

Impact based fairness framework for socio-technical decision making*

Mattias Brännström*, Lili Jiang, Andrea Aler Tubella and Virginia Dignum

Umeå University, Universitetstorget 4, 901 87 Umeå, Sweden

Abstract

Avoiding bias and understanding the consequences of artificial intelligence used in decision making is of high importance to avoid mistreatment and unintended harm. This paper aims to present an impact focused approach to model the information flow of a socio-technical decision system for analysis of bias and fairness. The framework roots otherwise abstract technical accuracy and bias measures in stakeholder effects and forms a scaffold around which further analysis of the socio-technical system and its components can be coordinated. Two example use-cases are presented and analysed.

Keywords

Fairness, socio-technical factors, decision-making system, information-flow.

1. Introduction

Artificial Intelligence (AI) is rapidly becoming more capable and becoming more and more integrated in everyday life and society. Algorithmic decision-making is now commonly used in many contexts, from the personal sphere, to business, and public and private organisations [1, 2]. This makes tools for understanding and analyzing AI systems of critical importance. Much work has been and is being done on the transparency and explainability of algorithmic decisionmaking [3] but the problem at hand goes much deeper than transparency.

Several recent and comprehensive reviews make clear that there is a strong connection between large-scale societal change and development and adoption of AI [4, 5]. There has been significant effort to approach the problem from the top-down view, focusing on identifying and agreeing on high-level values and principles [6, 7]. This approach is taken by most AI ethics guidelines, standards and regulatory approaches. While some issues around the responsible use of AI can be approached in this manner, several of the most complex and impactful interactions are those which exist across the gap between the technical properties of systems and the socio-technical sphere where it is embedded. Problems of fairness, algorithmic accountability, contestability and trustworthiness among others fall into this transcendental category [8, 9]. Yet, much work on algorithmic fairness has had a strongly technical focus with biases in data

Aequitas 2023: Workshop on Fairness and Bias in AI | co-located with ECAI 2023, Kraków, Poland


*Corresponding author.

✉ mattias.brannstrom@umu.se (M. Brännström); ljiang@cs.umu.se (L. Jiang); andrea.aler@umu.se (A. Aler Tubella); virginia.dignum@umu.se (V. Dignum)

🆔 0000-0003-3113-2631 (M. Brännström); 0000-0002-7788-3986 (L. Jiang); 0000-0002-8423-8029 (A. Aler Tubella); 0000-0001-7409-5813 (V. Dignum)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

and models at the forefront [10, 11]. Together with the named biases are so-called fairness definitions focusing on different calculations over the confusion matrix [10, 11, 12]. These are also the approaches taken by leading tools for analysis of bias [8, 13].

The current prevailing directions with growing taxonomies of bias types and fairness definitions or measures have met critique on several levels [14, 15]. The fairness definitions are getting disconnected from their rationale by focusing only on the output, nor is there a clear reason why one should be picked over another [15]. The highly technical focus aids ethics washing and notions that software can be measured to be fair in a disconnected way from how it actually affects any stakeholder [15]. Many of the bias definitions make use of protected categories, but without a closer look on what the respective social role of such protected features are [15]. Another problem is a growing disconnect between expressing bias in terms of deviations from truth and fairness definitions as stand-alone normative objectives.

The aim and contribution of this paper is to link the technical notion of algorithmic fairness to the socio-technical sphere where the connection between technical accuracy and stakeholder impacts can be established. Focusing on stakeholder impacts, AI fairness is realigned to avoid disproportionate harm to real individuals or groups. Since potential impacts can be connected to measurable effects, informed actions can be focused on the areas and problems which are actually causing harm.

This paper will be laid out as follows. Section 1 explores the background on fairness and impact assessment. Section 2 presents the main theoretical framework. In Section 3, we demonstrate the system on two illustrative case studies. Section 4 concludes the paper with a discussion on future directions.

2. Background

The main current in AI Fairness focuses on the one side on taxonomies of biases which can occur in relation to data collection, datasets, algorithms, evaluations, and other components of technical or social systems. These biases are generally expressed as deviations from truth or leading to an incorrect outcome [10, 11]. On the other side we find fairness definitions, which are normative formulas calculated from accuracy measures mainly in the form of a Confusion Matrix - true positive, true negative, false positive and false negative [16]. Such a confusion matrix can be seen in Figure 1b. Over 20 of these fairness definitions are documented and many of them are mutually incompatible given particular cases [12, 10].

