five is to make sure that the application is reliable, safe and transparent. Step 6 is about whether the AI application is ethical, the benefits and the identified risks must be considered jointly to assess the AI. Step seven ensures everything is documented and it is known who is responsible for what. Finally, in step eight, it is ensured that the impact assessment is reviewed periodically.

3.2.6. Other Ethical Frameworks

It should be noted that there are certainly more frameworks, such as the Asilomar AI principles [17] and The Ethics Guidelines for Trustworthy Artificial Intelligence [18]. We do not discuss them here, because they do not, in our opinion, add new insights compared to the frameworks already mentioned. Not discussed either because of the practical approach of our research but nonetheless interesting for a more philosophical view on moral dilemmas regarding AI is the report by the European Group on Ethics in Science and New Technologies [19].

3.3. Related Research on AI Issues

The ethical frameworks refer to several different ethical issues such as bias and privacy. Although it is beyond the scope of this paper to provide an overview of scientific literature on these ethical issues, we provide a brief overview here of surveys and other important papers on these ethical issues.

One prominent issue is *privacy*. Interesting surveys are [20,21] which emphasize that privacy is not only an issue for learning algorithms, but that learning algorithms may also be used to help protect privacy as well. The different kinds and purposes of AI ask for their own specific perspectives on privacy protection [22]. An important warning is issued by [23] that any attribute can be identifying in combination with others, hence that a separation between personally identifiable (quasi-identifiers) and non-identifying attributes is meaningless. They propose that differential privacy is a promising solution direction [24], where data are more aggregated. Another interesting solution direction is federated learning: instead of sending the data to the model, the model is distributed and sent to the data thus avoiding certain privacy risks [25,26].

The issues of beneficence and non-maleficence are not specific to algorithms. It has been argued early on that in the design of computer technology we should pay attention to human values while designing systems as understood from an ethical standpoint [27,28]. An interesting survey is [29] which discusses the relationship between AI and human rights, sustainability and beneficence in general. A commonly overlooked issue with advanced machine learning methods is their energy consumption: model training and development likely make up a substantial portion of the greenhouse gas emissions because of excessively large models used, for example, in natural language processing [30]. An important solution direction is to use special hardware designed for energy-efficient learning [31,32].

A well-studied AI issue is algorithmic bias: Outcomes may be skewed by a range of factors and thus might be considered unfair to certain groups or individuals [33]. For example, learned algorithms are not perfect—they also do make mistakes just like humans—but they should not be more error-prone for one group versus another. Being trained on natural human language, learned algorithms in natural language processing exhibit similar biases and are equally prone to discrimination as humans are [34,35]. Sources of bias can be found in the data, the annotation process, input representations, models, and the research design [36]. Removing such biases, i.e., obtaining truly fair algorithms, is still an open problem (see, for example, [37]). Note that AI can also be used to battle human bias and discrimination and other maliciousness humans cause, for example, in detecting discriminating posts and fake news.

The AI issue that is gaining more traction in research is about transparency and explicability. It is well-known that this issue is one of the main obstacles for adoption of AI in high-stake domains such as healthcare [38]. Part of the research on transparency and explicability was sparked by the European GDPR legislation which states that subjects have a right to an explanation [39]. The field of Explainable AI (XAI) actively researches how to explain the outcomes of learned algorithms [40]. Most approaches, however, provide post hoc explanations, i.e., attempts at explaining a model or outcome after the model has been learned in an interpretive manner. An important new direction in this area are the intrinsically interpretive models that develop machine learning methods that by design are explainable [41,42]. To be able to communicate well to users of algorithms, consumer

labels are envisioned to, for example, enable lay persons to assess when human intelligence should overrule artificial intelligence [43].

3.4. Brainstorm with Experts

The framework constructed from the AI4People framework supplemented with aspect from the other frameworks we analyzed, was used as a basis for a discussion with various AI researchers and developers. The brainstorm had a dual rule: in part it validated the constructed framework, but its main purpose was to fine-tune it. Three main changes were made as a result of the brainstorm:

- It was added to the framework that the algorithm should be fair. Although it can be considered as implicitly part of ‘the algorithm respects the rights and interests of all parties impacted’, it was found so important as to include it explicitly.
- The term “AI” was replaced by the term “algorithm”. It was discussed that applying a simple algorithmic function raises similar ethical concerns. An “AI” is a more advanced and more complex algorithm, but the core of the concern lies in the fact that the computer takes over (parts of) the evaluation and decision making process. The term “algorithm” was chosen as the more general term which would not exclude the ethical evaluation of non-AI decision making support.
- The wording of some of the detail questions in the framework were adjusted. For example, ‘as little privacy infringement of people’ was changed to ‘as little *and justifiable* privacy infringement of people’.

4. Results

4.1. Proposed Ethical Framework

The proposed ethical framework can be found in Figure 2. It is constructed from the AI4People framework supplemented with aspects from the other analyzed frameworks and a brainstorm with experts. Besides striving for more completeness, we also added more detailed questions for applicability in practice.

The main use of an ethical framework is as a checklist for organisations. A more detailed checklist was created as well and can be found in the appendix of the thesis on which this article was based [44]. It is not an aim to answer all questions with a clear yes, but to become aware of ‘no’ or ‘yes but’ answers. This allows one to analyse the issue and explicitly make design choices that are thought through and discussed. It is, for example, hard if not impossible to make an algorithm completely unbiased, so it is important to design preventive and compensating measures into the system. Considering all aspects of the framework and documenting them properly can achieve robustness against external scrutiny.

What is ethically ‘good’ is not absolute. Ethics are in part culturally defined and viewpoints may change over time. An ethical framework should therefore be open to adaptation and evolution. An organisation could, for example, decide to add an aspect or question. It is more important to discuss and carefully formulate an organisation’s viewpoint for important ethical concerns, than to aim to ‘solve’ all such concerns. Design decisions and their justifications should be documented. It is also important to re-evaluate the viewpoints from time to time to see whether they still match the organisation, society and other (implicit) rules.

Figure 2. The proposed ethical framework.

In the following part, we describe all aspects and their questions in detail.

4.1.1. Beneficence

According to Floridi et al. [12], a principle that can typically be found at the top in the various frameworks, although expressed in different ways, is that AI technology should be designed such that it is beneficial to humanity. What exactly is considered ‘beneficial’ or ‘good’ is not set in stone. There are many theories on what is good, for example, utilitarianism aims for a best outcome for a majority of affected citizens, while virtue ethics is based on the question “What would a ‘good’ person do?” (as in DEDA). The organization should define as clear as possible what it views as beneficial to humans and choose an ethical theory to follow, so that its view can guide decisions for all aspects.

Foundations for human beneficence can be found in human rights declarations and conventions. As the interpretations may change over time and since laws are not always inherently good or may not inherently lead to human well-being, these declarations and conventions must be further evaluated and elaborated upon (see also Explicability in Section 4.1.5).

In the example case of social security assistance, beneficence is about helping people regain their self-sufficiency and independence and about the fairness of the social security system. It is important to look beyond economic benefits, such as to subjective well-being [45]. An illustrative example of this was given by one of the law enforcers involved in the municipality of Amersfoort case: “If a recipient of unemployment benefit is committing fraud with Marktplaats (Dutch E-Bay red.), i.e., realizing as an unemployed person a trade volume more like a business than like for private exchange, hence generating a steady income, the straightforward way of handling such fraud, is terminating the unemployment

benefit and giving a fine. The beneficence of a system with such a rule is sub-optimal as it does not necessarily help these people. It might be better for these clients to provide support in creating their own business (registered at the chamber of commerce)", so that they legitimately do not need the unemployment benefit anymore.

Another important side to this aspect is the planet's well-being. For example, any benefits for the environment should be discussed explicitly.

As a summary, our framework incorporates the question regarding "Well-being if humans, society and planet put upfront" as a means to provide more detail for effective practice application of the framework.

4.1.2. Non-Maleficence

The other side of the coin is that AI technology should also be designed such that it does no or little harm. A special point of attention regarding AI technology is personal privacy. There is always a risk of invasion of privacy when person data is being used. The key question is whether these risks are acceptable given the purpose. According to Floridi et al. [12], it is not entirely clear whether it is the people developing or using AI or the technology itself that should be made not to do harm. Here, both are taken into account. The aspect of non-maleficence is detailed with four separate questions described below.

