2022 International Conference on Multidisciplinary Research
https://doi.org/10.26803/MyRes.2022.24

Towards an Integrated Approach for Preserving Data Utility, Privacy and Fairness

Mortaza S. Bargh
Rotterdam University of Applied Sciences,
Rotterdam, The Netherlands

Sunil Choenni
Research and Documentation Center (WODC),
Ministry of Justice and Security,
The Hague, The Netherlands

Abstract
Data reusability has become a distinct characteristic of scientific, commercial, and administrative practices nowadays. However, an unlimited and careless reuse of data may lead to privacy breaches and unfair impacts on individuals and vulnerable groups. Data content adaption is a key aspect of preserving data privacy and fairness. Often, such content adaption affects data utility adversely. Further, the interaction between privacy protection and fairness protection can be subject to making trade-offs because mitigating privacy risks may adversely affect detecting unfairness and vice versa. Therefore, there is a need for research on understanding the interactions between data utility, privacy and fairness. To this end, in this contribution, we use concepts from causal reasoning and argue for adopting an integrated view on data content adaption for data driven decision support systems. This asks for considering the operation context wholistically. By means of two cases, we illustrate that, in some situations, local data content adaption may lead to low data quality and utility. An integrated wholistic approach, however, may result in reuse of the original data (i.e., without content adaption, thus in higher data utilization) without adversely affecting privacy and fairness. We discuss some implications of this approach and sketch a few directions for future research.

Keywords: algorithmic fairness, bias, causal reasoning, data utility, privacy, trade-offs.

1. Introduction

There is a growing interest in the utilization of data by reusing the collected data for various purposes, as we witness by the rise of open data, data sharing initiatives and data-driven applications. In fact, data reusability has become a distinct characteristic of scientific, commercial, and administrative practices nowadays. It enables an evidence-based optimization of practice, allows reanalyzing the existing evidence for

verifying previous results, and minimizes effort duplication via building on the work of others. However, it has been recognized that an unlimited and careless reuse of data may lead to privacy breaches and unfair impacts on individuals and groups (see, e.g., Bargh and Choenni, 2013).

Often data sets contain too much personal data than the amount needed for the data reuse purpose. Having this excessive personal data stems from the ways that these data sets are collected and used. Sometimes, like when conducting statistical analysis and/or scientific research, the collected data contains too much personal data due to inappropriate research design. Excessive personal data can also be resulted from linking data items in various data sets. There are other times where data sets are collected for one purpose but are used for another legitimate purpose. Examples of such cases are to reuse patient data, which are collected to document medical treatments, for medical research, or to reuse offender data, which are collected to substantiate judicial procedures, for criminology research. In addition to having excessive personal data, the collected data may have biases due to the data representing an unjust and unjustifiable situation in the real world (caused by, for example, systemic, systematic, structural, and institutional discrimination) or the data representing a justifiable situation unfaithfully (due to, for example, erroneous observations or unrepresentative data sampling). In addition to biased data, the way that the data is processed and/or the result of data processing is interpreted may lead to unfair impacts on individuals and groups if the outcomes of data processing are applied to practice carelessly.

An important concern in (re)using data is how to deal with excessive personal data and its unfair impacts. According to privacy laws and regulations, like General Data Protection Regulation (GDPR, 2016), it is necessary to minimize the amount of personal data in data sets to the data needed for the (legitimate) data usage. Not adjusting the amount of personal data to the data usage purpose can lead to privacy breaches and may impact individuals, groups, and society adversely. It may also inflict reputation damages upon organizations responsible for such privacy breaches, bring lawsuits against them, and impose financial penalties on them. GDPR has a comprehensive view on personal data and considers it as any information that relates to an identified or identifiable natural person. Specifically, "an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person", see Article 4 of GDPR (2016). Further, one of GDPR principles regarding the processing of personal data is that it should be "processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency')", see Article 2(1) of GDPR (2016). On the other hand, GDPR allows for processing personal data for legitimate purposes such as archiving purposes in the public interest, for scientific or historical research purposes, or for statistical purposes. This processing, however, should be subject to appropriate safeguards by ensuring that technical and

organizational measures are in place (particularly with respect to personal data minimization and fair data processing).

Data content adaption is one of the key aspects of preserving data privacy and fairness. Often such content adaption affects the data utility adversely. This adaption may range from being very strict to being loose. The former, which asks for a severe adaption of data content such that none of the original data characteristics are preserved, leaves limited room for data reuse. The latter, which asks for a slight or no adaption of data content, leaves the data open to privacy and fairness threats.