The fairness definitions generally do not directly correlate to the degree the various classified biases are expressed but depend on a great variety of additional factors, such as ground truth availability or explanatory variables [17]. They also do not take into account — besides the choice on which definition to use — the socio-technical environment in which they are embedded or how the system affects its stakeholders [15].

2.1. Algorithmic Impact Assessment

Algorithmic Impact Assessments (AIA) are methods for *ex ante* estimation of the societal consequences of an AI system before its development and iteratively during its life-cycle [18]. There is no clear consensus on the exact way to perform AIA but impact assessments in

general typically build on stakeholder identification, participation and risk assessment with accountability as a main result [18]. Algorithmic impact assessment is also at least partially overlapping with Data Protection Impact assessments (DIPA) which are involved in GDPR as well as in Fundamental Rights Impact Assessments (FRIA) [18].

Conducting AIA takes the first step to answering the ‘question zero’ of AI ethics, namely if AI should be used for this case at all by exploring the potential foreseeable consequences [19, 18]. A focus on social impacts to affected stakeholders has the potential to bridge the abstract notions of algorithmic fairness with actual social effects [18, 15].

2.2. Bridging the Context Gap

Another hindrance to the application of high-level ethical guidelines to AI ethics is what is termed as the abstraction gap, i.e., the seeming disconnect between values and guidelines on an abstract level and the specific technical features of a particular AI project. This also applies to algorithmic fairness where all biases and cases of unfair treatment need to be considered for a general and undefined application, but only a small subset of them apply to any real application.

A way to bridge this gap is to capture the context in such a manner that it singles out the issues which are relevant for this particular case. RAIN [6] is an example of a formal framework which breaks down high-level ethical values from AI guidelines into a graph structure of context and stakeholder dependent norms and so bridges this abstraction gap. These norms — representing possible threats to the values from the perspective of the stakeholders — can then be assessed and evaluated in a context-aware manner.

In order to approach algorithmic fairness in a similar manner it would be necessary to characterize the socio-technical domain and links to stakeholders in a way to connect them with the technical measures of fairness and bias. This is what is done in the current paper.

3. The ImpactFramework

We propose the Impactframework to specify a structural scaffold over the information flow of a Socio-Technical System (STS) into which a Technical System (TS), i.e., an automatic decision making or decision support system, is embedded. Into this structure, the outputs, technical and social factors are connected to stakeholder impacts. This structure provides a reference frame for connecting technical measurements to actual effects and to track the risk and effect of potential sources of bias.

This framework does not aim to replace the existing work on stakeholder elicitation, algorithmic impact assessment or accuracy measurements. Rather, it connects the outputs and insights given by these approaches and tools. The modular dependency on these other methods means the framework can be applied in varied use-cases where different approaches might be ideal.

Applying the Impactframework in practice goes through the following steps:

1. Model the information flow.
2. Identify affected stakeholders and impacts.
3. Analyze the structure to determine bias and measures.
4. Perform testing to verify how the system performs.
5. Take action to handle risks of negative impact.

In the following the focus will be on the first three steps, first describing the capture of the system structure and information flow followed by presenting connecting impacts and potential measures to this structure.

3.1. Decision Making Systems

We will formally characterize a Socio Technical System (STS) using the elements of Information Channel Theory [20, 21] that defines an *information network* consisting of three types of objects *sites*, *types*, and *channels*. *Channels* connect *sites* with each other and *sites* are of a *type* [20, 21]. In this paper, types will be assumed to be unique to each site and we can thus combine the notion of type and site.

We will see each *system*, S — STS or TS — as a series of interactions through which some information becomes a particular output decision. Using the Information Channel Theory elements, sites will represent information passing through the system while channels will represents sub-systems which alter the information, like by making decisions.

Minimally, a system on this conception which takes some information I as input and have outputs O can be written as $I \xrightarrow{S} O$. The output O needs to relate to the input I such that all information relating to a particular case upon which the decisions taken by S to give O can be found in I . We can say that $O \subseteq I$ and this requirement impose a partial ordering over the sites. A system can consist of any number of sub-system components, each of which is modeled as input-output maps as described above. Together these nested and chained systems form a *directed acyclic graph* from the base information I through social or technical sub-systems (channels) until finally reaching one or more output sites O representing the decisions taken.

3.2. Modeling the Information Flow

A socio-technical decision-making system (STS) can be seen as a system on the above conception, where at least some component is a Technical System (TS), e.g., an artificial intelligence component embedded into otherwise social components, i.e., people.