As Little and Justifiable Privacy Infringement of People

The purpose of this question is to ensure an aim for minimality. Even though certain data may legally be used, that does not mean all allowed data should be used. All data used should have a relevant contribution, be proportional to the goal (impact should be justifiable), and no data should be used that can excessively hurt people. Whether data has a relevant contribution is not an absolute: is 0.5% gain in accuracy sufficient relevance or should it be more than 1%? Design decisions regarding such thresholds should be motivated and documented. Furthermore, a good question to ask is whether the same goal can be achieved in different way, for example, using different less-sensitive data.

For learning algorithms, a training set is needed. The question arises whether it is fair to the people included in the data set that their data is used for training the algorithm, while, most probably, the data set used for training is contains old, possibly outdated, data. Furthermore, people included in the training data no longer need to be part of the target group of the algorithm. There exist preventive techniques for such problems. For example, a limited data set can be used to determine a statistical distribution with which a larger representative training set can be generated which satisfies this distribution, hence contains the same patterns for the algorithm to learn. Anonymisation or pseudonymisation of data are other useful techniques. A solution with the least privacy infringement is preferable.

Being a false positive or negative can have profound impact, for example, being marked by an algorithm to be a potential fraudster while you are not. The design of the whole socio-technical system should be resilient against such often unavoidable errors of the algorithm. In an application such as fraud detection, it is important that the design includes safeguards and procedures guaranteeing that human inspectors are involved whose responsibility it is to properly investigate and to make the final decision (possibly against the outcome of the algorithm). The subject must be protected as much as possible against unnecessary or invasive investigations (e.g., a house visit or questioning of neighbors can be very humiliating). See also Section 4.1.4, Justice.

Future developments should also be considered, for example, what can be considered anonymous now might not be 20 years from now, or in combination with other (open) data sets of background information. DEDA advises to ask the question "Can you imagine a future scenario in which the results of your project could be (mis)used for alternative purposes?"

Finally, in governmental applications it is advised to conduct a DPIA (see Section 3.2.5) for better understanding of data and goals. Moreover, it is advised to involve a data privacy officer if one is available or can be externally arranged.

No Misuse Can Be Made of the Algorithm

According to Floridi et al., [12], there are three kinds of misuse:

Accidental misuse (overuse): An example of accidental misuse is when an algorithm is used for a different goal than for which it is designed. To prevent accidental misuse, clear boundaries and protocols should be in place specifying what the algorithm and humans are allowed to do and how they should interact. For example, it should be clear under which circumstances humans are allowed to intervene or question the algorithm.

Deliberate misuse (misuse): To prevent deliberate misuse, sufficient security measures should be in place.

Underuse: To prevent underuse, it is important to get important stakeholders to see the importance of the algorithm.

No Misuse Can Be Made of the Algorithm: Now and in the Future

As mentioned before, the future may bring developments such as new technology, new data, or newly found security vulnerabilities that may provide new opportunities for misuse. Therefore, it is important to anticipate certain developments and to establish a protocol for regular re-evaluation.

No Misuse Can Be Made of the Algorithm: Algorithm Is Secure

An important vulnerability of an algorithm is its data. Sensitive data may be attractive for exploitation by others, so it is important to secure the access to the data and to possibly anonymize or encrypt it. The threat not only comes from outside: also users of the algorithm may misuse the data (accidentally or not), hence it is important to provide access to only the data that they need and/or by logging access to data. Furthermore, the algorithm itself may be the target of misuse: people (from inside or outside) may, for example, attempt to replace the algorithm with an altered one.

No Misuse Can Be Made of the Algorithm: Algorithm Will Be Kept up to Date

As circumstances may change, in the organization or with the subjects of the algorithm, it is important to keep it up-to-date. First, the data itself needs to be kept up-to-date. Protocols for this purpose should be in place, such as a regular request to subjects of the algorithm to update their information. Incorrect data poses a risk that the algorithm produces incorrect results leading to incorrect decisions. Second, for a learning algorithm, it may need regular retraining with training data that supplements recent cases. In general, data quality and the algorithm's behavior should be monitored. Furthermore, all protocols and monitoring should be documented.

Algorithm Is Developed by Skilful, Rational and Responsible People

Developers of an algorithm and the designers of the entire socio-technical system have a significant responsibility. They should be sufficiently skilful as to not make mistakes with, for example, security measures. They should also be rational to be able to make objectively justified decisions and to document these properly. They should also be responsible being aware of potential risks and threats and feeling a duty to prevent them. Defining a code of conduct may be desirable. Finally, thorough testing and validation should be standard. Standards and checklists for the development and use of algorithms have been or are being developed. It is of utmost importance that these are known and are applied. For example, many algorithms developed for diagnosis and prognosis of COVID-19 reviewed by [10] were found to be not usable for clinical use, which their developers could have determined themselves if they would have used the CLAIM checklist [46].

Only Professionals Can Access the Data Bound by Transparent Protocols

Only professionals should be able to access data and only on a need-to basis, especially data in decrypted form and keys needed for decryption or coupling of data sources. The IEEE Global Initiative on Ethics of Autonomous and Intelligent System states that access should be granted on a case-by-case basis [45]. As a preventive measure, the group of people with the ability to access the full data should be kept to a minimum. Access permissions should be well-document as well as the protocols governing the provision of permissions. IEEE Global Initiative also recommends that a regulatory body oversees the entire process of using and maintaining algorithms [45].

4.1.3. Autonomy

Algorithms should not have the autonomous power to hurt people. Humans must be able to always stay in or regain control. It does not mean that an algorithm should never hurt a human—hurting people may be involved in stopping crime, for example—but this should not happen by sole decision of an algorithm. At the same time, humans are known to be rather weak in making good rational decisions, which an algorithm could be better at [47,48]. So, allowing humans to deviate too much from an algorithms result may also not be desirable. Nevertheless, a human should be responsible, should have some form of choice, and should be instructed to perform checks and/or investigate situations with considerable uncertainty.

Furthermore, it is usually beneficial that the subjects of an algorithm receive a right to know that they are a subject, a right to challenge a decision based on an algorithmic result, or even a right to opt out or decide for themselves. Note, however, that how noble these rights may sound and how attractive it may seem to lift them to general principles that should always hold, there are circumstances that call for protocols violating such principles. For example, for the purpose of fraud detection with social security, citizens should not be allowed to choose whether they may be investigated or not: it is by law determined that they should be. It may not even be desirable to let them know as that may incite counteractions that may threaten the investigation. Nevertheless, subjects should have some rights to protect themselves, such as a right to an explanation (see Section 4.1.5, Explicability) or to challenge a decision. Furthermore, protocols could be designed that ensure that subjects are informed at defined moments in time and explained their rights [12].

4.1.4. Justice

(The development of) algorithms should promote global justice. The foundation for this can be found in the laws and regulations. Note again that laws and regulations need not always be ethically good. In the detailed questions below, we highlight the main aspects for obtaining global justice when algorithms are involved.

Algorithm Respects the Rights and Interests of All Parties Impacted

It is important to realize that the rights, interests, and priorities of different parties may be different but should nevertheless all be respected. For example, a data scientist may be mainly interested in making the ‘best’ algorithm, a tax payer may be mainly interested in a fair social security system, while a benefactor of social security may be mainly interested in their privacy. All such interests and priorities should be properly weighed in the design of the whole socio-technical system.

Algorithm Respects the Rights and Interests of All Parties Impacted: Algorithm Is Fair

An important aspect here is fairness of the algorithm: the algorithm should have no bias and should not discriminate. This is laid down in the GDPR, but it may require substantial effort to achieve. Examples abound of (usually unintentional) bias of algorithms. For example, a well-known case is from Amazon who attempted to use an algorithm evaluating CVs to reduce manual effort in recruitment. By using historic training data of employees of Amazon, they unintentionally taught the algorithm that the established employment

patterns are the desired ones, which is not what they intended. For example, women are under-represented in technical positions, but it was not their intention that it should stay this way, i.e., that the algorithm should discriminate against them. Floridi et al. [12] mention that an algorithm has the potential to eliminate past discrimination in the long run, if implemented properly. Methods and tools exist for checking for bias, for example PROBAST [49] which has been effectively used to show that of the 62 algorithms for diagnosing or prognosing COVID-19 passing other quality criteria, only six had a low risk of bias whereas the others either had high (45) or unclear (11) risk of bias [10].

It was discovered in the Amersfoort case that a similar risk was also present in their data. In the past, they used a project-based work practice based on selecting a theme and then within the project specifically investigate the subjects related to the theme. For example, a project checking up on people who did not request special social assistance for more than 3 years or people of whom (one of) the parents are unknown. An algorithm using historic data may detect these patterns and erroneously use them for new cases thereby discriminating against people who do not know how to ask for special social security assistance.