Today, to realize GDPR appropriately, organizations search for a balanced way of preserving data utility while protecting against privacy and fairness threats. In this paper, we aim at the adaptation of the data content such that, on the one hand, the quality of data content is good enough for the purpose in mind and, on the other hand, the data content does not disclose illegitimate privacy-sensitive information nor lead to unfair treatment of individuals. More specifically, our objective is to seek for a systematic approach for making trade-offs among data utility (e.g., precision and accuracy of predictions), privacy disclosure risks, and unfairness risks. It is worthwhile to note that also the interaction between privacy protection and fairness protection can be subject to making trade-offs (Balayn et al., 2021). As (some of) the privacy sensitive and fairness sensitive attributes are common (like age and gender), adapting such attributes for mitigating privacy risks may adversely affect the use of such attributes for detecting unfairness in observed data. Therefore, scholars and practitioners emphasize the need for research on understanding the interactions between privacy and unfairness, see (Balayn et al., 2021; Jagielski et al., 2019).

In this paper, in line with (Choenni et al., 2018), we argue for the importance of context for data content adaption for data-driven decision support systems. By means of two cases, we show that an adequate description/modelling of the operation context in which the data will be reused can result in higher data utilization compared to the case where the context is missing. In the first case, data will be adapted in a data preprocessing stage without taking the operation context into account. It will be shown that the quality of data degrades substantially if one wants to preserve privacy and fairness adequately. For the second case, we provide an example where adapting the data content is unnecessary. To substantiate the second case, we propose extending causal models with additional variables that represent the operation context (i.e., data usage context and data usage impact). Using this extension, we provide an argumentation structure for (not) using the original data (with a high data utility) without jeopardizing the privacy and fair treatment of individuals and groups. These cases indicate the importance of approaching the data content adaption problem in a way that data processing and the outcome usage (i.e., operation) are considered together (i.e., having an integrated approach for data content adaption).

The remainder of this paper is organized as follows. In Section 2 we explain data-driven decision-support systems to describe the problem setting. Subsequently in Section 3, we present the (theoretical) principles used in the paper. In Section 4, as a

benchmarking example, we describe a commonly used strategy to modify data content for protecting privacy and fairness. In Section 5, we use causal reasoning theory and propose some extensions to traditional causal models to argue in favor of or against using protective attributes in some situations. We discuss the results in Section 6 and draw conclusions in Section 7.

2. Problem setting: data-driven decision-support systems

A data-driven decision-support system exploits available data to provide organizations and humans insights to take an informed decision and carry out an operation or action in the real world. To this end, as shown in Figure 1, the decision-support system observes some input data attributes denoted by X and possibly an output attribute denoted by Y. During its training phase, the system builds a model $\mathcal{M}$ based on some already observed instances of attributes X and Y, denoted by $\boldsymbol{X}$ and $\boldsymbol{Y}$. During its operation phase, the system uses the (latest) model $\mathcal{M}$ together with the current values of input variables X to yield a prediction of the output variable Y, denoted by $\hat{Y}$. This is called interpreting the model for the current input data point X in the figure. Subsequently, the predicted output $\hat{Y}$ is used together with some contextual information (possibly by a human agent) to create an action, affecting a real-world phenomenon. The latter is called outcome interpretation (possibly by a human agent) in the figure. We use the term *extended data-driven decision support system* to refer to the whole process of making decisions and acting (i.e., the combination of the data-driven decision support system and the human intelligence), as shown in Figure 1.

