cy. center for algorithmic transparency

Document Title	**Summary of core technical concepts for end-users and developers**
Project Title and acronym	Cyprus Center for Algorithmic Transparency (CyCAT)
H2020-WIDESPREAD-05-2017-Twinning	Grant Agreement number: 810105 — CyCAT
Deliverable No.	D4.1
Work package No.	WP4
Work package title	Promoting Algorithmic Transparency
Authors (Name and Partner Institution)	Avital Shulner-Tal, Veronika Bogina, Alan Hartman, Tsvi Kuflik -University of Haifa
Contributors **(Name and Partner Institution)**	
Reviewers	Kalia Orphanou (OUC) Lena Podoletz (UEDIN)
Status **(D: draft; RD: revised draft; F: final)**	F
File Name	D4.1_Summary_Core_Concepts_M18
Date	31 March 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810105.

Draft Versions - History of Document

Version	Date	Authors / contributors	e-mail address	Notes / changes
v1.0	17/09/2019	Avital Shulner-Tal Tsvika Kuflik	tsvikak@is.haifa.ac.il	Initial version
v2.0	8/11/19	Alan Hartman	ahartman@is.haifa.ac.il	
v3.0	22/11/19	Tsvika Kuflik	tsvikak@is.haifa.ac.il	adding index + creating semi final version
v4.0	12/1/2020	Avital Shulner-Tal, Tsvik Kuflik	avitalshulner@gmail.com; tsvikak@is.haifa.ac.il	Final version

Abstract

D4.1 provides a review of core concepts for user groups. It involves taking an inventory of what technical concepts surrounding algorithms can and should be understood by core user groups (children, teens, adults / public employees) as well as for system developers who may be in a position to affect the transparency of algorithms (e.g., user interface designers and developers of system-user interaction loops). It will be used to inform the choice of intervention(s) to be developed in WP5 as well as the educational materials.

Keyword(s):	Algorithmic bias, Discrimination Discovery, Explainability promotion, Fairness promotion.

Contents

1. Executive Summary

D4.1 integrates and abstracts the findings of WP3. In WP3, we surveyed the scientific literature in the emerging field of Fairness, Accountability and Transparency (FAT), characterizing the problem and solution spaces described by FAT researchers within the information and computer science disciplines. The current document aims to structure and extend the findings from WP3, resulting in a more holistic understanding of the *key concepts* surrounding FAT, from the point of view of various stakeholders involved in these processes (i.e., not only developers of algorithmic systems, but also those who use them and those who monitor and/or regulate their behaviors).

D4.1 first presents a set of case studies of algorithmic systems as an introduction, gradually increasing in potential of risk/damage as a result of bias, motivating the need to better define the technical concepts surrounding FAT (problems and solutions), as well as clarifying the needs of various stakeholders (Section 2). Following that, Section 3 details a holistic model for discussing FAT, from the problem space (the components of an algorithmic system and the potential risks to fairness) to the solution space (detecting and mitigating risks). While these spaces were discovered and described briefly in WP3 (D3.1), here, they are analyzed in more detail, taking into account their relation to three broad classes of stakeholders (Observers, Developers and Users), who have very different perspectives of the system. The model presented is then mapped back to the set of case studies presented in Section 2. Finally, Section 4 presents a stakeholders' body of knowledge, which clearly articulates a mapping between stakeholder role and the need-to-know concepts surrounding FAT.

Thus, D4.1 lays the groundwork for the other four deliverables in this work package. As described below, each deliverable extends / focuses on particular aspects and audiences that are important to the overall goals of the CyCAT project.

- D4.2 is an easy-to-read guide for educators, summarizing the core technical concepts surrounding algorithmic system transparency and explaining how these can be taught to secondary school students. In particular, this document "translates" the framework for end-to-end fairness management in algorithmic systems, presented in D4.1. It explains to teachers the importance of raising students' awareness of algorithmic processes and algorithmic bias, by contextualizing the CyCAT framework within familiar pedagogical principles. Finally, this document provides example lesson plans (developing the objectives, materials, activities, and finally, the evaluation) enabling teachers to implement them in their own classrooms.
- D4.3 elaborates more on the types of data issues, identified in D4.1 Section 3.2, that can present risks to fairness in an algorithmic system and suggest ways to train relevant stakeholders, specifically raise awareness of developers and regulators to such issues and train them how to examine, identify and mitigate such issues.
- D4.4 aims to provide more in-depth analysis of the FAT processes. It provides suggestions for educating stakeholders, mainly developers about the actual processes aplied by algorithmic systems and their development for increasing awareness for potential biasses and discrimination and motivating the mitigation of these challenges.
- D4.5 focuses on describing in more depth the body of FAT concepts that system users need to know.

2. Illustrative examples

While we cannot claim that the case studies in D3.2 represent an exhaustive search of the popular press coverage on algorithmic bias, analyzing the contents of our dataset can provide some insight into what kinds of systems and problems catch the public's attention and those that may be more likely to result in harm to particular individuals and/or groups.

We choose three illustrative examples as case studies, presenting them in terms of the increasing order of the severity of the impact on the user most affected by the outcome of the algorithmic system, as a way to introduce the user to the actual issues and concepts used later in the document..

Case study 1: AD_SERV: Assume you are using the WAZE (https://www.waze.com/) navigation App and you are getting specific recommendations for businesses nearby. Why do you get these ads?

You may assume that these are the best recommendations for you according to what the system knows about you (personal characteristics and preferences) + the contextual information it has (location, day of the week, time of day, weather, whether you are in a hurry or not, etc) and how they fit the recommended items.

However, this may be actually true or it may happen simply because the specific advertisers paid for their ads to be delivered to people like you…

Technically we have here an **algorithmic system** that recommends a list of advertisements to present on a webpage or to present on a mobile app used by an **end user**. The recommendations are based on information about the identity and attributes of the viewer of the webpage, their recent activity on that page or app and additional contextual aspects, as noted above. When considering advertisements, each one is equipped with a set of desired targets and with its history of display and click-through that are used for targeting users - so it is not only for the benefit of the end user...Hence the recommendations may be **explicitly biased.**

Case study 2: CREDIT_RATE: Assume you apply for a loan from a bank, because you need some money to open/extend your business. You visit your bank, talk to the person that deals with loans. This person does some magic with her desktop and at the end lets you know whether you are entitled at all to get a loan (you may simply be denied a loan as "the bank assumes that you or your request are too risky"), and if so, what are the terms of the loan.

Technically, again, it is an **algorithmic system** that recommends to a **bank officer**, who is its **actual user**, whether or not to grant a loan to a **custome**r, who is the **user that is being impacted** by the system. The decision is based on information about the particular customer (credit history, personal wealth…), the conditions of the loan requested, and a **database** of prior, both successful and unsuccessful, applicants, including the repayment behaviour of successful loan applicants. This information may also include post codes of the customer that may be correlated by the system with large concentration of customers that cannot be trusted - hence they are **proxy attributes.** These may be mostly customers from a certain social/ethnical group and if their ethnicity is recorded, then the system uses **protected attributes** and causes violation of **group parity.**

Case study 3: COMPASS - a system that intended to predict recidivism (see [Skeem & LowenKamp 2016] and [Larson et al. 2016]). The (**algorithmic**) **system** was intended to support judges, probation and parole officers (**system users**) to assess a criminal defendant's likelihood of becoming a recidivist (a term used to describe criminals who reoffend).

There are dozens of these risk assessment algorithms in use. Many states have built their own assessments, and several academics have written tools. There are also two leading nationwide tools offered by commercial vendors. Larson et al. analysis of COMPASS (Correctional Offender Management Profiling for Alternative Sanctions), showed that black (**protected attribute**) defendants were far more likely than white (**protected attribute**) defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk. They compared the recidivism risk categories predicted by COMPASS to the actual recidivism rates of defendants in the two years after they were scored, and found that the score correctly predicted an offender's recidivism 61 percent of the time, but was only correct in its predictions of violent recidivism 20 percent of the time. In forecasting who would re-offend, the algorithm correctly predicted recidivism for black and white defendants at roughly the same rate (59 percent for white defendants, and 63 percent for black defendants) but made mistakes in very different ways. It misclassified the white and black defendants differently when examined over a two-year follow-up period (**implicit bias**). The violent recidivism analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 77 percent more likely to be assigned higher risk scores than white defendants.