Social systems often have non-linear and dynamic interactions between their participants. They can consist of feedback loops and back and forth passing of information. In this framework we will not be modeling this interaction directly but rather follow the transfer of information on an abstract level such that the model contains but underdetermines the actual interactions. Each channel represents a *change* between one state of information or site to another state of information rather than individual interactions of participating agents and components.

The simplest possible STS can be illustrated by a single user interacting with an application illustrated in Figure 1a. Seen as an information flow, we see the initial information state I and the channel S_1 representing what the user will share with the application (which is constrained by both the user and the application). A channel TS is determined by the application, modeling what contribution or change the interaction produces. Finally we have the users channel S_2 to the output site O which describes how the information now with the user contributes to a decision. This simplest socio-technical system could be described in this way

$$I \xrightarrow{S_1} I_1 \xrightarrow{TS} I_2 \xrightarrow{S_2} O,$$

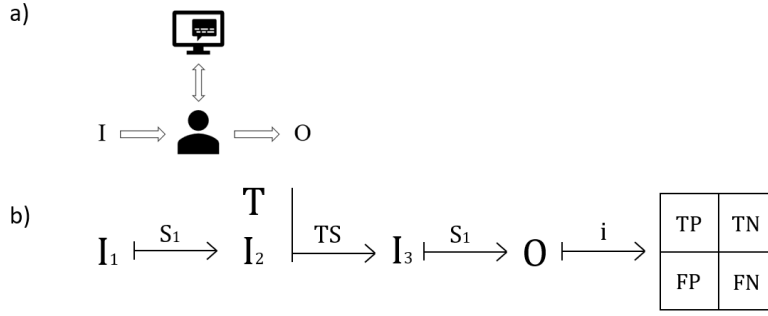


Figure 1: a) A minimal socio-technical decision making system. b) The information flow and associated impact function and confusion matrix of the same system.

where S_1 and S_2 both represent the same user, but in different roles with regards to the information flow — and TS represents the behaviour of the application. A schematic over this information flow is presented in Figure 1b. A real system can be much more complicated and contain both more elements and branches. Because sub-systems and systems are modeled in the same manner we can imagine any number of levels of nesting where systems are in turn described using sub systems. This nesting is very useful for transparency-modeling, but we will focus on single level descriptions in this paper.

Since we are mapping information flow, it is a requirement that $O \subseteq I$, i.e., that the output does not contain information which is not in the input. A case of particular interest for AI decision making systems is *training data*. Suppose that the application in the previous example employs an AI model with training set T . We then write

$$I \xrightarrow{S_1} I_1, (I_1 \cup T) \xrightarrow{TS} I_2 \xrightarrow{S_2} O$$

making it explicit that T is part of the input to the system TS .

Informally, a useful rule of thumb when building graphs like these is to associate each system that has a functional description with a by-phrase explaining how the output is reached from the input. For example, “ TS gives I_2 by comparing I_1 with T ”. This describes TS functionally, and if such a description is easy to give, then the relevant information is accounted for, otherwise something is missing. By making it clear in this manner what information a decision is made upon it is also made clearer which biases and problems might apply to such a decision. Transparency thus builds the foundation of responsible AI [3].

3.3. Identify Stakeholders and Impacts

There are many types of stakeholders for most systems, including decision makers, data subjects, and bystanders. An important class of stakeholders have already been identified, namely the participants of the decision making process — but the focus here is to identify those *impacted* by the system. This can be done using any appropriate stakeholder elicitation method [22]. Details of which are out of scope for this paper and use case dependent.

Impacted stakeholders fall into two categories. Firstly those who are impacted but in a way largely independent of particular decisions taken by the system - we call these *independent subjects* and they include for instance privacy invasion of the subjects of training data, effects on the general public or bystanders for the mere existence or widespread use of the system. These types of impact take an important role in answering the so called Question Zero; 'Should AI be applied here?' [19]. Secondly there are the stakeholders whose impacts are directly dependent on the decisions taken by the system and we call them *dependent subjects*, U . These are the ones which the following framework will foremost focus on and which constitute the recipients of the systems outputs O directly or indirectly. It is important to note that the outputs here refer to the outputs of the socio-technical system, not merely embedded technical systems.

All impacts — dependent or independent — apply to a group of subjects, and both outputs and impacts are distributed *over* subjects as well as over subject attributes. If all subjects are indistinguishable from each other on some attribute then any impact will be homogeneously distributed on that attribute by the system. This holds just as well for social and technical components of the system.