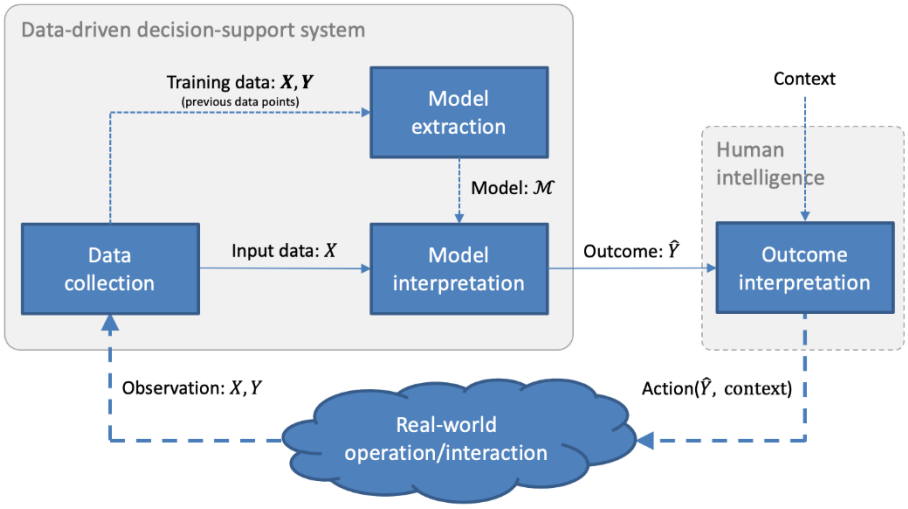

Figure 1: An illustration of an extended data-driven decision support system

The use of (already available) data in extended data-driven decision support systems should be done in a responsible way. To this end, many protection mechanisms should be established such as those for security and privacy, explainability (and interpretability), and fairness (Chorąś et al., 2020). As mentioned above, we focus on those protections that deal with data content adaption. Such mechanisms mainly affect the data utility, privacy and fairness aspects of responsible data-driven systems).

A closely related concept to fairness is bias. The observed data (i.e., X and Y) and, consequently, the predicted outcomes $\hat{Y}$ and decisions Action($\hat{Y}$, context) – whether made by people or systems – may show bias. According to (Balayn et al., 2021), a bias exists when certain data classes have different distributions for the values of some label attributes (e.g., output attribute Y and/or its prediction $\hat{Y}$) systematically. A data class represents a group of data instances, which typically share certain attribute values (e.g., the data instances representing females between 30 and 40 years old). A bias can be problematic or not, which is mostly established based on human judgment (Balayn et al., 2021). As such, Balayn et al. (2021) distinguishes three bias types, namely:

- Desired bias, referring to those that are part of correct system functionality.
- Undesired bias, relating to the classes of protected attributes (like, gender, race, religion and sexual orientation). Protected attributes are considered as sensitive according to laws and/or societal or ethical norms. This bias type is often perceived as *unfair* by the stakeholders, especially those impacted by the system.
- Unimportant bias, referring to those not being problematic according to laws or societal discourse. This bias often relates to the classes defined by contextually meaningless attributes, for example, the data class representing individuals wearing sunglasses and T-shirts.

The undesired bias in the data collection stage in Figure 1 can be due to the data representing a situation in the real world that is undesirable (due to, for example, systemic, systematic, structural and institutional discrimination), or the data representing a desirable situation unfaithfully (due to, for example, erroneous observations or unrepresentative data sampling), see (Choenni et al., 2018) for an example. The output of the data processing stage (i.e., the model extraction and model interpretation blocks in Figure 1) can be biased due to, for example, the use of unfit and unsuitable algorithms to process the data, e.g., for making output bias and variance trade off (Bishop, 2006).

3. Relevant concepts and principles

In this section we describe the (theoretical) foundations needed for our discussions in the following sections.

A data processing model and its extension

Dealing with privacy and fairness issues can be done at various parts of the extended data-driven decision support system shown in Figure 1. The measures dealing with fairness issues are traditionally categorized at three stages of a data driven-data decision support system, namely before, within and after a data analytics algorithm (Balayn et al., 2021; Berk et al., 2021). Specifically, as shown in Figure 2, the fairness protection measures are categorized as

- Pre-processing measures, which are applied to the input data (i.e., to X and Y),
- In-processing, which encompass the in-algorithm treatments, and
- Post-processing, which are applied to the output data (i.e., outcome $\hat{Y}$).

In literature fairness measures taken in the pre-processing stage are part of data management domain (or within data engineering community) and those taken in the in-processing and post-processing stages are part of data analytics domain (or within data science community).

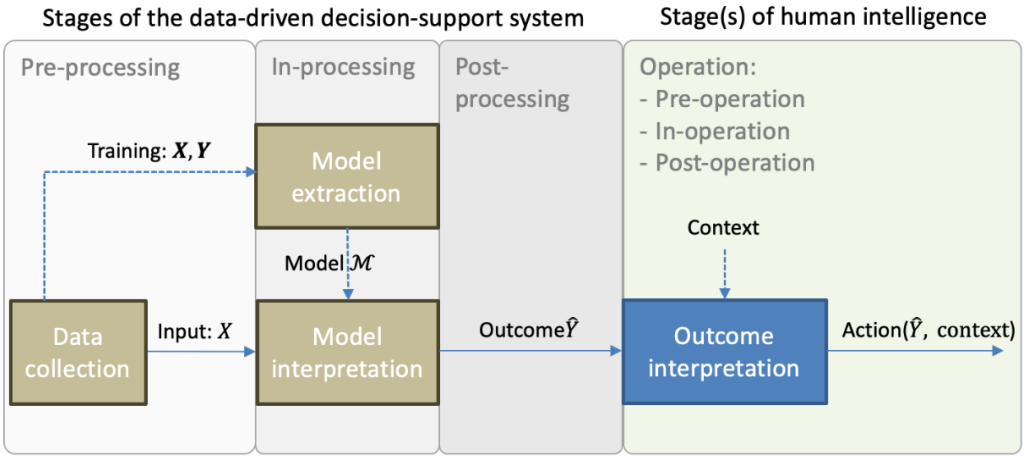

Figure 2: An illustration of the issue discovery and mitigation stages in an extended data-driven decision support system