3. Model for Algorithmic Transparency and Fairness

In order to have a complete and stand alone document, we do not rely here on aspects that were partially developed during earlier stages, but present here a holistic approach, that is based on previous steps and enhanced with knowledge gained during the analysis of the literature review results.

3.1 Algorithmic Systems

In order to discuss algorithmic systems, we first present a model that enables us to discuss the potential sources of risks to fairness and transparency and to suggest methods to mitigate these risks. An abstract algorithmic system has five main sources of risk (enhancing the model that was initially suggested by WP3):

Input (I) - the particular values input to a specific run of the algorithm

Output (O) - the value(s) produced in response to the input

Algorithm (M). The algorithmic core that, given a particular instance (after being trained), performs computation based on this **Input (I)** and provides an **Output (O)**. In some learning models, the algorithm itself undergoes change due to the addition of the Input (I) to the store of **Training Data (D)**. For instance, the addition of data may come from third parties, and their interactions with the system (i.e., implicit behaviours and/or constraints, see below).

Training Data (D). Data which is used to train the **Algorithmic (M)** when some machine learning techniques are applied.

Third Party Constraints (T). Implicit and explicit constraints, given by third parties (not necessarily set by developers), that may impact the design and performance of the

Algorithmic (M). These include operators of the system, regulators, and other bodies which influence the use and outcomes of the system.

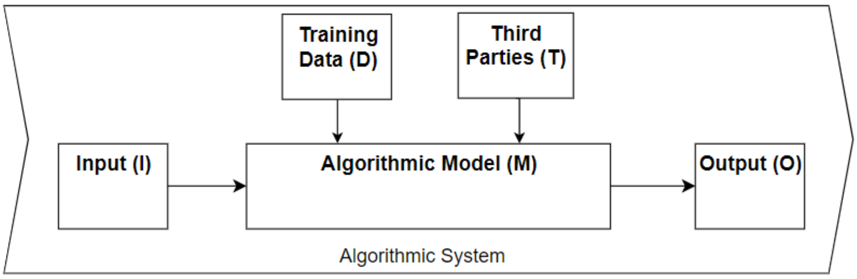


Figure 1. Algorithmic System Model

3.2 Risks to Fairness

We classify algorithmic biases on the basis of the causal factor: data bias, processing bias, and human bias

- **Data Bias.** Biases that can appear in the input (I) or training data (D) (Danks and London, 2017)].
  - **Input Bias.** The input data may contain information about sensitive attributes in an implicit or explicit way. This category also refers to insertion of incorrect or incomplete information by the user [Danks and London, 2017].
  - **Training Data Bias.** I○ Information about sensitive attributes of people may be contained in the training data and such information may be unbalanced and discriminatory to particular groups of people. The training data may also be based on an unrepresentative set of instances, and may also suffer from inaccurate or biased classification (i.e., inaccurate “ground truth” / annotation) [Danks and London, 2017].
- **Algorithmic Processing Bias**. Biases that can appear during algorithmic model (M) learning and processing [Danks and London, 2017]. (e.g. information from insensitive attributes can infer in some way the values of the sensitive attributes, and the algorithms can exhibit discriminatory behavior unintentionally [Madaan et al. 2018]).
- **Human Bias**. Biases that relate to humans in inappropriate system development or usage
  - **Third Party Bias.** Biases that are caused by the **Third Parties (T) (**Implicit and explicit constraints regarding the design and performance of the system) that are not directly related to the development process [Tal et al., 2019].
  - **Transfer Context Bias.** Applying the algorithmic system in a context which is both different from its intended use and inappropriate
  - **Developer Bias.** Expecting certain outputs, can result in unconsciously incorrect handling of the data or incorrect development of the algorithm. (e.g. focus on specific data while ignoring generic data [Klauer et al. 2000], knowledgeable systems often fail to see the users' point of view [Volker 2003], developer's stereotyping).

We also concern ourselves with a discussion of the perception of bias - justified or not - but observable to users or other observers of the algorithmic system.

- **Perceived Bias**. Biases that are related to the perception of the input-output correlation within and between users [Chiu et al. 2009] (e.g. cognitive biases [Pohl, 2004]). The perceived bias can be related to the diversity of knowledge, which can influence the training data and may create an algorithm that is perceived to be "unfair", as well as to the diversity of users, which may have different perceptions of the world and can interpret the system behaviors in different ways [Giunchiglia, 2006]. (e.g. diversity dimensions [Maltese et al., 2009])

Figure 2 illustrates the risks to fairness and transparency according to the algorithmic system they are related to – the system components that were presented at Figure1. are augmented with the biases that may be introduced to the system.

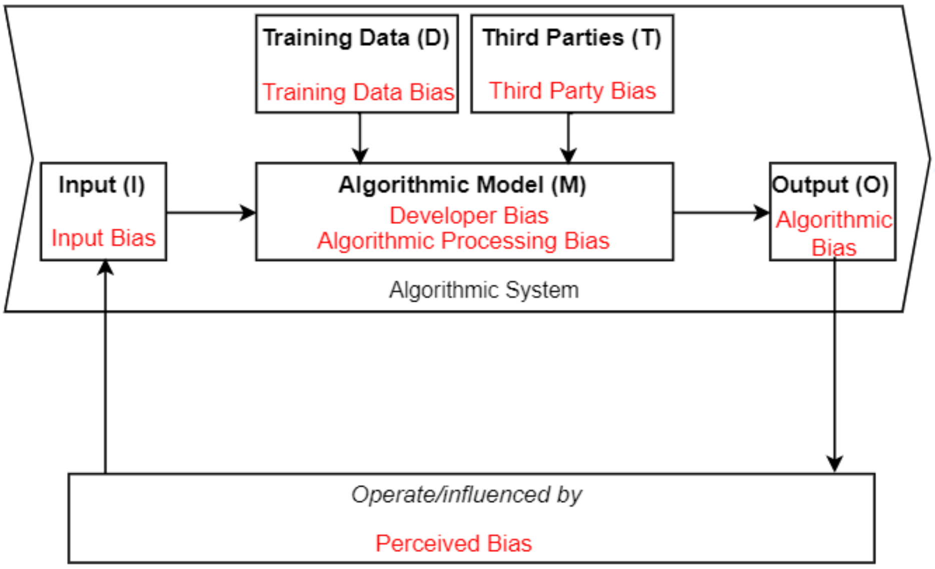


Figure 2. Risks to fairness and transparency. Algorithmic system components potential biases and where they may occur including perceived bias. Arrows represent data flow.

3.3 Risk Detection and Mitigation

In this section we discuss the detection and mitigation strategies for risks to fairness and transparency. In order to address risk detection and mitigation we consider the following definitions:

- Protected (sensitive) Attribute. Refer to attributes that contain confidential information about a specific individual such as race, sex, language, religion, political or more.

- Protected Group. Refer to a group of people that is qualified for special protection by law, policy, or similar authority.

- Proxy Attributes. Refer to features that are proxies of sensitive attributes (e.g. neighborhood may infer the race or income level of an individual). Those attributes and may cause bias, even if the sensitive attributes are excluded [Zhang et al., 2019].

We first present the formal aspects of measurement and strategies that can demonstrate fairness compliance with two rigorous definitions – [Dwork et al., 2012]: ***Individual fairness*** – "Any two individuals who are similar with respect to a particular task should be classified similarly". ***Group fairness*** – "The property that the demographics of those receiving positive (or negative) classifications are identical to the demographics of the population as a whole" (also known as statistical/demographic parity [Gajane and Pechenizkiy 2017]). Later we discuss the less formal aspects of perceived fairness and how to address it.

We consider a system to be fair and transparent once both formal and informal aspects of fairness (as presented below) have been addressed.