If the subjects are dissimilar however, they could be treated differently. We will say that subjects are *discriminable* when I distribute unequally over U on some attribute. If so, then the output O will also distribute unequally over the same attribute of U . We can see this as for every sensitive attribute of U , I will be subdivided into subgroups and any corresponding impacts will be similarly grouped. This goes in both ways so in order for O to discriminate over some subject sub-groups then I must also be discriminable over the same subgroups. Discriminability can be both positive and negative as there is a potential for biased impact in both unfair discrimination and lack of discrimination where such is needed [10, 11]. Examples of the latter is when medical treatments need to distinguish between protected attributes for proper care, or when the impact of a decision would affect some subgroups differently and disproportionately.

Sensitive attributes of U might be explicit like a gender attribute or indirect in the form of proxy attributes from which sensitive attributes can be derived. It is because of proxy-attributes and unequal representation not always be immediately clear which subgroups exist within a site. It is therefore beneficial to have extra control data — attributes which are part of datasets or measurements but are not necessarily used for the decision making or training data. Such attributes can be used to explore the behaviour of a system and help detect proxy attributes. An example of this is that the presence of a gender attribute in a dataset, which is not used for training or input, can be used to assess the existence of proxy attributes for gender in the data which is used. If no such gender attribute exist associated with the dataset, this examination might be impossible.

Having made an initial identification of the dependent subjects it is possible to consider impacts. Impacts are any normatively valued effects of the system, such as benefits or harms and can not occur without connection to a subject. There is always someone harmed or benefited for the impact to exist. Different groups of subjects of a decision making system can be impacted in different ways. Identifying the impacts in a particular case can be done using Impact Assessments [23].

Like with subjects, *dependent* impacts are derived from the systems output O together with the Confusion Matrix (CM) [16]. The CM is both a theoretical tool and a frame for measures derived using control data with correct answers over many test runs. As a theoretical tool CM

categorizes outputs into true positives, true negatives, false positives and false negatives. Given the CM we can describe an *impact function*, i , which is unique for each STS and for each group of dependent subjects and possibly for some discriminable subgroups. The impact function assigns valuations to all CM outcomes. These impact valuations can be based on user studies, expert knowledge or in some cases direct measurements. Quantifying impacts in comparative ways is inherently problematic and it might be best to separate impacts into categories if they represent dramatically different levels of harm. Yet from an analysis perspective it is helpful to assign numerical values to i so that statistical distributions of CM can be multiplied with i , as long as it is done with care to avoid utilitarian pitfalls.

4. Analyze Bias and Measurements

There are several kinds of measurements which can be made on a system which will show how bias propagate and explain effects on impacts.

This paper will focus on the distribution of the input subjects over discriminatory attributes and the confusion matrix with respect to the end result or partial results. These measures are relatively easily obtainable and as objective as the control data they are compared to, but taken by themselves it is unclear what they mean, if anything. By combining these measurements with their valuation of the impact functions they become contextualized. Valued results can be traced backwards to their source biases. Testing and measurements, as well as elicitation of stakeholders and bias needs to be repeated throughout the system life-cycle.

In order to obtain the confusion matrix of the system output, test data ideally containing the output of every subsystem as well as comparison data representing the correct output for each input is necessary. In some cases it might be difficult to obtain a correct answer, then some other attributes might fill that role by proxy — if so, conclusions must then be modified accordingly.

Given test data the CM describing the distribution of outcomes can be calculated. In case there are several outputs and impacted groups, a confusion matrix for each case needs to be calculated separately. By applying the impact functions to the outputs we can obtain a measure of the impact of the system on the specified dependent subjects. It is not sufficient to test just the technical component in isolation to understand the outputs from the full socio-technical system. Pre-selection of subjects before the technical system as well as how technical outputs are transmitted into final outputs might make a lot of difference to the impacts, where effects of e.g., social biases, trust in algorithmic decisions (low or high) and steps to confirm conclusions might play significant parts. For this reason, test data should ideally be of the outputs of each subsystem for every input. With such data available it is possible to see how the final CM and impacts develop and originate. Some steps can well be seen as partial outputs in their own right if they have independent impacts.

Impacts are typically not best seen as a single quantity but rather as distinct classes of impacts, some of which might be more severe than others and might be best seen as risks of harm. Rather than looking at all impacts together it might be more informative to understand particular classes of harmful impacts, how they are caused and possibly prevented.