One can also categorize privacy protection measures in the three stages mentioned above (although there might be no privacy protection measure categorized in the in-process stage yet). Therefore, we propose the stages of a data-driven decision support system, as shown in Figure 2, to be applicable to both fairness and privacy protection measures. Further, as shown in Figure 2, we argue for and advocate adding a new stage corresponding to the operational stage of a data-driven decision support process, wherein an Action($\hat{Y}$, context) is defined and carried out. The action Action($\hat{Y}$, context) is possibly based on some human intelligence, which interprets the outcome of the decision support system $\hat{Y}$ based on the situation (i.e., the operation context like the objective and the relevant environment characteristics). In turn, the measures in the operation stage can further be divided into pre-operation, in-operation and post-operation substages (corresponding to what should be done before, during and after applying an action to a real-world phenomenon). An elaboration of such subdivision is the subject of our future research.

Dealing with privacy and fairness issues in each stage in Figure 2 requires conducting two types of actions: detecting the issues and subsequently mitigating them. Having such a distinction is inspired by the literature analysis of (Balayn et al., 2021) that results in identifying six main directions of research on unfairness, which we suspect to be applicable for privacy as well, namely:

1. To define, formalize and measure unfairness,
2. To identify cases on unfairness in datasets,
3. To develop ways to mitigate the unfairness within such datasets,
4. To test unfairness in the outputs of machine learning based software systems,
5. To understand how humans perceive the unfairness of data-driven decision-support systems, and
6. To investigate how humans might create certainty, given the biases found in the outputs of the systems.

Clearly, the directions mentioned in items 2, 4 and 5 relate to the detection aspects and those in items 3 and 6 relate to the mitigation aspects. Considering the stages identified in Figure 2, we attribute a pair of discovery and mitigation actions to each stage. For example, in Section 4, we analyze a scheme from literature for detecting and mitigating both fairness and privacy issues simultaneously in the pre-processing stage. There are also suggestions to detect such issues in early stages and mitigate them in later stages, see for example the vision depicted in (Stoyanovich, et al. 2017) and indicated in Figure 3. The literature does not provide guidance in the selection of the detection and mitigation methods, i.e., which method(s) and where to apply. As mentioned in (Balayn et al., 2021) about where to apply detection and mitigation: "it seems to primarily depend on the notion of fairness to optimize for, and on the actual context of the application." To provide evidence for this context dependency, we sketch a scheme in Section 5 that considers the impact of actions at the operation stage and underlines the situations that do not require any detection and mitigation.

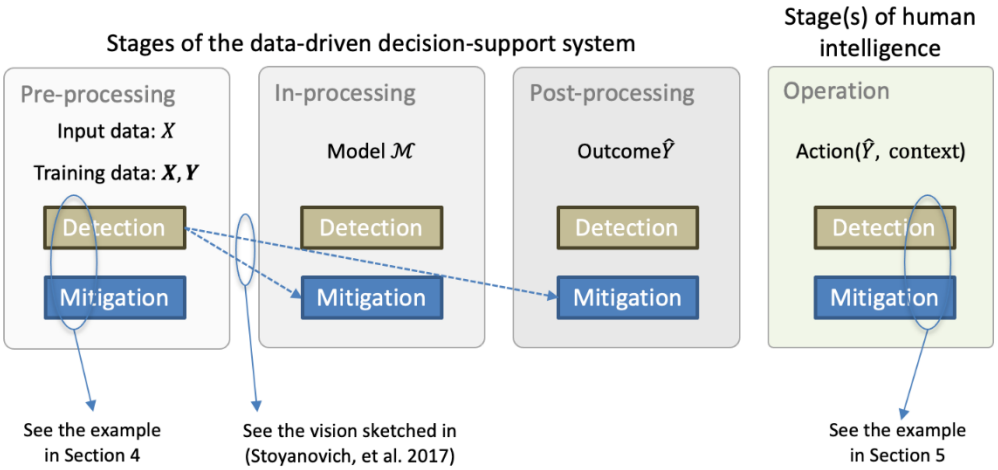

Figure 3: An illustration of the detection and mitigation measures taken in various stages of an extended data-driven decision support system