3.3.1 Formal Fairness Detection and Mitigation

The issue of discrimination discovery was discussed by [Pedreschi et al., 2009] who presented the following model - based on a case similar to our CREDIT_RATE example.
Figure 3 presents a discrimination discovery model [Pedreschi et al., 2009] that builds on a set of rules elicited from historical decisions of a Decision Support System (DSS) for providing credit for potential applicants. The discrimination analyses flow starts with the input pool (case attributes like age, job type and etc.) that is based on the historical records of the applications, sometimes enriched with an external data (this is also the input to the DSS - a black-box trained system). The output – the result of the system, together with the input data becomes a training set for a rule induction process. Using this set, a set of classification and association rules are extracted. In the end, the output is the set of the potentially discriminatory patterns that can unveil the contexts of the groups' discrimination. Moreover, as can be seen from the figure, the process is iterative.

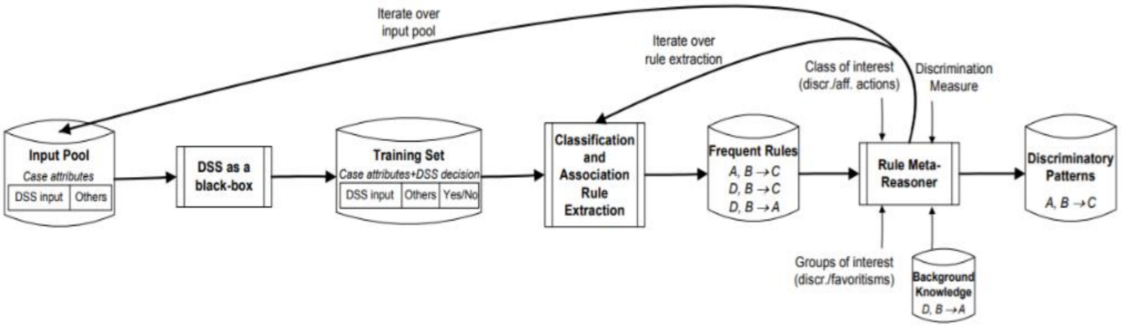


Figure 3. Iterative discrimination discovery model for providing credit to applicants [Pedreschi et al., 2009]. The input of the model consists of historical transactions of credit, and the classification of the original system. Then association and frequent rules are extracted and the output of the model is the set of discriminatory patterns for this group of applicants (e.g. the group discriminating patterns that are found as reflected by the inferred association rules given the data and classification).

When considering discrimination discovery, two aspects of discrimination are considered, implicit and explicit discrimination discovery:

- **Discrimination Discovery.** The ability to identify discrimination against sensitive groups in the population, caused by biases in an algorithmic system. It can be divided into the following kinds:
  - ○ **Explicit** (direct) **Discrimination Discovery.** The ability to identify discrimination which is caused by both data biases and inappropriate use of sensitive attributes in algorithms [Hannák et al. 2017].
  - ○ **Implicit** (indirect) **Discrimination Discovery.** The ability to identify discrimination which is caused by algorithmic processing biases and human biases due to the fact that some insensitive attributes are very informative about sensitive attributes [Speicher et al. 2018].

With respect to discrimination discovery, we can consider case studies 1, 2 and 3 and assume that by applying (explicit) discrimination discovery techniques, we can identify the unfair behavior of the system in case 1. The use of (implicit) discrimination discovery techniques may enable us to identify the unfair behavior of all 3 systems.

The mitigation of threats to fairness is a process we call fairness management:

- **Fairness Management.** The ability to ensure fairness with regard to sensitive groups in the population by applying predefined fairness measures that quantify an undesired bias in the training set or in the model. Fairness management processes include:
  - **Fairness Assurance.** Three strategies have been identified for improving the (objective) fairness of an algorithmic system (Kilbertus at al. 2018).
  - **Fairness Formalization.** Formalizing fairness (Kusner et al. 2017) and defining different ways for addressing fairness (Gajane and Pechenizkiy 2017). Figure 4,5, present different kinds of fairness formalization models that can help to select the most suitable fairness measures to each context. Figure 4 presents a "fairness tree" that provides guidelines for selecting the relevant fairness metric(s) based on the context of the problem [Saleiro et al., 2018]. It helps the decision maker to decide whether she cares more about the distribution of false negatives results or false positives results based on a few questions (e.g. whether the model predicted labels can be changed, and whether those interventions will help people or hurt them) and provides the relevant fairness metric(s). Figure 5 presents a "fairness matrix" that helps selecting the most suitable fairness measures for ML prediction problems based on the answers to the following two questions: (1) whether fairness is considered as achieving parity (provide equal probabilities for individuals across groups) or satisfying preferences (considering individual choices within groups) and (2) whether fairness needs to be measured in the treatment (reference to certain protected demographic groups) or in the impact (results) [Gajane and Pechenizkiy, 2017]. Furthermore, Verma and Rubin [Verma and Rubin, 2018] gathered 25 useful definitions of fairness measurements for classification problems. They analyzed and demonstrated the rationale of these definitions and abstracted them into 3 classes: (1) Definitions based on the predicted

outcome for various demographic distributions of subjects. (2) Definitions based on the predicted outcomes for different demographic distributions (that also compare it to the actual outcome that is recorded in the dataset). and (3) Definitions based on predicted probabilities and actual outcome (that considers both the actual outcome that is recorded in the dataset and predicted probability score for a certain classification).

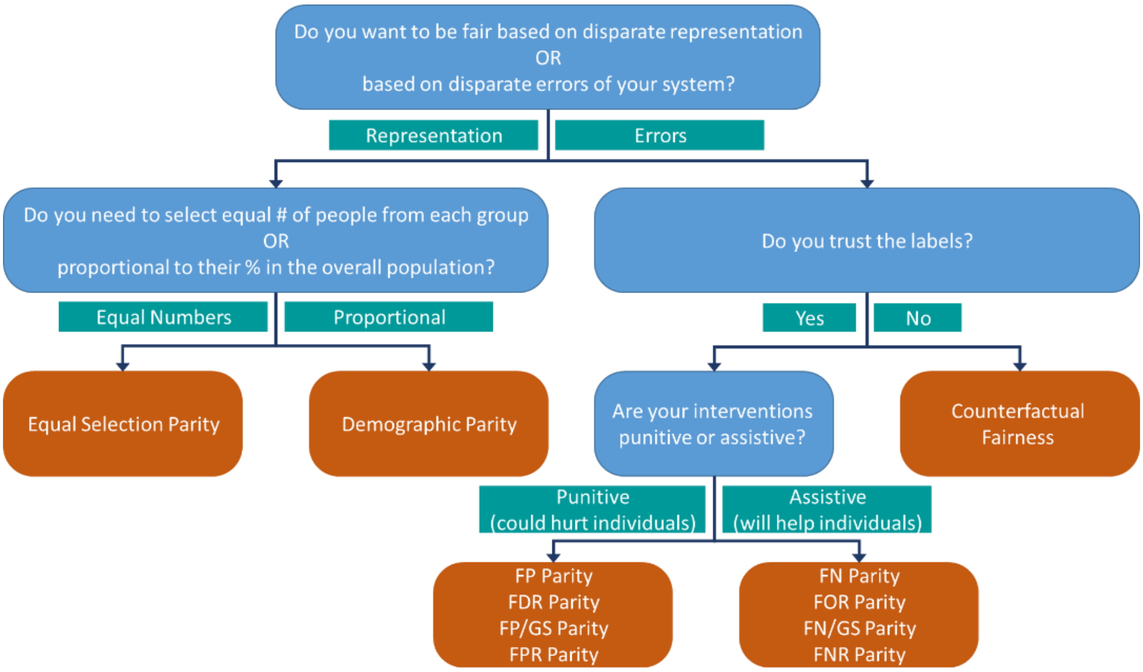
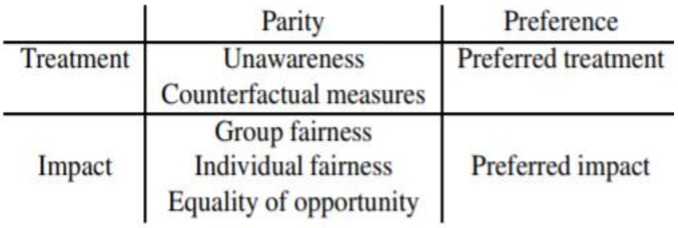

Do you want to be fair based on disparate representation
OR
based on disparate errors of your system?
Representation
Errors
Do you need to select equal # of people from each group
OR
proportional to their % in the overall population?
Do you trust the labels?
Equal Numbers
Proportional
Yes
No
Equal Selection Parity
Demographic Parity
Are your interventions punitive or assistive?
Counterfactual Fairness
Punitive (could hurt individuals)
Assistive (will help individuals)
FP Parity
FDR Parity
FP/GS Parity
FPR Parity
FN Parity
FOR Parity
FN/GS Parity
FNR Parity

Figure 4. Fairness formalization model [Saleiro et al., 2018]. The fairness tree provides guidelines for selecting the relevant fairness metric based on the context of the problem.