Dependent subjects are often not a homogeneous group and impacts might be very unequally

distributed. For this reason it is not fully sufficient with test data which tracks correct or incorrect decisions but also how discriminable attributes propagate. With control data in addition of the information actually passing through the system it is possible to determine how impacts distribute, and also to what degree proxy attributes contribute to make the input and output discriminable. For example, with control data with gender information about subjects it is possible to determine if impacts are equally distributed over genders – and also where in the information flow such bias appear.

In effect, a separate CM should be calculated for every discriminable attribute as well as intersections of attributes. These sub-group CMs with associated impacts will describe how impacts distribute over the subjects. Aggregation bias i.e. incorrect generalisation on an individual level from group data could be detected by an uneven sub-group distribution [10].

Further common biases can be detected by comparing the distributions over sensitive attributes of training sets and other reference inputs (recall that training sets are treated as sub-system inputs). If static or reference information is not representative for the distribution in the information it is compared to the result might be biased or plain incorrect. Sampling bias, population bias and representation bias among other problems can be detected this way and connected to impacts. The connection to impacts help determine if a possible problem is actually a problem in the particular use case of the system.

It would be possible to take this analysis much deeper to track how subsets within the information at each stage contribute to respective decision, but this analysis is outside the scope of this paper. It is worth mentioning however that there are interaction of this type which can not be easily analysed if looking at only a single sub-system, such as a singular AI model, but rather become available for analysis when the interaction of multiple steps and sets of information is considered. An example of this is when information of different levels of generalisation are separated or combined to produce a joint output (which risks aggregation bias, overgeneralization and stereotyping). On a similar vein additional analysis when it comes to system transparency and accountability can also be performed given that the system structure is charted and connected to impacts.

5. Example Case Study

Here two illustrative example cases are presented. The cases are fictional, and while inspired by real situations they are devised to highlight issues which demonstrate the use of the Impact-framework. Were they real, the example case would fall under European AI Act's High Risk classification ¹.

¹<https://artificialintelligenceact.com/title-iii/chapter-1/article-6/>

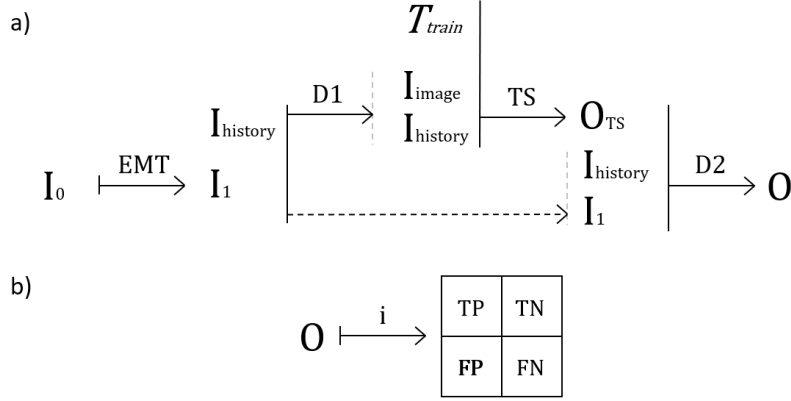


Figure 2: This case is characterized by the combination of a low signal ratio and severe false-positive impact.

5.1. Abuse & Neglect

This example case concerns a decision support for assisting doctors in potential cases of child abuse or neglect and is of interest because it is an example of a case with low signal ratio and a high negative impact for false positives. When children coming to emergency services with suspicious bruises or burns they might be flagged by emergency technicians (EMTs) and will also be given a doctors examination. Into this situation a proposed decision making support system is added. The system will be given image data of injuries as well as demographic data for risk estimation, under the assumption that abuse is more common in some demographics.

A schematic for this case is given in Figure 2 where a) presents the information flow and b) is the impact function with confusion matrix.

$$\begin{aligned}
 I_0 &\xrightarrow{EMT} I_1, \\
 (I_1 \cup I_{history}) &\xrightarrow{D_1} (I_{image} \cup I_{history} \cup T) \xrightarrow{TS} O_{TS}, \\
 (O_{TS} \cup I_1 \cup I_{history}) &\xrightarrow{D_2} O \xrightarrow{i} \mathbf{Impacts}
 \end{aligned}$$

Sites:

- I_0 The injured individual.
- I_1 The injured individual with potential EMT flag
- $I_{history}$ Medical journal and medical history.
- I_{image} Photos of injuries, demographic identity, patient history.
- O_{TS} TS recommendation.
- O Decision to classify as malice or accident.