Another example was given by one of the law enforcers in the Amersfoort case. It is a fact that people with a non-Dutch origin have a house abroad more often. Although having a house abroad may in general be an indicator for monetary gains of fraud or crime, using this indicate may risk discrimination against citizens with a foreign origin. So, not only ethnicity is not a desirable indicator, also proxies for ethnicity may not be desirable as well. Additionally, very subtle injustices may happen. To illustrate from a personal story of a reputable person, he investigated after having been denied a credit card several times and it turned out that due to an algorithmic decision, their address was blacklisted primarily based on overdue payments by someone living in their house. The rule has some logic to it: people living together on one address are typically families and if one has payment problems, they tend to all do, but it is surely not generally true, hence may unjustly deny some people a credit card.

Preventive measures include identifying discriminating indicators or variables, including proxies for them, and exclude these from the (training) data. Potential discriminating factors should also be documented [50]. As mentioned before, training data could be generated from a distribution: it is potentially possible to correct a distribution for historic discrimination and then generate the training data from the corrected distribution.

DEDA provides an approach for predicting potential bias threatening justice. A comparison could be made between the outcomes (human) project members expect for certain cases and the decision of the algorithm.

Algorithm Is Developed with Consultancy of Important Stakeholders

In software engineering, it is standard practice to perform a stakeholder analysis and involve the most important stakeholders in the development. This is as important if not more important when developing algorithms, for example to gain trust in the algorithm. For our example case of social security fraud, the interests of the important stakeholders lay far apart: the municipality would like to strive for highest possible accuracy which is obtained by using as much data as possible while the citizens would like to have the least privacy infringement possible. It is advisable to involve a client counsel representing all citizens in this case. Concerns, priorities, requirements should again be documented carefully.

Algorithm Is Accurate

Perfect accuracy is quite possibly never reachable. Therefore, an algorithm makes to a certain degree errors. It is common practice to estimate the accuracy of an algorithm by conducting experiments with a test data set of cases the algorithm has not seen before, i.e., cases that are not included in the training set.

Knowing that an algorithm makes errors, the whole socio-technical system should be designed with compensating measures. In our example case of social security fraud,

there are mainly two kinds of errors: a false positive where an innocent person is suspected of fraud, and a false negative where a guilty person is not detected. As mentioned earlier, being marked and investigated as a suspect can have a profound impact, so it is important to explicitly consider a positive only as a *potential* fraudster and carefully design procedures to protect the suspect as much as possible against negative consequences. The other kind of error can be mitigated, for example, by, besides investigating the cases the algorithm marks as suspect, to also investigate a random selection of cases or a selection of cases the algorithm determines as borderline. Note that this infringes their privacy and cause some harm to them, hence the trade-off with addressing the false negative problem should be carefully weighed and documented. In this way, also another problem is mitigated: the problem of fraudsters always being a step ahead and creatively finding ways to avoid detection by the algorithm. Finally, an organization using an algorithm is ultimately responsible for its own decisions also when these are based on outcomes of algorithms [50].

Algorithm Is Effective and Efficient

Being effective means that the goal is best reached with this particular algorithm. Being efficient means that the resources involved are justified by and optimal for the goal. Therefore, it is advisable to define a benchmark with metrics for effectiveness and efficiency, so a comparison can be made between alternative solutions. One should first consider the alternative of not using an algorithm, for example, whether it is more effective and/or efficient by hiring more human capacity. Additionally, alternative algorithms using for example different sets of variables can be evaluated for effectiveness and efficiency. According to the IEEE Global Initiative on Ethics of Autonomous and Intelligent System, focus must be placed on how to interpret the metrics for effectiveness and efficiency [45].

In our example case, the goal is to reduce fraud. However, this comes at the cost of privacy infringement, IT infrastructure and maintenance, investment into development and maintenance of the algorithm, human resources needed, and so on. Imagine that after the first run, the algorithm identifies 400+ potential fraudsters. It may be impossible to properly investigate and prosecute all. Strategies should already be in place for achieving efficiency such as capacity for 30 investigations per quarter (or a top 25 and 5 randomly selected ones). The benchmark should include metrics like number of investigations and fraudsters discovered to allow evaluation of efficiency and effectiveness, respectively. One should also be aware that all this is paid by public money which creates an extra responsibility to carefully evaluate the resources used both in terms of money and manpower. Furthermore, a possible environmental impact should be considered.

4.1.5. Explicability

One should be aware that a small group of people can create an algorithm that impacts the lives of many. We already established the importance of involving all important stakeholders in the design including representatives of the affected many. Therefore, explicability to all is of utmost importance. Everyone needs to know what the algorithm is and does, so all can be convinced that it is beneficent and does no harm [12]. Furthermore, responsibilities, procedures and regulations should be clear to all. In our opinion, other frameworks were under-developed on this aspect, which is the reasons we added it explicitly and worked it out with several questions.

Who the Stakeholders Are and How They Are Affected Is Clear

According to the AIIA [16], it must be clear who the stakeholders are, what their position is and what their expectations and wishes are. This information can be obtained by performing a stakeholder analysis. Included should be parties and organizational roles involved in the development and design of the algorithm, roles involved in its use and the handling of its outcomes, roles involved in its technical functioning and maintenance, roles involved in its management and governance, people whose data is processed by it or

otherwise affected, and everyone who directly or indirectly benefits, may be harmed, or provides for some necessity.

The more indirect stakeholders are often not easy to identify hence risk to be forgotten. To illustrate, for the Amersfoort case, it is the Dutch taxpayer who ultimately pays for the algorithms and the whole Dutch society benefits from a well-functioning anti-fraud system. Additionally, the group of people affected is often larger than one may at first think. For example, Wynsberghe et al. [13] describe an algorithm meant to find social media accounts of a given list of people, but not only these people were effected by the algorithm. In order for it to search and resolve ambiguities, it invaded the privacy of many more people who just happen to have a similar name or address by also monitoring their social media activity.

How stakeholders are affected must be addressed from a broad spectrum of viewpoints: ethical, social, economic and legal. In the SyRI case, for example, it is clear that even though the SUWI decree allows the use of certain data, people felt that their privacy and autonomy was wrongfully impeded, so strong as to start a campaign against it. Just adhering to the law does not suffice; thought must go into the other viewpoints as well for additional requirements and desires. Furthermore, again, since the law and the other viewpoints may change over time, the requirements and desires should be re-evaluated regularly.

Who the Stakeholders Are and How They Are Affected Is Clear: The Context Is Clear

For being able to make a proper value trade-off between the requirements, benefits and concerns of the various stakeholders, the context should be clear. It matters whether personal data are used for marketing purposes or for fraud detection [13]. Furthermore, it also matters where the data goes and which assurances are active. To illustrate the latter, Sohail et al. [51] describe the general health case practice of sending blood and tissue samples *with accompanying data* to labs that are independent organizations without proper assurances. If the context changes or if a boundary is crossed to a different context, everything should be re-evaluated with this other context in mind, such as DPIA (see Section 3.2.5), ethical and legal checks.

Algorithm Is Transparent

According to the IEEE Global Initiative on Ethics of Autonomous and Intelligent System, transparency of an algorithm means that one should be able to determine how and why an algorithm came to a certain outcome, for example, why someone is considered a fraud risk [45]. There are three reasons why transparency is important according to the Dutch Ministry of Justice and Security: gaining trust, facilitating verification and help, and it can lead to improved adherence to rules and regulations.

The field of 'Explainable AI' develops and studies methods and techniques for being able to explain the reasoning and outcomes of algorithms, notably learning algorithms. There are two main categories: post hoc explanation and transparency by design. The former takes the algorithm (mostly) as a black box whose and attempts to explain a certain outcome (local explanation) or full reasoning of the algorithm (global explanation). The latter provides for the development or learning of an algorithm whose reasoning is directly understandable (white box), which typically involves the loss of some accuracy compared to black box algorithms. Kahneman shows in their book "Thinking, Fast and Slow" that humans are not good at making unbiased objective decisions and already very simple easily explainable algorithms such as weighing several independently assessed criteria can already significantly improve their decision making [48].