Considering case study 3, as a (simplified) example, considering figure 4 and assuming we do have **training data that we trust** (as these are historic evidence), hence **no errors are assumed to be in the system** and we test the representation, then it is needed to select **equal numbers of different group members** to avoid the bias towards ethnic groups.



	Parity	Preference
Treatment	Unawareness Counterfactual measures	Preferred treatment
Impact	Group fairness Individual fairness Equality of opportunity	Preferred impact

Figure 5. Fairness formalization model. Gajane and Pechenizkiy [Gajane and Pechenizkiy, 2017] defined fairness through addressing two main questions: (1) Parity or preference? (2) Treatment or impact? They classified the fairness formalization considering these two main questions by creating "fairness matrix".

Considering case study 3 (again), as a (simplified) example, considering figure 5 and assuming we do test the **Impact** on **Parity** we need/may select Group fairness measures and/or Individual fairness measure and/ or Equality of opportunities - for checking that there is no impact on individuals/groups give the reasoning of the system with respect to different groups and between individuals with identical characteristics, but different protected attributes (e.g. race)..

- **Fairness Sampling.** Sampling a subset of data from the training set in order to reduce the data bias by training the model on different distributions samples [Zemel et al. 2013; Torralba and Efros 2011]. A simple example in this case is ensuring equal numbers of cases with similar distribution of labels for both ethnic groups.
- **Fairness Learning.** Train an algorithm to be fair under the given fairness constraints [Zafar et al., 2015]. figure 6 suggests the fairness pipeline [Bellamy et al., 2018] that aims to make fair predictions. They distinguish between three main paths: (1) fair pre-processing, that can be used when modification of the training data is allowed and removes underlying discrimination from the data before any type of modeling is preformed [d'Alessandro et al., 2017] (2) fair in-processing, that can be used when the traditional learning algorithms for a ML model can be modified in order to address discrimination during the training procedure [d'Alessandro et al., 2017] and (3) fair post-processing, in any other case, e.g. changing the labels of the class or the confidence of the classification rule [Romei et. al. 2013] . those algorithms can handle fairness at different stages of the training of the model by transforming the original dataset into a fairer dataset [d'Alessandro et al., 2017]

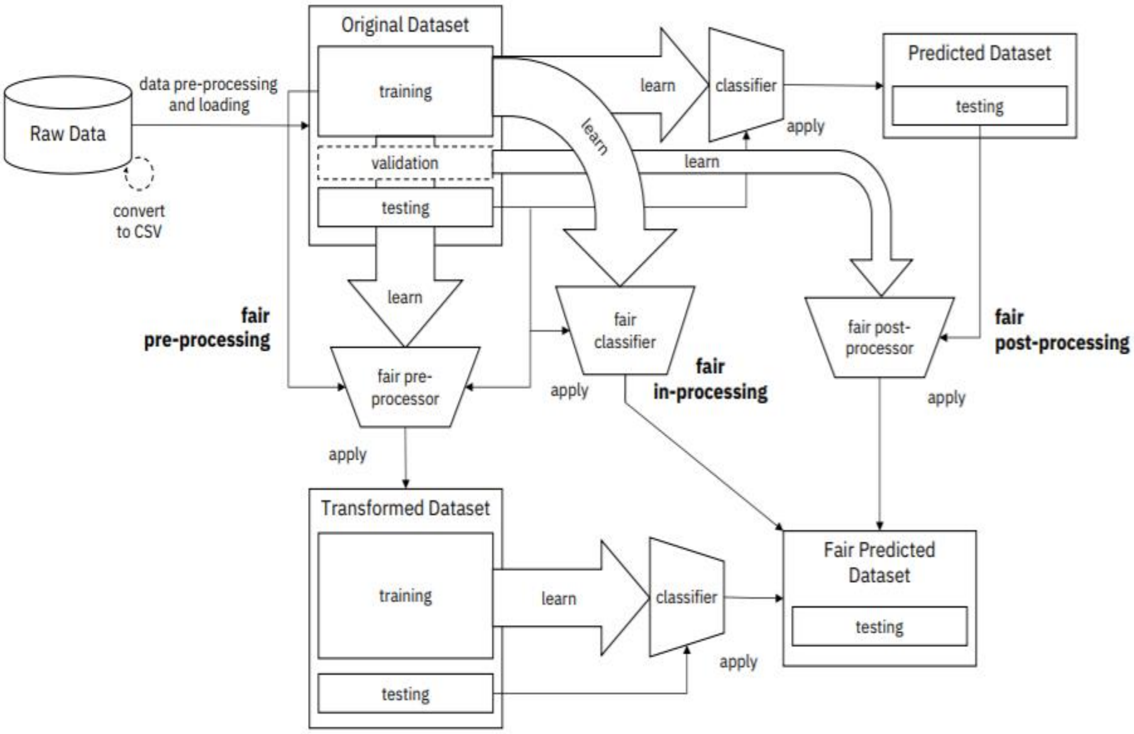


Figure 6. Fairness learning model. Bellamy et al., [2018] presented a fairness pipeline to achieve fair dataset. The pipeline includes three possible paths (bolded) when the user can choose one of them.

- **Internal Fairness Certification.** Once no unintended discrimination is discovered after implementing the fairness assurance (if intended discrimination was discovered make sure that this is by design), the developer verifies whether algorithm's output satisfies fairness constraints that were defined in fairness formalization, and then the developer can consider and certify the algorithmic system as fair.

- **Auditing.** The ability to audit the process and results of the algorithm by an external regulator in order to assess compliance with specifications, standards, contractual agreements, or other criteria (e.g. study the correlation between inputs and outputs (Eslami et al. 2017; Sandvig et al., 2014).
- **Formal Fairness Certification.** A regulator or certification authority can decide whether to certify the fairness of an algorithmic system based on the auditing results and the internal fairness certification (Kilbertus et al. 2018).

Considering case study 3 again, we may test the three options - the training set itself, the process and the outcome. Depending upon the results, we may fix the dataset, if it found to be biased, we may check the algorithm in we have training data that we trust and we may take corrective actions to fix the output if needed.

3.3.2 Informal Fairness Detection and Mitigation

It is not enough to apply a set of measures to ensure the fairness of an algorithmic system, as its users need to be convinced that it is fair. One component of this is transparency – providing a reasonable level of explanation about the process and the results of the algorithmic process. Another means to improve the perceived fairness of a system is by certification using a trusted external observer.

Explainability management can create a more transparent and interpretable system and therefore can increase the fairness of the system as shown in Figure 7.

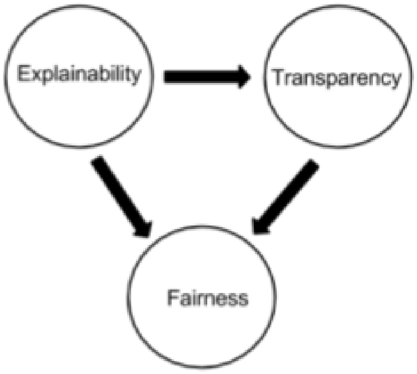


Figure 7. Explainability, transparency and fairness relationship [Abdollahi and Nasraoui, 2018]. Explainability is needed for transparency and both are required for the fairness of the system. Abdollahi and Nasraoui relate to fairness as formal fairness yet it can also be interpreted as perceived fairness - Explainability and transparency can increase the fairness and the perceived fairness of the system [Ribeiro et al., 2016]

Zhang et al. (2019) refer to machine learning interpretability as the extent of understanding the reason for the decision of a system by an observer. They claimed that interpretability contains both transparency and post hoc explanations and that interpretability is required by regulators due to the user's legal right to explanation.