Channels, i.e sub-systems can be described as:

- *EMT* Medical personnel making an initial labeling.
- D_1 Doctor making the decision on to use the support system and what information to enter.
- *TS* Decisions support system doing image analysis and risk analysis.
- D_2 The final decisions are made by people but will be influenced by the systems output.

A particularity of this case is that while false negatives i.e failure to detect and actual case of neglect or abuse is terrible, so is false positives which might cause wrongful accusations, problems and pain for affected families and children. Another is that most cases are not abuse.

Given this impact estimation it is essential in this case to keep the false positive rate low, while still detecting the true positives. Using this lens we can examine the use of deographic information in the training data T . The rationale for including this sort of information at all would be because abuse and neglect might be more common in certain demographics.

By considering the impact function, we can also see that if even in the highest risk demographic the abuse cases are not the majority it means *any* direct application of a demographic derived factor will lead to higher false positives in the high risk group and higher false negatives in low risk groups. It thus seems, just from the impact distribution, that using demographic information in a mixed manner with the image analysis will always have detrimental impacts regardless of how well the system otherwise performs.

We can also observe from the structure in Figure 5.1 a) that it will be difficult for the doctor to understand when O_{TS} is based upon I_{image} or $I_{history}$ if the graph represents the structure of the real system. This is a transparency problem and could be solved by separating the two modalities into two systems, allowing D to make an informed choice between injury assessment and risk assessment. Beyond this structure analysis, real measurements can be made of at any sub-system. With such measurements the real ratios between outcomes can be established. Combined with the system structure it can be determined how real impacts are caused by the actual system and it's parts.

It is also important to point out that addressing the impacts could also be approached by adding another social step after the outputs specifically to carefully handle the risk of false positives. If such efforts would reduce the negative impact of false-positives significantly, that would in turn potentially allow for more sensitivity in TS (more TP and FP).

5.2. Reinforcement Learning in Hiring

In this example case a reinforcement learning (RL) filter is proposed to be added to a hiring system. Prior to this addition the hiring system is constructed as a filtering pipeline, narrowing down the list of applicants in several steps until the final candidates are presented to the employer and selected for interviews. The first steps are performed using rule based filters with the express aim to not treat protected groups or attributes differently.

After passing through this filtering the potential employers make a shortlist from applications and finally call candidates for an interview.

The proposed RL system would use scores from the interview and shortlist steps to perform a better filtering on applicants before they are presented to the employers. Figure 3 gives an overview of the structure of both the original and the proposed system.

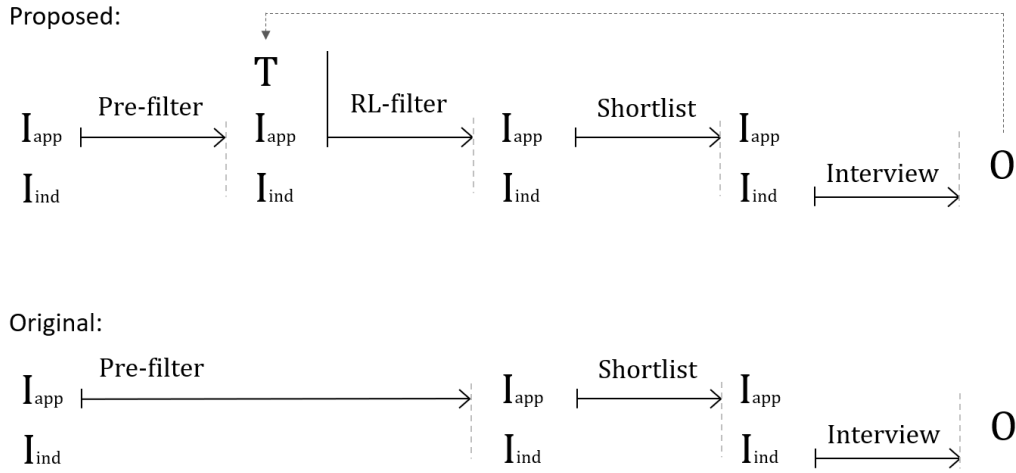


Figure 3: The hiring system of the case study where the application goes through several filtering steps. The Original system uses a pre-filtering step, a manual shortlisting step and an interview. The Proposed system adds another filtering step based upon the shortlist and interview steps.

$$\begin{aligned}
 I_{app} &\xrightarrow{Pre} I_{app}, \\
 (I_{app} \cup T) &\xrightarrow{RL} I_{app} \xrightarrow{Shortlist} I_{app}, \\
 I_{ind} &\xrightarrow{Interview} O
 \end{aligned}$$

Sites from $I_0, I_1, I_2, I_3, I_4, O$ and their corresponding information can be described by these elements:

- I_{ind} The individual job applicant.
- I_{app} The job application.
- T A training set based on interview scores.