Note that full transparency to all stakeholders may not be desirable. For example, knowing the full reasoning of a fraud detection algorithm may give fraudsters an undesirable opportunity to find ways to prevent detection. Transparency to some stakeholders is, however, important for verifying the correctness of the outcomes of the algorithm and whether or not it is biased in some way. Furthermore, not every part of the reasoning of an algorithm needs to be transparent; it may suffice that it is testable. As an analogy, we do

not understand fully how the eye and brain allow humans to recognize objects and faces, yet we trust a witness when they have seen something or someone, because through test and time we know the reliability of our visionary system. Additionally, the training process of a learning algorithm can often not be explained, because it is similarly ill-understood, but also here we trust it through rigorous testing and validation. The comparison with human learning, recognition and reasoning also makes clear that alternative solutions without the use of algorithms and based solely on humans are not fully transparent nor devoid of bias and error either. Therefore, one should not over-criticize algorithms, but strive for improvement of transparency and performance compared to a fully human-based solution. The Dutch Ministry of Justice and Security states that the focus of information provision should be on describing the goal of the algorithm, which criteria and rules the decision was based on, and the sources of data that are used including their quality and how they are combined [50].

Transparency not only pertains to the algorithm itself, but also on the rest of the socio-technical system. The protocols, procedures, rules, and regulations should also be transparent. It should be clear who can access certain data, who is responsible and for what (for example, when a data breach occurs) and what should be done when something goes wrong. It is also advisable to have exit and change protocols and protocols for what to do when the algorithm produces unexpected outcomes. Note that a well-thought through and transparent design of the socio-technical system the algorithm is embedded in, can be very instrumental for its acceptance. It became very clear in our analysis of the SyRI case (see Section 5.1) that the public outcry was mostly directed not at the algorithm itself but at the flaws and intransparency of the rest of the socio-technical system.

Algorithm Is Transparent: Users of the Algorithm Can Explain the Algorithm

As an extension to the ‘Who the Stakeholders are and how they are affected is clear’, the users of an algorithm should have sufficient understanding of how the algorithm works and why it produced a certain outcome. They have the responsibility to monitor that it keeps behaving properly. It is advisable to educate the users of an algorithm on how much they (should) rely on the algorithm; they should be aware of how dependent they are on it. It is a risk to become too dependent and too reliant.

Algorithm Is Transparent: Stakeholders Are Educated on Purpose and Functioning of Algorithm

According to Partnership on AI, the public must also be educated and should have a say [8].

When talking about social security fraud, the public is always a (indirect) stakeholder. As already established, full transparency is not desirable, because it would make it easy for citizens to prevent being caught for fraud. Therefore, an intermediate level of transparency should be sought, where stakeholders are informed and educated to the maximum extent without being able to circumvent detection. Included in this information should be what is the goal of the algorithm, why a certain type of algorithm is used, which data are used, how often the algorithm will check this data, what consequences may happen, who is responsible for the analysis and which quality checks happen [50]. If decision trees or decision rules are used, it must be clear why a certain threshold was chosen. Someone should be responsible for communicating with and to the general public and other stakeholders. This should be done in a concise, understandable and easily accessible manner. According to the Ministry of Justice and Security [50], the public must be informed via the municipality’s website about:

1. The fact that the municipality performs data analyses
2. Why they do this
3. What consequences may occur for affected citizens
4. Whether or not machine learning is used and an explanation of this
5. What the legal basis is
6. Which data sources are used

7. What the role of third parties is in the process
8. What checks on quality are performed
9. If there is a human intervention in the process
10. Which assessment frameworks are present and how these are used

To illustrate the thin line between an innocent algorithm and a not-so-innocent one, let us do a thought experiment. Suppose you are taking a shower and your Roomba cleaning robot would come in to vacuum the bathroom floor. Would you mind? Probably not. Suppose now this Roomba has a camera to more accurately see the walls and other bumps. Would you mind now? What if the camera would back-up to the cloud? Note also that these questions are not specific to algorithms. What if a blind and deaf man walks into your bathroom? What if the man is not blind and deaf, he walked into the bathroom on purpose to see you showering and to share this information, but when he runs outside, he gets run over by a truck and dies, so he could not share the information. Would it still be an invasion of privacy? This shows that much depends on information received, how this information is received, what the goal is and how sensitive the information is.

Another stakeholder who may need to be informed and educated are the employees of the organization even beyond the users of the algorithm. Any feelings of unease may obstruct a project aimed at developing an algorithm or may send undesired signals outside the organization. Therefore, it is important to discuss quite broadly (see also Section 3.2.3).

Algorithm Is Transparent: Citizens Are Well Informed When Researched by the Algorithm

People should be informed when an algorithm is inspecting their data and especially when the outcome of the algorithm may have consequences for them, for example, when they are marked as a potential fraudster. This could be done periodically with an e-mail or letter, as also enforced by the GDPR. It could be advisable to inform them more often, for example, periodically or when something relevant changes. Moreover, they should be informed when their data is at risk or when it is looked at more extensively than normal. According to DEDA (see Section 3.2.4), they should definitely have the right to object against the outcome if an algorithm, especially if the outcome results in a decision affecting them negatively.

Algorithm Is Accountable

Finally, accountability for the outcomes of the algorithm is important. It should be clear who designed and developed the algorithm and who is responsible for the way it works [12]. This includes the responsibility of choosing the data that are used, for how the algorithm works, and for what to do in case something goes wrong. It also includes the responsibility of creating protocols and other peripheral matters. Additionally, included should be why the algorithm was created in the first place, why certain data were chosen, and who can access which data (and why). It should even be known who is ultimately responsible for the entire project. This should be documented carefully.

5. Validation

The goal of this research was to come to a more complete ethical framework with added detail questions for improved applicability in practice. We therefore validate it mainly on the aspect of completeness on two real-world cases:

- SyRI: The “Systeem Risico Indicatie”. This algorithm-based system for fraud risk assessment evoked much public uproar in The Netherlands, parliamentary discussions were held about it, and a court ruled that it was against the European Convention on Human Rights (ECHR) [52].
We chose this case, because substantial information is publicly present from the media and the court ruling to answer the questions of the framework.
- Amersfoort: The fraud detection algorithm under development by the municipality of Amersfoort, one of the partners in this research project.

For SyRI, we analyzed the publicly available information from the court case and what was written about it in the media on the lookout for ethical concerns. For Amersfoort, we did the same based on information from the media and interviews with people designing and developing the system. Based on the information, we attempted to answer all questions of the framework and then compare what was raised with what the framework asked for as a means to assess the framework's completeness. As a concrete metric, we counted the number of concerns/questions raised/answered. Since it turned out that from the court case, in the media and interviews nothing was raised that was not also a question within the framework (our main validation result), we present the results as counts of questions answered, partially answered, not answered, and not answerable (due to lack of information).

5.1. SyRI

5.1.1. Analysis of the Court Case and Media

We present the analysis guided by the main aspects of Figure 2.

Beneficence It is clear why SyRI was developed in the first place: the benefits are to stop fraud from happening and to (re)gain trust from society in the system. The amount of money lost is substantial. Achieving this goal with the use of an algorithm seems like a good way to go (see also justice), because of its effectiveness and efficiency. Some mentioned lapses in beneficence could be found, namely that the SyRI law was established with SyRI already in mind.

Non-maleficence Due to the SyRI law, practically all personal data that can be thought of are allowed to be used. The contribution of the various data attributes to the decision making process has not been made clear, hence their necessity for the purpose as justification for the privacy infringement is unclear. Important to realize is that not directly suspected citizens are also analyzed by the algorithm, since it works on neighborhood level. It has been clearly established that the impact on someone being falsely accused is huge. The SyRI algorithm had been subjected to a DPIA which was reviewed by independent parties.

Autonomy People marked by the system could not oppose this; they were not even informed of the fact.

Justice SyRI was much criticized on the aspect of transparency (also an issue of explicability). A large concern is that there is practically no information available on the development process of the algorithm. It is unclear how it had been developed including things like whether the interests of all stakeholders were taken into account or not. Another major concern was that there was no proof available that it would actually work. Two cases were made public, of which one had only false positives. There are, however, similar projects that have been proven to work.

Explicability First, it is unclear whether or not a stakeholder analysis had been performed or not. Additionally, other important information on this aspect is unknown. As explained before, full transparency is not desirable with fraud detection. The mistake made by SyRI is that this non-transparency was carried too far; the aspect of educating the public was ignored. There was only limited public information available and it was unclear where to go for more information. People marked by the system were not informed of this. People were sent an informative letter, but it did not make clear when and how often would be used within a two year period of time. Looking at it from the perspective of the proposed framework, it is quite ironic that half of the questions that could not be answered were under explicability (see Section 5.1.2).