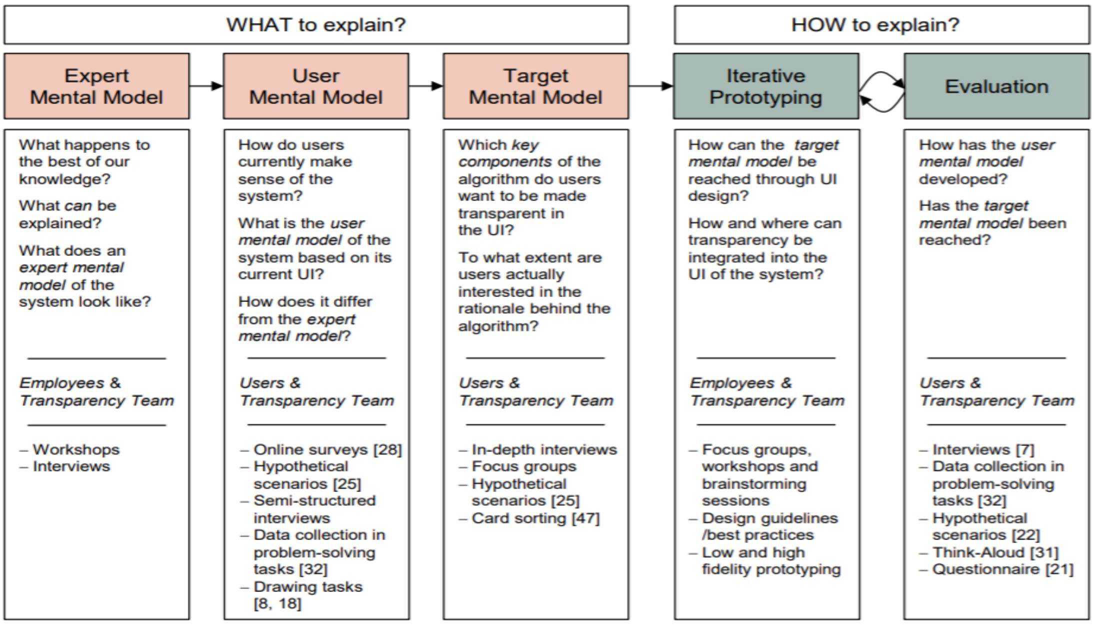
- **Explainability Management.** The ability to explain the decisions made by algorithmic systems to users in order to increase the assessment of trust of the users. Ribeiro et al. presents a novel explanation technique which can explain the output of any classifier in an interpretable and faithful manner by learning an interpretable local model. They also proposed a method for presenting the representative individual predictions [Ribeiro et al., 2016]. Figure 8 suggests a five stages explanation model which deals with creating the “what” of the explanation (stages 1-3) and its presentation format - “how” (stages 4-5) of the design process for the integration of transparency of intelligent systems. It presents the stages of the explanation process by referring to guideline questions, importance, different stakeholders that are involved, outcomes and exemplary methods for each stage in the process. The first stage (Expert Mental Model) is used for gaining common understanding about data collection and processing methods. The second stage (User Mental Model) deals with the users' beliefs about the system logic and transparency in order to create a list of differences between the user and expert mental model. The third stage (Target Mental Model) deals with the trade-off between transparency (displaying more information) and the visual or cognitive load which can be formed. The fourth stage (Iterative Prototyping) aims to create several prototypes for integrating the explanations into an existing UI, and the fifth stage (Design Evaluation) deals with evaluating of the different prototypes with respect to design changed that can improve users' mental model [Eiband et al., 2018]. Explainability can be classified as follows:
    - ○ **White-box Explanation**. White-box algorithms reveal their structure therefore it is easier to explain both the model and the outcome as it appears from their definitions. (e.g. decision tree).
    - ○ **Black-box Explanation**. Such explanations fill an intention gap between user's needs and interests and the system’s goals [Wang and Benbasat, 2007]. A comprehensive survey of black-box model explanations can be found at [Guidotti et al., 2018].

      The explanation falls into two categories:

        - ■ **Model Explanation.** Explaining the logic of vague classifier by using an interpretable and transparent model that can mimic the behavior of the black-box model. The interpretable model (e.g. decision trees, decision rules, the contribution of features to the decision and so on) uses an interpretable global predictor, that can be derived from the black-box itself, and instants of the dataset of the black-box, that can be extracted by using random perturbation or random sampling. Figure 9 suggest a way to provide an interpretable and transparent model that can be understandable to users. Furthermore, Figure 10 suggest that the explanation should occur in the modeling phase since it can help in designing a more transparent models [Abdollahi and Nasraoui, 2018].
        - ■ **Outcome Explanation.** Explaining the correlation (reason for prediction) between a particular input and its output, without explaining the whole logic of the black-box model. An interpretable local predictor (e.g. decision rule classifier) is built for every test instance and the explanation is the specific rule that was used for classifying this instance as can be seen

in Figure 11. The outcome explanation should occur at the prediction phase by presenting a justified result to the user and as a result the decision can be more transparent to the user which may increase the fairness [Abdollahi and Nasraoui, 2018], as suggested in Figure 12.

When considering the three case studies, explanation management can be used for explaining the results of the process, if the system is a “black box” system - in this case a process similar to the one described by Figure 3 may be used to elicit the reasoning rules in every specific case and these rules may be used for constructing a human understandable explanation. The details of the explanation creation process will follow the 5-stages of Fig. 8. For instance, in case study 1 - the explanation should be about how the system got to the specific recommendation to the user and this may lead to the decision how to do that - textual short explanation/visual/audio...



WHAT to explain?

HOW to explain?

Expert Mental Model

What happens to the best of our knowledge?

What *can* be explained?

What does an *expert mental model* of the system look like?

Employees & Transparency Team

– Workshops
– Interviews

User Mental Model

How do users currently make sense of the system?

What is the *user mental model* of the system based on its current UI?

How does it differ from the *expert mental model*?

Users & Transparency Team

– Online surveys [28]
– Hypothetical scenarios [25]
– Semi-structured interviews
– Data collection in problem-solving tasks [32]
– Drawing tasks [8, 18]

Target Mental Model

Which *key components* of the algorithm do users want to be made transparent in the UI?

To what extent are users actually interested in the rationale behind the algorithm?

Users & Transparency Team

– In-depth interviews
– Focus groups
– Hypothetical scenarios [25]
– Card sorting [47]

Iterative Prototyping

How can the *target mental model* be reached through UI design?

How and where can transparency be integrated into the UI of the system?

Employees & Transparency Team

– Focus groups, workshops and brainstorming sessions
– Design guidelines /best practices
– Low and high fidelity prototyping

Evaluation

How has the *user mental model* developed?

Has the *target mental model* been reached?

Users & Transparency Team

– Interviews [7]
– Data collection in problem-solving tasks [32]
– Hypothetical scenarios [22]
– Think-Aloud [31]
– Questionnaire [21]

Figure 8. Explanation model [Eiband et al., 2018]. Eiband et al., 2018 suggested a pipeline of five stages (what to explain: stages 1-3, how to explain: stages 4-5) for participatory design process. All stages consider different relevant stakeholders and questions.