Channels, i.e sub-systems can be described as:

- *Pre* Pre-screen applications by comparison between I_{app} and the job posting.
- *RL* RL-screen applications by comparison between I_{app} and T .
- *Shortlist* Employer makes shortlist for interview based on I_{app} .
- *Interview* Employer makes selection based on interview (I_{ind}).

The main impacted stakeholder groups in this system are the applicants and the employers. This case is a good example of where there is no simple true answer if a job applicant fits for a job or not. A proxy 'truth' which can be used however is if the candidate was hired.

Using this we can construct a confusion matrix with eight fields representing where in pipeline the candidate was rejected. This test data can be obtained by letting a number of random candidates pass through the whole system and noting if they would be hired or rejected. From the employers perspective, it is a bad outcome if a candidate who would have been hired is eliminated early and also a bad outcome if a candidate who would not have been hired is

passed to interview or shortlist, because this increases the employers workload and potentially obscures other more suitable candidates. From the applicants perspective, getting eliminated early represents a loss of opportunities, contacts and feedback even if they were not hired. If there is no chance of being hired, getting eliminated early might save time for the applicant, but they are also shut out from direct interaction.

The applicants are not a homogeneous group however. We can assume there are favored and disfavored sub-groups. Such favored groups could disproportionally be chosen over the disfavored group at the interview step and/or the application step. There might also be groups calling for their own impact function — a severe risk which become apparent with a broader perspective is that some groups might be effectively shut out from entering the job market at all and a badly constructed hiring system might exacerbate that problem. This can be illustrated with high impact for early and systematic rejection, especially on the RL step.

We can see that if *Pre* makes a pre-sampling without considering protected attributes then the subjects after this step will be homogenized on non-protected attributes. If so, then the selection by the employer in the shortlist and interview will to a large degree be about the protected attributes (since these are now what mostly set the applicants apart). The RL feedback-loop then amplifies this section by effectively applying the same bias twice. Taken together this means there is high risk for very unequal treatment between groups and *especially so* because of the combination of the *RL* step with the prior *Pre* filtering step.

These steps combined mean that the reinforcement learning must operate on the subset of attributes which are *not* made similar by the initial selection but *are* reflecting some part of the employers judgement and can be found in the application information directly or via proxy.

This situation can be confirmed (or rejected) as influencing by using control data which allows tracking of protected attributes. It can then be determined if there is an unequal distribution at each step. Such data can also confirm biases which occur due to pre-selection e.g who need to use a system like this and who are allowed to use it?

While the above analysis focused on the RL part of the system, the so called fair filtering could also act as a gatekeeper if the criteria e.g particular degrees, are not equally obtainable. This situation can be analyzed with measures in a similar way.

6. Discussion

In this paper, context has been described in terms of information flow and impacts. With a general way to describe context, it is made clear which specific features are problematic and what can be addressed to devise a solution. The distribution of impacts over the technical system could also be used to guide choices of the fairness definitions to apply if that approach is desired.

Distribution and aggregation problems on the level of datasets which are commonly discussed roots of algorithmic biases are also put into context by relating them to their role in the socio-technical system and the flow of information.

6.1. Transparency and Robustness

The framework is to some sense primarily addressing the *transparency* of algorithmic fairness by contextualizing where biases occur and how they connect to impacts. Just like transparency is essential for well-motivated trust in algorithmic systems, it is also essential for understanding a systems role in contributing or alleviating social problems. It is not possible for the users or the affected subjects of an algorithmic decision to apply their own faculties of judgement unless the process behind the decision is sufficiently transparent. One way of achieving such transparency is subdivision into meaningful and meaningfully connected parts [24, 25].

The connection to impacts also allow clearer connections from fairness to risk and safety. Taken in isolation biases and their distributions tend towards the balancing of unavoidable abstract quantities. Looking at impacts however opens other avenues. It might be possible to prevent some harmful impacts entirely just by adding or changing steps, such as extra validation of conclusions with the risk of harmful mistakes. While some issues of fairness really are about balancing resources, some can be meaningfully seen from a safety perspective where risks of mistreatment or wrongful conclusions are contained and removed as much as possible rather than evenly distributed.

7. Conclusion and Further Work

The presented framework provides means for modeling the information flow of a socio-technical decision system such that technical measures and stakeholder impacts can be connected. This initial work however leaves many directions unexplored.