In short, the goal behind SyRI is clear and good. There is something to say for not having full transparency in this case. There are quite some parts of SyRI, however, that were not worked out well. These were mainly the amount of data used, the accuracy of the algorithm, the impact on someone when they are a false positive, the transparency of the development phase and the education of the public.

5.1.2. Comparison of the Analysis to the Framework

A summary of the completeness results on the comparison of the analysis of the court case and media on SyRI with the proposed framework can be found in Figure 3. More information on the topics can be found in Figure 4. The topic of pseudonymization/anonymization was omitted from the table, because it is in most of the cases too ambiguous to ascertain which of the two is meant. Note that the intended use of the framework is *during* the development of an algorithm and not afterwards, but the amount of information available from the court case and media attention made it very suitable for validation. The difference to be noted is, that if the framework would have been applied by the organization that developed SyRI itself, more (inside) information would have been available to them.

	✓	□	✗	Not answerable (□)
Beneficence	4	1	0	0
Non-maleficence	10	10	3	9
Autonomy	0	1	1	1
Justice	4	4	4	3
Explicability	11	5	9	11
Total	30	22	17	22

(a)

(b)

Figure 3. Results of a comparison of specific ethical concerns between the framework, the court case and the media on whether questions in the framework could be answered with a yes, maybe, no, or were not answerable. (a) Exact numbers. (b) Percentages.

Topic	In framework and not in analysis	In analysis and not in framework	In both framework and analysis	Partial overlap	Not answerable
Reasoning behind SyRI			✓		
SUWI legislation mentioned as issue			✓		
A lot of personal data present			✓		
Impact on subjects			✓		
Reviewed by independent parties				✓	
Unclear how SyRI was created			✓		
Lack of accuracy			✓		
Lack of transparency			✓		
Lack of education to the public			✓		
Risk cases are not informed			✓		
A lot of unclarity is present			✓		
Results of SyRI are unclear/unverifiable					✓
Bias present in SyRI					✓
Human interference should be present				✓	

Figure 4. Comparison of specific ethical concerns between framework, the court case and the media on whether questions in the framework could be answered with a yes, maybe, no, or were not answerable.

There are several discrepancies between the author's application of the framework with the analysis of the general media and court case.

- The media and the author applying the framework are of the opinion that it is doubtful that this large amount of data is necessary for the purpose of the algorithm. The court also mentioned this briefly, however, it did not judge on this.
- The issue of independent parties reviewing the system is not a black and white case. It can be established that it has been reviewed by multiple governmental people. The media seem to agree on this, however, they mentioned that many of these governmental people are not educated enough to make decisions on this. There is something to say for that, looking at the number of questions that were asked about SyRI, but nonetheless some critical questions were asked by governmental people. Competence of (certain people within) the government is often something of debate and it seems not to be different here.
- The media claimed that the results of SyRI are unclear and unverifiable. The court case does not agree with this. It was mentioned that the algorithm is nothing more than a simple decision tree, which should be sufficiently transparent for the people working with it. The author applying the framework evaluates this as a case of a truth that is probably somewhere in the middle.
- The media often emphasize that SyRI is discriminating. It is difficult to say something about this since so much of SyRI is kept a secret. The court left it somewhere in the middle and the author chose the same approach with the framework, with a mention that the decision to apply SyRI only to vulnerable neighborhoods does feel doubtful. The framework's conclusion might have been different if it was applied by a team member of SyRI with more information. For example, the question 'Are there any potential discriminating factors in the algorithm?' could very well uncover more issues besides this decision to apply it only to vulnerable neighborhoods.

5.1.3. Comparison of General Media with the Court Case on SyRI

It is interesting to evaluate the differences between what seem the most important arguments in the court case vs. the most important arguments in the media as it may illustrate differences between legal and public viewpoints. According to the court case, the purpose behind SyRI is good [52]. The court did establish that there was a lack of transparency, especially about which data are used and information provision when someone is considered a risk case. The court takes into consideration that the impact on someone when they are a risk case is huge and ruled that this consideration is not in balance with the transparency provided. The court confirmed that full transparency to everyone regarding the rules governing a decision of the algorithm was undesirable. However, when someone is considered a risk case, they ought to know why the algorithm decided they are.

The general media emphasizes in part different aspects. *Bijvoorbachtverdacht* is a good source of opinions. In one of their articles, they mention six points as main issues [53].

- It was undesirable to introduce the law surrounding SyRI while ignoring advice from important stakeholders.
- It is weird that citizens cannot inspect their data.
- The volume of data used for the algorithm is problematic, which is a point also highlighted by *Tweakers* [54].
- Data used by SyRI is not anonymized but merely pseudonymized, although it is unclear whether they consider this necessarily as a bad thing.
- The results of SyRI are unclear and not verifiable.
- The economic benefit of the algorithm is relatively low.

In another article, they explain that they think it is weird that the focus of SyRI was put on so-called problem neighborhoods as well as the large impact it has on an individual when one is being considered a risk case [55]. The lack of accuracy is also mentioned as an issue. The largest issue is that they feel that every citizen is already a suspect before there is an indication of any wrongdoing.

Computable.nl places the latter issue in another perspective by observing that social benefit checks via an algorithm like SyRI indeed makes a citizen suspicious in advance, but that speed checks on the highways and a similar algorithm of ABN Amro, a Dutch bank, do the same, without any complaints by the public [56]. According to Bits of Freedom [7], who awarded SyRI the cynical Big Brother privacy award, the algorithm would create a stigma to the neighborhoods it would be applied to. They also mention that citizens could not defend themselves. For Amnesty, the most important issues are that citizens are kept in the dark and the lack of transparency of the algorithm [57]. Interestingly, they mention that the algorithm cannot be checked, while the state mentioned during the court case that it can be. In the reactions on the earlier mentioned Tweakers article, people mainly want human interference with the algorithm, as in, the algorithm should not make decisions without a human checking these. Another topic that is often discussed in the general media, is that of the lack of education of the public as well as the users, for example, the governmental people who accepted the law. An interesting addition is done by computable.nl, who find it beneficial to have the goal communicated more clearly as well as to have information on the development roadmap.

A few things stand out in comparing the attention to different topics in the court case and the general media.

- Fewer topics have been discussed during the court case and consequently conclusions were drawn for less topics. This makes sense, because the ruling of the court has a narrower scope. One reason is that the court cannot make assumptions, whereas the media can. The court also often gives the benefit of the doubt whereas the general media would not do that.
- The reactions on Tweakers.nl were very clear about the requirements for human involvement. Although both the media and the court mentioned this aspect, the court did not judge on it.
- The main point of Bijvoorbautverdacht.nl was that citizens would be suspect in advance, which we categorized under impact on subject. From the court case, the large impact on citizens was the main reason SyRI was ruled to be stopped. The court did not necessarily rule that people were suspect in advance, they did rule that the impact was not proportionally to the goal of detecting fraud.

5.1.4. Conclusion on Validation with SyRI

For this specific case, the framework was found to work, since it covered all the flaws found by the court case and the media. These were the amount of data used, the accuracy of the algorithm, the impact on someone when they are a false positive, the transparency of the development phase and the education of the public. The framework discovered more flaws than the media and court case. The other way around this was not the case, often because the assumptions were made by the media and not by the author applying the framework. If more information would have been present, this could have been avoided, i.e., if the framework would have been applied by a designer or developer of SyRI as intended.

5.2. The Algorithm of the Municipality of Amersfoort

A summary of the completeness results on the comparison of the analysis of the media on the algorithm of Amersfoort with the application of the proposed framework by the municipality of Amersfoort can be found in Figure 5. More information on the topics can be found in Figure 6.

Figure 5. Results of a comparison of the analysis of the media on Amersfoort with the application of the proposed framework by the municipality of Amersfoort on whether questions in the framework could be answered with a yes, maybe, no, or were not answerable. (a) Exact numbers. (b) Percentages.

Topic	In framework and not in analysis	In analysis and not in framework	In both framework and analysis	Partial overlap	Not answerable
Reasoning behind using a fraud detection algorithm			✓		
Some sensitive data present			✓		
Protocols need to be updated or created	✓				
(Lack of) Communication with the public			✓		
Customers do not know automatically why they are a risk case			✓		
No periodic reminder present	✓				
Algorithm needs training			✓		
Impact on customer				✓	
Risk of being an early adapter of such an algorithm			✓		
Inaccuracy present in the algorithm	✓				
Context may change			✓		
Users are able to explain the algorithm				✓	
Human interference must be present			✓		
Which data are used and why are clear			✓		
Algorithm should not make decisions for us			✓		
It is discoverable why the algorithm made a certain decision			✓		

Figure 6. Comparison of specific ethical concerns between framework and the media.