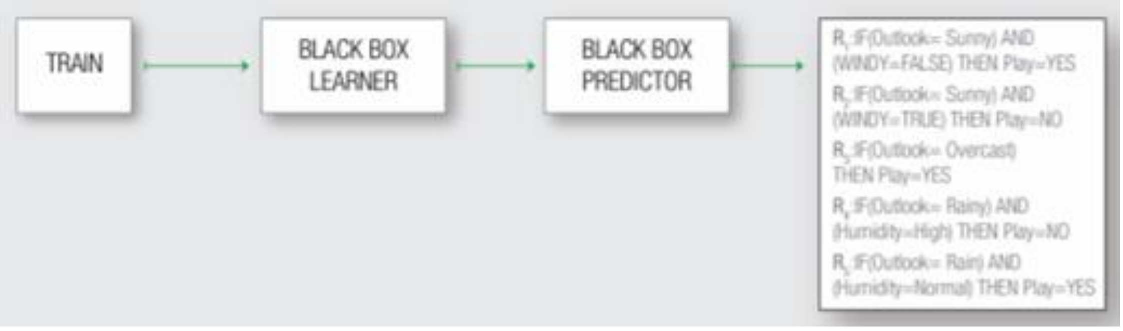

TRAIN
BLACK BOX LEARNER
BLACK BOX PREDICTOR
R_1: IF(Outlook= Sunny) AND (WINDY=FALSE) THEN Play=YES
R_2: IF(Outlook= Sunny) AND (WINDY=TRUE) THEN Play=NO
R_3: IF(Outlook= Overcast) THEN Play=YES
R_4: IF(Outlook= Rainy) AND (Humidity=High) THEN Play=NO
R_5: IF(Outlook= Rain) AND (Humidity=Normal) THEN Play=YES

Figure 9. Black-box model explanation through imitating the black box behavior by a set of rules [Guidotti et al., 2018].

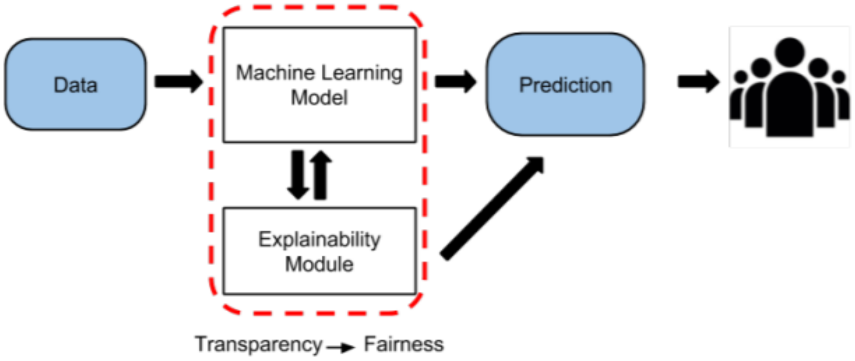


Figure 10. Black-box model explanation [Abdollahi and Nasraoui, 2018]. The machine learning transparency is achieved via explainability in the modeling phase.

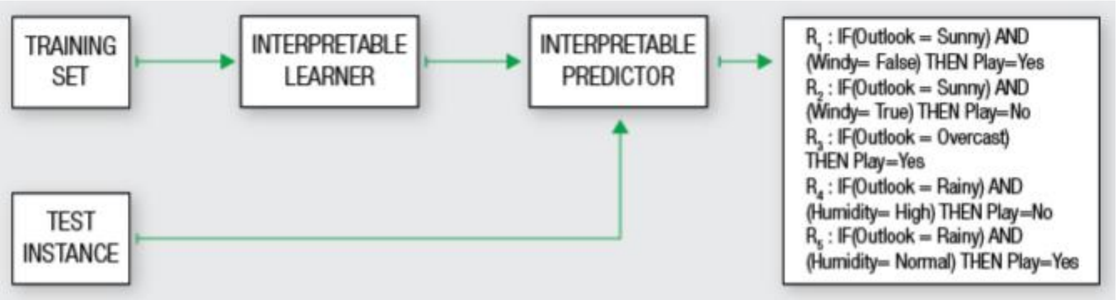

TRAINING SET
INTERPRETABLE LEARNER
INTERPRETABLE PREDICTOR
TEST INSTANCE
R_1 : IF(Outlook = Sunny) AND (Windy= False) THEN Play=Yes
R_2 : IF(Outlook = Sunny) AND (Windy= True) THEN Play=No
R_3 : IF(Outlook = Overcast) THEN Play=Yes
R_4 : IF(Outlook = Rainy) AND (Humidity= High) THEN Play=No
R_5 : IF(Outlook = Rainy) AND (Humidity= Normal) THEN Play=Yes

Figure 11. Black-box outcome explanation [Guidotti et al., 2018]. The received explanation is about the reason for choosing this particular instance and there is no explanation about the logic (process) behind the black box.

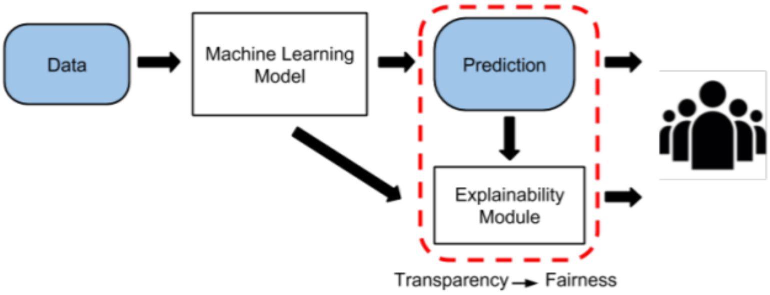


Figure 12. Black-box outcome explanation [Abdollahi and Nasraoui, 2018]. The ML transparency is achieved via explainability in the prediction phase.

- **Perceived Fairness Management.** The perceived bias of the outcome can impact the perceived fairness of the system. Perceived fairness can be measured through questionnaires and statistical tests [Lee, 2018]. It can also be affected by the system explanations [Binns at al., 2018] (whether for a model or for an outcome). A descriptive approach for identifying the notion of perceived fairness for machine learning was suggested by Srivastava et al. [Srivastava et al., 2019]. They argued that the perceived fairness of the user is the most appropriate notion of algorithmic fairness. Their results show that the formal measurement, demographic parity, most closely matches the perceived fairness of the users and that in cases when the stakes are high, accuracy is more important than equality.

3.4 Abstract Model for Algorithm Fairness and Transparency

The final element we introduce into our model is the stakeholders. We consider three broad classes of stakeholders of an algorithmic system – developer, user and observer. We note that there are a variety of users of a system with different perspectives. The bank officer and bank client are both users of the CREDIT_RATE system with completely different perceptions of fairness. Similarly there are a variety of observers (e.g. regulators, certification bodies, ethics committees) and a variety of roles in the developer class (architect, validation team, implementation team, product owner, marketing manager, etc.).

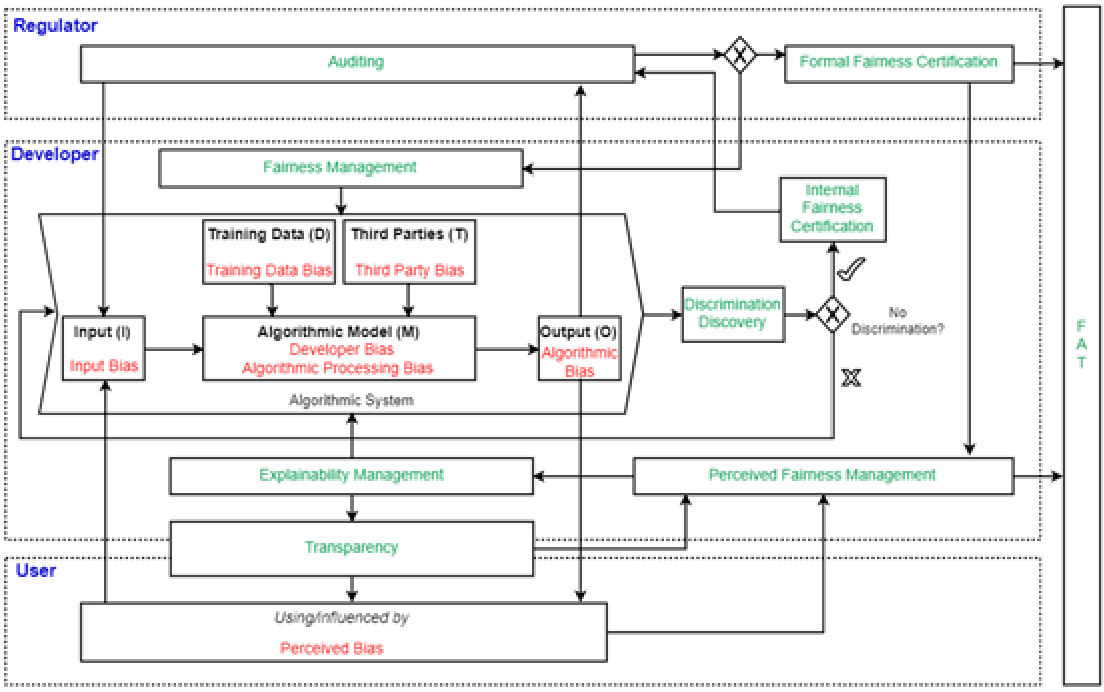


Figure 13. Algorithmic system true fairness (ASTF) framework. The framework considers different stakeholders (in blue) and their influence on the system (dashed boxes), the components of the algorithmic system (in **bold black**), the possible risks to transparency and fairness and their detection(in red), the suggested mitigation actions (in green). Arrows describe the flow of influences between different elements of the model. The system is considered to be **truly fair** if both fairness certification from the regulator and perceived fairness management are applied and approved.