As mentioned above, deeper analysis of the connection between sub-sets within the information flow and datasets will further the detection of biases and their contribution to the impacts. Related to this is also impact driven proxy attribute detection which emphasize that discriminability of attributes can take on both positive and negative roles based upon the composition with other social and technical systems — a dimension which is difficult to capture without an overarching connecting structure.

On the socio-technical side a closer integration with existing initiatives in stakeholder elicitation and impact assessments would be valuable. It might be possible to derive input functions and information flow together and in tandem with procedures for impact assessment and other assessment and evaluation tools. Similarly valuable would be a closer association with frameworks for algorithmic accountability, contestability and risk-analysis. Further validation work is of critical importance.

Acknowledgments

The work has been supported by the AEQUITAS project funded by the European Union's Horizon Europe Programme (Grant Agreement No. 101070363).

References

- [1] A. Theodorou, V. Dignum, Towards ethical and socio-legal governance in ai, *Nature Machine Intelligence* 2 (2020) 10–12.
- [2] D. S. Rubenstein, Acquiring Ethical ai, *Florida Law Review* 73 (2021).
- [3] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information fusion* 58 (2020) 82–115.
- [4] R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S. D. Langhans, M. Tegmark, F. F. Nerini, The role of artificial intelligence in achieving the sustainable development goals, *Nature communications* 11 (2020) 1–10.
- [5] G. D. R. Castro, M. C. G. Fernández, Á. U. Colsa, Unleashing the convergence amid digitalization and sustainability towards pursuing the Sustainable Development Goals (SDGs): A holistic review, *Journal of Cleaner Production* 280 (2021) 122204.
- [6] M. Brännström, A. Theodorou, V. Dignum, Let it rain for social good, *AI Safety* (2022).
- [7] J. Cobbe, M. S. A. Lee, J. Singh, Reviewable automated decision-making: A framework for accountable algorithmic systems, in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 598–609.
- [8] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: A toolkit for assessing and improving fairness in ai, *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).
- [9] A. Aler Tubella, A. Theodorou, V. Dignum, L. Michael, Contestable black boxes, in: *International Joint Conference on Rules and Reasoning*, Springer, 2020, pp. 159–167.
- [10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, *ACM Computing Surveys (CSUR)* 54 (2021) 1–35.
- [11] E. Ntoutsis, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al., Bias in data-driven artificial intelligence systems—an introductory survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (2020) e1356.
- [12] S. Verma, J. Rubin, Fairness definitions explained, in: *Proceedings of the international workshop on software fairness*, 2018, pp. 1–7.
- [13] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al., Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, *IBM Journal of Research and Development* 63 (2019) 4–1.
- [14] V. Dignum, The myth of complete ai-fairness, in: *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, Springer, 2021, pp. 3–8.
- [15] L. Weinberg, Rethinking fairness: an interdisciplinary survey of critiques of hegemonic ml fairness approaches, *Journal of Artificial Intelligence Research* 74 (2022) 75–109.
- [16] S. Visa, B. Ramsay, A. L. Ralescu, E. V. D. Knaap, Confusion matrix-based feature selection, in: *in Proc. Midwest Artif. Intel. Cognit. Scienc. Conf.*, 2011, p. 120–127.
- [17] K. Makhoul, S. Zhioua, C. Palamidessi, On the applicability of machine learning fairness notions, *ACM SIGKDD Explorations Newsletter* 23 (2021) 14–23.

- [18] A. Calvi, D. Kotzinos, Enhancing ai fairness through impact assessment in the european union: a legal and computer science perspective, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1229–1245.
- [19] L. Munn, The uselessness of ai ethics, *AI Ethics* (2022).
- [20] J. Barwise, J. Seligman, et al., *Information flow: the logic of distributed systems*, Cambridge University Press, 1997.
- [21] J. Barwise, D. Gabbay, C. Hartonas, On the logic of information flow, *Logic Journal of IGPL* 3 (1995) 7–49.
- [22] G. I. Pacheco C., A systematic literature review of stakeholder identification methods in requirements elicitation, *J. Syst. Softw.* 85 (2012) 2171–2181.
- [23] H. D, Impact assessment methodologies for microfinance: theory, experience and better practice 28 (2000) 79–98.
- [24] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., *Queue* 16 (2018) 31–57.
- [25] A. Páez, The pragmatic turn in eXplainable Artificial Intelligence (XAI), *Minds and Machines* 29 (2019) 441–459.