Before starting the description of the results of the municipality of Amersfoort, it should be noted that these results are not precisely comparable to SyRI. This is mainly because 19 more questions were answerable as there were people involved with inside information. It is not reasonable to assume that these 19 questions would all have been answered positively or negatively for SyRI. Another difference is that the SyRI algorithm was already in use, whereas the algorithm from the municipality of Amersfoort is still in the development phase. This means that some aspects simply have not been arranged yet, because they were planned for a later stage. This means that some questions are marked red or orange in the framework, which will turn green when re-evaluated later.

The application of the proposed framework was done by interviewing various stakeholders involved in the design and development of the algorithm. These included law enforcers, project management, team management, information provision advisor, data scientist and lawyers. As part of the interviews, an evaluation of 12 cases was done: 10 cases were selected by a preliminary prototype algorithm and two cases were selected randomly (hence most likely non-fraudulent). The interviewees did not know which were the randomly selected cases.

We present the analysis again guided by the main aspects of Figure 2.

Beneficence: Analogous to SyRI, it is clear why the municipality of Amersfoort chooses for a fraud detection algorithm: to aid in the reduction of fraud within the social security assistance context. This, in turn, to increase trust in the government and the tax system as a whole. The mindset of the municipality of Amersfoort is that they feel that the earlier fraud is detected, the earlier people can be helped. This focus on helping rather than punishing is strongly addressing beneficence. They also aim to reduce the differences between the human law enforcers.

Non-maleficence: As with any learning algorithm, historic data is needed for training the algorithm, which is not necessarily a bad thing, but awareness of this fact should be created among the stakeholders involved. There are a few factors in the algorithm that could be considered sensitive information, such as gender and special notices, as identified by the team. It still has to be evaluated whether or not these criteria sufficiently add to the accuracy of the algorithm; if they do not, then they may be excluded from the data. As this is also a fraud detection algorithm, being considered a high risk case may have significant impact on someone. The difference with SyRI is in the protocol: when customers, as they are called, are identified by the algorithm, the first step is a desk investigation by a human law enforcer. Only if suspicion is raised by the law enforcer is the customer invited for a conversation. The impact hence is not necessarily larger than a system without an algorithm. This point also illustrates the importance of a socio-technical approach to design: the involvement of a human as explicitly designed as a protocol in the system effectively addresses an important ethical concern with the use of algorithms. A final concern under non-maleficence, is that the algorithm has not been tested on real-world cases yet, so some risk of an ill-behaving algorithm is there; but this falls under ‘no innovation without risk’.

Autonomy: The intended protocol addresses the autonomy concerns: humans stay in control. The desk investigation by the law enforcer ensures that there is an opportunity for a knowledgeable user to go against the outcome of the algorithm, even before the customer is impacted. The subsequent conversation with the customer also provides an opportunity for the customer to defend themselves by providing logical explanations for suspicious data or even counter-evidence.

Justice: Practically every algorithm is imperfect: it will produce false positives and negatives. The evaluation of the 12 cases showed that in this case, a limited inaccuracy of the algorithm has a beneficial effect: it keeps the law enforcers on their guard when investigating cases marked by the algorithm.

Explicability: Here also, the context could change. Awareness should be created with all stakeholders of this fact. Moreover, customers are not automatically immediately notified when they are considered a risk case. They are only notified after the desk investigation. The desirability of a protocol allowing them to request this (with a risk score and reasons why) still needs to be evaluated and if positive, such a protocol needs to be designed. It has been decided that law enforcers are not informed why the algorithm marks a customer as a high risk case. The reason is to keep them as objective and on their guard as possible. There are currently no protocols envisaged for dealing with extreme public outrage. Such protocols may be useful, since there is quite some resistance to these types of algorithm after what the SyRI court case. Another point that one did not realise but came to light when applying the framework,

is that the citizens of the municipality of Amersfoort currently do not know that an algorithm is being developed. They will only be informed after the municipal council has been informed. The decision by the council whether to continue with the development of the algorithm or not, lies in the future at the time of writing. Related to this is the information to the public. Obviously this also has not been worked out yet, but the intention is to publish technical documentation with extra information. The framework recommends that this documentation should include why data analysis is performed, what the consequences are for citizens, the legal basis, and an explanation of the quality assurance. Finally, customers have a duty to provide certain information to the municipality such as when they move in with a partner or when they get a new job. No reminder is sent to customers currently. There is some legal basis to not inform customers when they are investigated, including a check by an algorithm. The municipality of Amersfoort is advised to think about the ethical basis, whether it is considered desirable to ask customers periodically to check and update their data among which the criteria the algorithm will be using.

In summary, the goal behind the algorithm of the municipality of Amersfoort is also clear and good. The focus on helping rather than punishing is a commendable beneficence. Many parts have already been thought of or worked out. Concerns to pay attention to are the sensitive data that may be included in the algorithm, the protocols that need to be updated or created, that citizens do not know about the algorithm yet, that the communication to the public must still be created in a good manner, that customers do not automatically get to know why they are considered a risk case and that it might be good to remind customers once in a while that the algorithm is used. For some concerns more awareness could be created such as that the algorithm needs training, the impact on customers, the risk that is present with using a brand-new algorithm, the inaccuracy that is deliberately present and the fact that the context may change. As mentioned earlier, not every part of the framework must be green, but the intention is to make sure that all aspects receive thought and attention.

5.2.1. Comparison of the Municipality of Amersfoort's Algorithm to the Media

It is, of course, a bit more difficult to analyze the performance of the framework on the municipality of Amersfoort, because no court case took place and less media attention is present. There has been some media attention to the initiatives of the municipality of Nissewaard in developing an algorithm. This algorithm has been developed by the same supplier and is rather similar, although it makes use of different criteria. In this section, we analyze what the media said about the Nissewaard algorithm. Details can be found in Figure 6.

The Federation of Dutch Trade Unions (known in the Netherlands as FNV) is the main adversary of the algorithms being developed by both municipalities. The general media often reports of confusion. For example, it was mentioned that 'a person with three cars who receives social security assistance is weird', while it should not be possible for the supplier developing the algorithm to have this information [58]. There has also been confusion about how much fraud is supposed to be detected by the algorithm [59]. FNV is of the opinion that every customer should be informed what their risk scores are, not only those who are considered a risk case [60]. The local political party "Volkspartij Gemeente Nissewaard" observes in the media that not much has been communicated to the citizens [61]. They also mention that the technical documentation is very difficult to read. Platform Burgerrechten adds to this that they believe the algorithm to be 'random'; the municipality should be able to explain why the risk cases are a risk [62].

An additional point, not directly from the media but mentioned in an interview with the information provision advisor, was that they saw drawbacks in sentiments of this technology being new and scary and in the complexity of an evolving context. In an interview, a data scientist confirmed that these algorithms indeed still need to prove themselves.

In comparison with SyRI, a few things stand out. There is much less information found in the media than during the analysis of SyRI. This makes sense, because of the outrage around SyRI. Moreover, there are many points raised in the framework which are not mentioned in the media nor during the interviews. This makes sense in this case, since inside information was present in contrast to SyRI, which is not known to the general public (yet). Other differences are for example that citizens do not know about the algorithm, the communication to the public must still be created, the confusion mentioned, and the unreadability of the technical documentation.

Two discrepancies can be observed. First, the impact on the customer was not mentioned explicitly by the media, but it was implied sometimes [60]. The second discrepancy is that users are able to explain the algorithm. This entails that the decisions of the algorithm must be clear. According to the media, in a reaction to the municipality of Nissewaard, this was not the case. It is unclear where this information comes from and whether this is based on facts or suspicions. The framework advises that explainability of the algorithm is important. The supplier is known to have worked on explainable AI within the algorithm. Finally, a lack of communication was mentioned in the media, but this was mentioned about the municipality of Nissewaard, not about the municipality of Amersfoort. It does highlight the importance of this aspect.

5.2.2. Conclusion on Validation with Amersfoort

First and foremost, during the interviews it was mentioned several times that a question by the framework had not been thought of, hence had not been discussed before, always with an associated intention to add such discussions to the project. Hence, the intended purpose of the framework to instigate additional discussion and decisions on ethical aspects to obtain a more complete treatment of ethical aspects, seems to be confirmed. An example is the protocol for when a public outrage might occur. Furthermore, three more (indirect) stakeholders were discovered during the interviews.