An algorithmic system gets input from the user, that from her side implicitly embeds her perceived bias into this input (like a search query into a search engine). Then given the training data (with the population data bias in it), input (user's data bias) and the third parties (human biases) the algorithmic system applies the relevant algorithmic model (with the algorithmic processing bias) and as a result outputs an outcome (this time with algorithmic bias).

First, we will explain how developer and user process this output and affect the true fairness of the system and, in the end, we will elaborate on the influence of the regulator.

After the user sees the output of the system, she can consider whether to learn from it and use it to change the input to the system to get a better desired output or decide that the output is fair enough (with respect to her perceived fairness) and that in her opinion the system is fair. An output explanation can affect the transparency of the system and therefore can reduce the user's perceived bias and as a result – affect her perceived fairness and judgement regarding the system true fairness.

From the developer’s side she is responsible for fairness management and should run discrimination discovery on the output of the system. If discrimination is discovered on any component of the system, the whole process is repeated, but with the updated component where the discrimination was found. If no unintended discrimination discovered, she receives the internal fairness certification that is passed to the regulator.

The regulator is involved in the auditing process, where she checks compliance with pre-defined rules for transparency and fairness. She provides a set of inputs to the system and gets a range of outputs, by which the system is validated. Once the system successfully passes the regulator’s process (which could include a wide variety of verification and validation as well as testing techniques) and also given the internal fairness certification, the regulator can officially approve that the system is fair by providing the formal fairness certification. The system is fair if both formal fairness certification from the regulator and perceived fairness management are applied and approved.

Our framework aims to address fairness on a more general level as algorithmic system true fairness (ASTF) as can be seen in figure 14. The ASTF is affected by the following: auditing which is performed by the regulator, explainability and formal fairness of the system which is implemented and verified by the developer and the perceived fairness of the user.

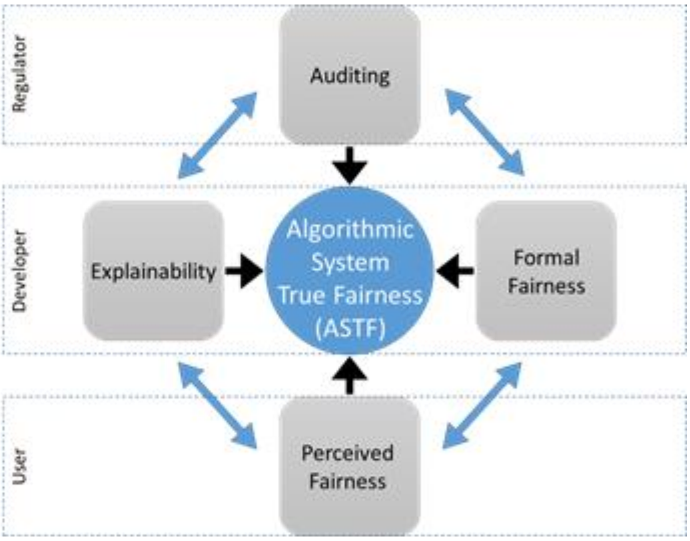


Figure 14. Solution space effect on algorithmic system true fairness (ASTF). Gray boxes represent the solution space components that lead to ASTF (black arrows point to blue circle). Blue arrows represent the relation between grey boxes and their relation to system stakeholders (dashed boxes).

3.5 Mapping the framework to the case studies

Case study 1: AD_SERV: An algorithm that recommends a list of advertisements to place on a webpage or mobile app. The recommendation is based on information about the identity and attributes of the viewer of the webpage, and their recent activity on that page or app. Each advertisement is equipped with a set of desired targets and with its history of display and click-through.

Stakeholders:

1. The **owner** of the AD_SERVE algorithm is an advertising agency that sells its services (usage charge for the algorithmic system) to advertisers and web content providers on the internet or mobile app owners. Its income is dependent on the successful targeting of end users of the web pages or apps.
2. The **web content provider** or **app owner** is paid by the advertiser if and when an ad displayed on their property is clicked. The web content provider or app owner pays the owner of the AD_SERVE algorithm for the right to use the system to populate their property with ads.
3. An **advertiser** is also a customer of the owner of AD_SERVE paying to list the ad and providing target directing information to the system.
4. The **end user** is the consumer of the web page or mobile app that is exposed selectively to ads from the repository of ads held by the owner of AD_SERVE. The information provided (not necessarily explicitly) by the end user to AD_SERVE could be anything from her location, her browsing history, her demographic details to the entire catalog of data publicly available on the internet, the advertiser's databases, the advertising agency's databases, the content provider's databases, to name just a few.

Components of the System

1. Input: Data provided by the specific end user, the specific content providers and advertisers.
2. Training Data: The algorithm is initially trained by data provided by the advertisers. It subsequently learns from the behaviour of all users, advertisers and content providers.
3. Third Party constraints: These constraints are supplied by the advertisers who target their marketing to specific market segments. Other constraints may be provided by the content providers who ban or encourage certain classes of advertisers from their sites or apps.
4. Algorithm: The algorithm provided by the owner attempts to maximize click through rates in order to satisfy its customers (the advertisers and content providers). It provides the third parties with ways to constrain the algorithm, but does not concern itself with fairness. It does however provide a way for the end-user to enquire "Why am I seeing this ad?" and provides an explanation of some sort.
5. Output: The algorithm provides a set of ads that are displayed on the content providers platform for a particular end user.

Risks to fairness and transparency:

1. **Explicit discrimination** in AD_SERVE certain forms of explicit discrimination may be observed due to third party constraints (e.g. Do not show my ad to female end-users.). Such discrimination may be legitimate if the advertiser is promoting male hygene products. It may also be evidence of improper discrimination, if the ad is for a position with a steel mill production line.

2. **Implicit discrimination** in an ad server has been noted by Sweeney (2013) who found that ads for research into criminal records were displayed disproportionately often on a search engine for when searching for names with a "black" preponderance (e.g. the name Latanya is predominately associated with black people in the USA, whereas "Jill" is not).

Mitigation of fairness and transparency risks

1. The system provides **Explainability Management** in the form of a response to the question "Why am I seeing this ad?". The response could be a simple "Inspired by your browsing history" which is a **Black Box Outcome Explanation.**
2. **Fairness Management** could be implemented for sensitive ads like those offering research into criminal records or other ads with potential for discriminatory display

Case study 2: CREDIT_RATE: An algorithm that recommends to a bank officer whether or not to grant a loan to a customer. The algorithm is based on information about the particular customer, the conditions of the loan requested, and a database of prior applicants both successful and unsuccessful, including the repayment behaviour of successful loan applicants.

Components of the System

1. Input: Data provided about the specific end user, by the bank officer the specific content providers and advertisers.
2. Training Data: The algorithm is initially trained using the bank's historical data.
3. Third Party constraints: These constraints are defined by the bank officers.
4. Algorithm: The algorithm implemented assesses the risks and benefits of approving the credit request given the historical data and the system configuration.
5. Output: The algorithm provides a decision whether to approve or deny the customer's request.

Risks to fairness and transparency:

3. **Explicit discrimination** may appear in the system has been configured to consider specific protected or proxi attributes as part of its reasoning (if this information is provided as input)..
4. **Implicit discrimination** may appear if the training set used by the system includes protected or proxi attributes and it is biased in the sense that these attributes correlate with final decisions..