6. Conclusions

In this paper, we propose a framework for identifying and assessing ethical concerns regarding the use of data-driven algorithms. The framework can be used during the design for identification of relevant concerns so that they can be addressed early in the design of the entire socio-technical system contributing to a more robust, complete, and effective introduction of such systems, improve the quality of their designs from the ethical perspective, and avoid injustice and associated uproar. The framework has been developed by merging ethical frameworks from the literature and enriching and fine-tuning it on the basis of expert interviews and brainstorm sessions. The resulting framework is based on five ethical principles that algorithms and the surrounding socio-technical system should respect: beneficence, non-maleficence, autonomy, justice, and explicability.

The framework has been validated by applying it to two fraud detection case studies: the *Systeem Risico Indicatie* (SyRI) of the Dutch government and the fraud detection system being developed by the municipality of Amersfoort. For SyRI, we analyzed media articles and the court case. We showed that all of the concerns mentioned there were also addressed by the framework, and more. Importantly, it showed that a substantial number of aspects were not or only partially addressed by the design of the SyRI system. About 24% of the questions asked by the framework could not be answered though, because of lack of inside knowledge as opposed to the Amersfoort case where only 3% of questions could not be answered. For the Amersfoort case, the framework showed that a much smaller but nevertheless still significant number of aspects were either not addressed or insufficiently addressed. However, also here, the validation showed that all concerns mentioned in the media and interviews we held, were addressed by the framework. Here too, the framework addressed more aspects that were not explicitly 'on the radar' of the designers of the Amersfoort system.

Common to both cases is that they are of high societal importance. Fraud is an important phenomenon and large sums of money are involved. Hence the reason why the solutions were or are created are more than clear. Second, in both cases the data subjects, citizens or customers, do not have a choice to participate. They cannot say that they do not want their data to be used. This makes that their autonomy is limited, hence there is need for proper safeguards and appeal protocols such that they receive sufficient control.

The framework proved to be an effective tool in both cases. Furthermore, it indeed seems useful to apply the framework early in the design phase having a diverse group of people involved and having sufficient knowledge of the design and realization of all components of the whole socio-technical system. The framework helped discover initially unidentified concerns in both cases, concerns that, were they identified in an early phase, could have been properly discussed and incorporated in the design. It can also hopefully avoid the introduction of new systems with ethical flaws and associated scandals such as happened with SyRI.

Future Work

The validation of the framework focused on two cases. One important part of future work should be focused on validation with more cases and gaining more experience with its use for the purpose of further improvement and fine-tuning. It is important to choose such cases in other important domains, such as healthcare and e-commerce, where algorithms are increasingly used and ethical concerns may have much impact as well. Furthermore, the framework is currently designed with a focus on algorithms with limited to no autonomy for the data subject. Future research could re-evaluate the desirability of context-dependent concerns like ‘stakeholders have the right to decide for themselves’ and how these should be integrated in the framework. Another possible extension is that the framework may benefit from adding specific security-related questions under non-maleficence, for example, about logging and access rights. Finally, an attempt could be made to generalize the framework to a broader range of technical inventions such as apps.

Author Contributions: Conceptualization, X.v.B. and M.v.K.; methodology, X.v.B. and M.v.K.; validation, X.v.B.; investigation, X.v.B.; resources, M.v.K.; writing—original draft preparation, X.v.B. and M.v.K.; writing—review and editing, X.v.B. and M.v.K.; visualization, X.v.B.; supervision, M.v.K.; project administration, X.v.B. and M.v.K. X.v.B. and M.v.K. stand for X.v.B. and M.v.K., respectit. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All information used in this study was found in mentioned sources.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Domo. Data Never Sleeps 6th Edition. 2019. Available online: https://www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf (accessed on 28 February 2021).
2. Doove, S.; Otten, D. Verkennend Onderzoek Naar Het Gebruik van Algoritmen Binnen Overheidsorganisaties. Available online: <https://kennisopenbaarbestuur.nl/media/255836/verkennend-onderzoek-naar-het-gebruik-van-algoritmen-binnen-overheidsorganisaties.pdf> (accessed on 4 February 2020).
3. Wong, P.H.; Simon, J. Thinking about ‘Ethics’ in the Ethics of AI. 2020. Available online: <https://revistaidees.cat/en/thinking-about-ethics-in-the-ethics-of-ai/> (accessed on 28 September 2021).
4. Raad van State. Participatiewet (NL). 2003. Available online: <https://wetten.overheid.nl/BWBR0015703/2021-07-01#Hoofdstuk2> (accessed on 1 July 2021).
5. van Geldrop, A.; de Vries, T. *Fraude Loont... Nog Steeds. Number NUR 980 in STT-Publicatie, Stichting Toekomstbeeld der Techniek*; University of Twente: The Hague, The Netherlands, 2015; ISBN 978-94-91397-10-3. Available online: <https://5dok.net/document/wq2vdery-fraude-loont-nog-steeds.html> (accessed on 24 November 2021).
6. Huisman, P. Hoe de toeslagenaffaire kon gebeuren. *Manag. Kinderopvang* **2020**, *26*, 36–37. [CrossRef]

7. de Jonge, A. SyRI Grootste Privacyschender 2019. 2019. Available online: <https://www.binnenlandsbestuur.nl/sociaal/nieuws/syri-grootste-privacyschender-2019.11606010.lynkx> (accessed on 28 September 2021).
8. Tweede Kamer der Staten-Generaal. Vragenuur (NL). 2018. Available online: <https://debatgemist.tweedekamer.nl/debatten/vragenuur-257?start=622> (accessed on 24 November 2021).
9. Shubhendu, S.; Vijay, J. Applicability of Artificial Intelligence in Different Fields of Life. *Int. J. Sci. Eng. Res. (IJSER)* **2013**, *1*, 28–35.
10. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.I.; Etmann, C.; McCague, C.; Beer, L.; et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **2021**, *3*, 199–217. [\[CrossRef\]](#)
11. Fox, W.M. Sociotechnical System Principles and Guidelines: Past and Present. *J. Appl. Behav. Sci.* **1995**, *31*, 91–105. [\[CrossRef\]](#)
12. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Lütge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [\[CrossRef\]](#) [\[PubMed\]](#)
13. van Wynsberghe, A.; Been, H.; van Keulen, M. To Use or Not to Use: Guidelines for Researchers Using Data from Online Social Networking Sites. RRI, Rict Responsible Innovation, UK. 2013. Available online: <https://research.utwente.nl/en/publications/to-use-or-not-to-use-guidelines-for-researchers-using-data-from-o> (accessed on 5 February 2020).
14. Partnership on AI; Our Tenets. 2021. Available online: <https://partnershiponai.org/about/> (accessed on 28 September 2021).
15. Utrecht Data School. Data Ethics Decision Aid. 2021. Available online: <https://dataschool.nl/deda/?lang=en> (accessed on 28 September 2021).
16. ECP—Platform for the Information Society. Artificial Intelligence Impact Assessment. 2018. Available online: <https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf> (accessed on 4 February 2020).
17. Future of Life Institute. Asilomar AI Principles. 2017. Available online: <https://futureoflife.org/ai-principles/> (accessed on 28 September 2021).
18. European Commission. Ethics Guidelines for Trustworthy AI. 2019. Available online: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (accessed on 28 September 2021).
19. European Group on Ethics in Science and New Technologies. Statement on Artificial Intelligence, Robotics and “Autonomous” Systems. 2018. Available online: https://ec.europa.eu/info/sites/info/files/european_group_on_ethics_ege/ege_ai_statement_2018.pdf (accessed on 7 February 2020).
20. Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* **2021**, *54*, 1–36. [\[CrossRef\]](#)
21. Li, J.H. Cyber security meets artificial intelligence: a survey. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 1462–1474. [\[CrossRef\]](#)
22. Curzon, J.; Kosa, T.A.; Akalu, R.; El-Khatib, K. Privacy and Artificial Intelligence. *IEEE Trans. Artif. Intell.* **2021**, *2*, 96–108. [\[CrossRef\]](#)
23. Narayanan, A.; Shmatikov, V. Myths and Fallacies of “Personally Identifiable Information”. *Commun. ACM* **2010**, *53*, 24–26. [\[CrossRef\]](#)
24. Dwork, C. A firm foundation for private data analysis. *Commun. ACM* **2011**, *54*, 86–95. [\[CrossRef\]](#)
25. Aledhari, M.; Razzak, R.; Parizi, R.M.; Saeed, F. Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE Access* **2020**, *8*, 140699–140725. [\[CrossRef\]](#)
26. Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; Gao, Y. A survey on federated learning. *Knowl.-Based Syst.* **2021**, *216*, 106775. [\[CrossRef\]](#)
27. Friedman, B. Value-Sensitive Design. *Interactions* **1996**, *3*, 16–23. [\[CrossRef\]](#)
28. Hendry, D.G.; Friedman, B.; Ballard, S. Value sensitive design as a formative framework. *Ethics Inf. Technol.* **2021**, *23*, 39–44. [\[CrossRef\]](#)
29. Kriebitz, A.; Lütge, C. Artificial Intelligence and Human Rights: A Business Ethical Assessment. *Bus. Hum. Rights J.* **2020**, *5*, 84–104. [\[CrossRef\]](#)
30. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 5–10 July 2019; pp. 3645–3650. [\[CrossRef\]](#)
31. Capra, M.; Bussolino, B.; Marchisio, A.; Masera, G.; Martina, M.; Shafique, M. Hardware and Software Optimizations for Accelerating Deep Neural Networks: Survey of Current Trends, Challenges, and the Road Ahead. *IEEE Access* **2020**, *8*, 225134–225180. [\[CrossRef\]](#)
32. Capra, M.; Bussolino, B.; Marchisio, A.; Shafique, M.; Masera, G.; Martina, M. An Updated Survey of Efficient Hardware Architectures for Accelerating Deep Convolutional Neural Networks. *Future Internet* **2020**, *12*, 113. [\[CrossRef\]](#)
33. Friedman, B.; Nissenbaum, H. Bias in computer systems. *ACM Trans. Inf. Syst.* **1996**, *14*, 330–347. [\[CrossRef\]](#)
34. Caliskan, A.; Bryson, J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **2017**, *356*, 183–186. [\[CrossRef\]](#)
35. Blodgett, S.L.; Barocas, S.; Daumé, H., III; Wallach, H. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 5–10 July 2019; pp. 5454–5476. [\[CrossRef\]](#)
36. Hovy, D.; Prabhumoye, S. Five sources of bias in natural language processing. *Lang. Linguist. Compass* **2021**, *15*. [\[CrossRef\]](#)