Mitigation of fairness and transparency risks

3. The system provides **Explainability Management** in the form of an explanation of its decision as it is ablack box, hence **Black Box Outcome Explanation** is provided**.**
4. **Fairness Management** could be implemented for ensuring group and individual parity

Case study 3: COMPASS: An algorithm that receives personal characteristics of a convict and predict the probability of recidivism. In essence the details of the representation and examination of the system is similar to case study 2, but with much greater risk .

4. Stakeholders Body of Knowledge

We analyse the educational needs of each of the stakeholder groups on an operational level by considering which problems these users need to solve. The first of these problems is that of awareness that there is a problem in the first place. Each of the stakeholder groups needs to have an educational module focused on awareness of the existence and implications of algorithmic systems which may be biased and or opaque.

Once awareness has been achieved, the specific problems faced by the stakeholder groups differ due to their role in the creation and usage of algorithmic systems. Developers, regulators and owners have key roles in the design, implementation and deployment of systems. Owners and regulators need the skills to demand and test for fairness and transparency, while developers need the skills to implement these requirements and to take actions to discover and prevent bias as it occurs in the development process. End users need skills to detect bias and interpret documents explaining the transparency features of the systems. [Eiband et al. 2018].

This leads us to the following classification of educational modules and their targets:

- **F&T** (Fairness and Transparency) **awareness** for all stakeholders - all stakeholders need to be aware of the potential biases and risks of algorithmic systems;
- **F&T requirements specification** for owners and regulators who need to produce such specifications, and developers who need to understand and implement them;
- **Discrimination and unfairness detection/testing** for all stakeholders – including end users;
- **Discrimination and unfairness correction tools** for developers
- **Explainability techniques** for developers to produce transparency evidence and for all other stakeholders to interpret such evidence.

The contents of these education modules should be varied according to the stakeholders' technical abilities and operational needs.

5. Glossary of terms and the "need to know"

The following table indicates the relevant concepts regarding algorithmic system fairness that each stakeholder should be familiar with. For example, the end user should be aware that a system can be discriminatory and that there are different ways to detect and mitigate the discrimination, without having to know all the technicalities. The table is meant to be indicative, but there may be exceptions for specific stakeholders in specific systems with different risks. For example, medical professionals using a diagnostic system need to know more concepts than a movie viewer needs to know about the Netflix recommender system.

Concepts	**Owner**	**End user**	**Developer / Auditor**	**Observer /Regulator**	**Educator**

General		Algorithmic System Model (Components)	x	x	x	x	x
		Algorithmic Transparency	x	x	x	x	x
		Computational Regulations (GDFR)	x	x	x	x	x
		Diversity	x	x	x	x	x
		Explainability	x	x	x	x	x
		F&T Awareness	x	x	x	x	x
Problem Space	**Biases**	Algorithmic Bias		x	x	x	x
		Algorithmic processing Bias			x	x	
		Developer Bias			x	x	
		Input Bias	x	x	x	x	
		Perceived bias	x	x	x	x	x
		Third Party Bias	x		x	x	
		Training Data Bias			x	x	
Solution Space	**Discrimination Discovery**	Discrimination Correction			x		
		Discrimination Detection	x	x	x	x	x
		Explicit Discrimination			x		
		Implicit Discrimination			x		
		Protected (sensitive) Attributes			x		
		Protected Group			x		
		Proxy Attributes			x		

	Fairness Promotion	F&T Requirements	x		x	x	x
		Fairness Certification			x	x	
		Fairness Formalization			x	x	
		Fairness learning			x	x	
		Fairness Sampling			x	x	
		Group Fairness/Parity			x		
		Individual Fairness			x		
	Auditing	Auditing methods			x	x	
	Explainability Management	Black Box Explanation	x	x	x	x	x
		White box explanation	x	x	x	x	x

6. Index of terms

7. References

1. Abdollahi, B., & Nasraoui, O. (2018). Transparency in fair machine learning: The case of explainable recommender systems. In Human and Machine Learning (pp. 21-35). Springer, Cham.
2. ACM Statement. Statement on Algorithmic Transparency and Accountability https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf
3. AI fairness 360. https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/
4. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.
5. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018, April). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (p. 377). ACM.
6. Chiu, C. M., Lin, H. Y., Sun, S. Y., & Hsu, M. H. (2009). Understanding customers' loyalty intentions towards online shopping: an integration of technology acceptance model and fairness theory. Behaviour & Information Technology, 28(4), 347-360.
7. d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. Big data, 5(2), 120-134.
8. Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. In IJCAI (pp. 4691-4697).
9. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226). ACM.
10. Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018, March). Bringing Transparency Design into Practice. In 23rd International Conference on Intelligent User Interfaces (pp. 211-223). ACM.
11. Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017, May). “Be careful; things can be worse than they appear”: Understanding Biased Algorithms and Users’ Behavior around Them in Rating Platforms. In Eleventh International AAAI Conference on Web and Social Media.
12. Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184.
13. Giunchiglia, F. (2006, September). Managing diversity in knowledge. In IEA/AIE (p. 1).
14. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 93.
15. Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., & Wilson, C. (2017, February). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In

Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 1914-1933). ACM.

16. Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P., & Weller, A. (2018). Blind justice: fairness with encrypted sensitive attributes. arXiv preprint arXiv:1806.03281.
17. Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. Psychological review, 107(4), 852.
18. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems (pp. 4066-4076).
19. Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica (5 2016), 9.
20. Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, 5(1), 2053951718756684.
21. Madaan, N., Mehta, S., Agrawaal, T., Malhotra, V., Aggarwal, A., Gupta, Y., & Saxena, M. (2018, January). Analyze, detect and remove gender stereotyping from bollywood movies. In Conference on Fairness, Accountability and Transparency (pp. 92-105).
22. Maltese, V., Giunchiglia, F., Denecke, K., Lewis, P., Wallner, C., Baldry, A., & Madalli, D. On the interdisciplinary foundations of diversity. Dipartimento di Ingegneria e Scienza dell'Informazione, 2009, University of Trento. technical report DISI-09-040. 20.5 Future Issues http://eprints. biblio. unitn. it/archive/00001651/01/040. pdf.
23. Pedreschi, D., Ruggieri, S., & Turini, F. (2009, June). Integrating induction and deduction for finding evidence of discrimination. In Proceedings of the 12th International Conference on Artificial Intelligence and Law (pp. 157-166). ACM.
24. Pohl, R. (Ed.). (2004). Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory. Psychology Press.
25. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). ACM.
26. Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review, 29(5), 582-638.
27. Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., & Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. arXiv preprint arXiv:1811.05577.
28. Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry, 22.
29. Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, Race, & Recidivism: Predictive Bias and Disparate Impact.(2016).
30. Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F., Arvanitakis, G., Benevenuto, F., ... & Mislove, A. (2018). Potential for Discrimination in Online Targeted Advertising Till

Speicher MPI-SWS MPI-SWS MPISWS. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*) (Vol. 81, pp. 1-15).

31. Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. arXiv preprint arXiv:1902.04783.
32. Sweeney, L. (2013). Discrimination in online ad delivery. Queue, 11(3), 10-29.
33. Tal, A. S., Batsuren, K., Bogina, V., Giunchiglia, F., Hartman, A., Loizou, S. K., Kuflik, T. & Otterbacher, J., "“End to End” Towards a Framework for Reducing Biases and Promoting Transparency of Algorithmic Systems," 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Larnaca, Cyprus, 2019, pp. 1-6. doi: 10.1109/SMAP.2019.8864914
34. Torralba, A., & Efros, A. A. (2011, June). Unbiased look at dataset bias. In CVPR (Vol. 1, No. 2, p. 7).
35. Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) (pp. 1-7). IEEE.
36. Volker, M. S. A. V. P. (2003). Sharing expertise: Beyond knowledge management. MIT press.
37. Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. Journal of Management Information Systems, 23(4), 217-246
38. Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. arXiv preprint arXiv:1507.05259.
39. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). Learning fair representations. In International Conference on Machine Learning (pp. 325-333).
40. Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (2019). Machine Learning Testing: Survey, Landscapes and Horizons. arXiv preprint arXiv:1906.10742.