37. Gonen, H.; Goldberg, Y. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings However, do not Remove Them. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 609–614. [\[CrossRef\]](#)
38. He, J.; Baxter, S.L.; Xu, J.; Xu, J.; Zhou, X.; Zhang, K. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **2019**, *25*, 30–36. [\[CrossRef\]](#)
39. Goodman, B.; Flaxman, S. European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation”. *AI Mag.* **2017**, *38*, 50–57. [\[CrossRef\]](#)
40. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [\[CrossRef\]](#)
41. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#)
42. Nauta, M.; van Bree, R.; Seifert, C. Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 14933–14943.
43. Seifert, C.; Scherzinger, S.; Wiese, L. Towards Generating Consumer Labels for Machine Learning Models. In Proceedings of the 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI), Los Angeles, CA, USA, 12–14 December 2019; pp. 173–179. [\[CrossRef\]](#)
44. van Bruxvoort, X. Towards the Design of Legally Privacy-Proof and Ethically Justified Data-Driven Fraud Risk Assessment Algorithms. Master’s Thesis, University of Twente, Enschede, The Netherlands, 2021. Available online: <http://essay.utwente.nl/85984/> (accessed on 24 November 2021).
45. The IEEE Global Initiative on Ethics of Autonomous and Intelligent System. In *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, 1st ed.; IEEE: Piscataway, NJ, USA, 2019. Available online: <https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf> (accessed on 24 November 2021).
46. Mongan, J.; Moy, L.; Kahn, C.E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Kimbrough, S.; Wu, D.; Zhong, F. Computers Play the Beer Game: Can Artificial Agents Manage Supply Chains? *Decis. Support Syst.* **2002**, *33*, 323–333. [\[CrossRef\]](#)
48. Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus and Giroux: New York, NY, USA, 2011; ISBN 978-0374275631.
49. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **2019**, *170*, 51–58. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Ministerie van Justitie en Veiligheid. Kamerbrief over Waarborgen Tegen Risico’s van Data-Analyses Door de Overheid-Kamerstuk-Rijksoverheid.nl. 2019. Available online: <https://www.rijksoverheid.nl/documenten/kamerstukken/2019/10/08/tk-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid> (accessed on 1 May 2020).
51. Sohail, S.A.; Bukhsh, F.A.; van Keulen, M.; Krabbe, J.G. Identifying Materialized Privacy Claims of Clinical-Care Metadata Share using Process-Mining and REA ontology. In Proceedings of the 15th Int’l Workshop on Value Modelling and Business Ontologies, VMBO 2021, CEUR Workshop Proceedings, Online, 4–5 March 2021; Volume 2835. Available online: <http://ceur-ws.org/Vol-2835/paper12.pdf> (accessed on 24 November 2021).
52. Haag, R.D. Court Case on Systeem Risico Indicatie (SyRI). 2020. Available online: <https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878> (accessed on 28 September 2020).
53. Bijvoorbautverdacht. Wat Is SyRI? 2018. Available online: <https://bijvoorbautverdacht.nl/wat-is-syri/> (accessed on 30 September 2021).
54. Hofmans, T. Moet SyRI Worden Stopgezet? Uitkeringsfraude Bestrijden met een Algoritme. 2019. Available online: <https://tweakers.net/reviews/7452/2/moet-syri-worden-stopgezet-fraude-bestrijden-met-een-algoritme-inzet-en-kritiek.html> (accessed on 30 September 2021).
55. Bijvoorbautverdacht. Bart de Koning over de Dubbele Moraal in de Fraudebestrijding. 2018. Available online: <https://bijvoorbautverdacht.nl/gedachte-burgers-potentiele-fraudeurs-zit-systemen-als-syri-gebakken/> (accessed on 30 September 2021).
56. Burg, P. SyRI Ligt nu al Onder Vuur. 2019. Available online: <https://www.computable.nl/artikel/opinie/computable-next/6829101/1509029/syri-ligt-nu-al-onder-vuur.htm> (accessed on 30 September 2021).
57. Janssen, G. Maxim Februari: ‘Je Moet soms Flink Tegen de Wet aan Duwen’. Wordt Vervolgd. 2019. Available online: <https://www.amnesty.nl/wordt-vervolgd/je-moet-soms-flink-tegen-de-wet-aan-duwen> (accessed on 30 September 2021).
58. Teitsma, T. FNV op Ramkoers met Nissewaard over Fraude-Opsporing: Rechtszaak Dreigt. *Algemeen Dagblad*, 27 April 2020. Available online: <https://www.ad.nl/voorne-putten/fnv-op-ramkoers-met-nissewaard-over-fraude-opsporing-rechtszaak-dreigt~a4bd9b0c/> (accessed on 30 September 2021).
59. Teitsma, T. Totta Data Lab Blijft Bijstandsfraude Opsporen, Ondanks ‘Ondoorzichtige’ Controles. *Algemeen Dagblad*, 17 April 2020. Available online: <https://www.ad.nl/voorne-putten/totta-data-lab-blijft-bijstandsfraude-opsporen-ondanks-ondoorzichtige-controles~a87b4743/> (accessed on 30 September 2021).

60. van Dijk, M. Nissewaard Belooft Burger Transparantie over 'Fraudescore'. 2020. Available online: <https://www.fnv.nl/nieuwsbericht/sectornieuws/uitkeringsgerechtigden/2020/06/fnv-sceptisch-nissewaard-belooft-burger-transparan> (accessed on 30 September 2021).
61. de Jonge, A. 'Rookgordijn' Rondom Fraudesysteem Nissewaard. Binnenlands Bestuur. 2020. Available online: <https://www.binnenlandsbestuur.nl/sociaal/nieuws/rookgordijn-rondom-fraudesysteem-nissewaard.13026523.lynkx> (accessed on 30 September 2021).
62. Huissen, R. Nissewaard Weigert Alsnog Beloftes Transparantie na te Komen. Platform Burgerrechten. 2020. Available online: <https://platformburgerrechten.nl/2020/09/22/nissewaard-weigert-alsnog-beloftes-transparantie-na-te-komen/?s=nissewaard> (accessed on 30 September 2021).