Personal data minimization

A core principle of privacy protection is to minimize the amount of personal data in the input and output data sets of data-driven applications to the level needed for the data usage in mind. The current fast growth of data and data-driven applications demands for using efficient mechanisms to detect personal data and to minimize it to the level needed. Personal Data Minimization (PDM) is generally done in the pre-processing stage on raw data (e.g., microdata) and/or in the post-processing stage on aggregated data (e.g., tabular data) or query replies. In this contribution we focus on microdata sets, which are structured as tables with rows, representing individuals, and columns, representing attributes (or features) of individuals like their age, gender and occupation. More formally, the columns of microdata set D_N, which comprises N records/rows, correspond to either the observed input attributes X or output attribute Y mentioned in the previous section. So, every record $d_n \in D_N$, where $n: 1, \ldots, N$, comprises M attributes denoted by a_m, where $m: 1, \ldots, M$. Attribute a_m assumes a nominal or ordinal value from domain $dom(a_m)$. A record d_n is in fact an element of super domain $dom(a_1) \times dom(a_2) \times \cdots \times dom(a_M)$, as a Cartesian product of the individual domains over which all attributes are defined.

One can distinguish between two main approaches for Personal Data Minimization (PDM): Syntactic and noise-based (Clifton & Tassa, 2013).

- Syntactic approach typically relies on generalizing data items until a syntactic condition is met. This means preserving the truthfulness of data in the sense that a modified value clearly specifies the group of possible original values. For example, when the original value of attribute age is generalized to the age range of 16-20, the data processor knows that the original value is surely one of the following values: 16, 17, 18, 19 and 20. The objective in syntactic methods is to restrict the ability of intruders to learn with a high enough certainty the identity of or private information about a data subject from a transformed microdata set. Example of the syntactic approach are k-anonymity and its complementary variants (like l-diversity, t-closeness) and HIPPA rules (Bargh et al., 2018; 2021; Annas, 2003).
- Noise-based approach: the PDM methods of this approach are not based on a syntactic condition. They add noise or randomness to a transformed data set (or to the outcome of a calculation on the original data set). The objective here might be to hide the influence of the data of a subject in the transformed data set (or in the outcome) and, as such, to preserve the subject's privacy. ε-differential privacy is an example method of the noise-based approach.

Syntactic models are in use for a while. This indicates their robustness and acceptance by stakeholders (although this does not imply that they are perfect). This wide acceptance can be attributed to, among others, their truthfulness preservation (Li et al., 2007; 2011) and their conformance with existing privacy laws and regulations. Truthfulness preservation means that the transformed data items are consistent with the original values (like, as mentioned above, age 19 years old changes to an age range like 16-20 years old). Conformance with existing privacy laws and regulations means that, like privacy laws, syntactic models aim at preventing identifiability and attribution through considering how pieces of information interact (Nissim & Wood, 2018). Therefore, we have considered syntactic models for our study.

Attribute mapping is one of the key steps of the family of k-anonymity, l-diversity and t-closeness methods that are based on the syntactic approach (Bargh et al., 2018; 2021). Attribute mapping refers to the process of assigning a type to every attribute in a microdata set. Depending on the contextual constraints and conditions, via attribute mapping one categorizes the attributes $a_1, a_2, \ldots, a_M$ of a microdata set into the following types:

- Explicit Identifiers (EIDs) are those attributes in the original data set that structurally and on their own could uniquely identify an individual, i.e., a data subject. Examples of EIDs are a data subject's name, home address and unique personal identification numbers like the 'social security number', 'national health service number', 'voter card identification number', or 'permanent account number'. Often EIDs in the original microdata set are removed (i.e., filtered out), replaced with an unrecognizable value (i.e., masked/suppressed), or replaced with a unique and unrecognizable value (i.e., pseudonymized).

- Quasi Identifiers (QIDs) are those attributes that, in combination, could potentially be used to identify an individual if these QIDs are found in other data sources together with EIDs (or anything that specifies or points to someone specific). The values of the QIDs can be used to link the EIDs that exist in the other data source with (some of) the records of the transformed microdata sets. Hereby (some of) the records in the original microdata set – even if its EIDs are removed – can be reidentified. An example of QIDs is the combination of birthdate, postal code and gender, as shown in Sweeney (2000, 2002). QIDs are usually generalized, suppressed or aggregated. As a result, the QIDs in the transformed microdata set assume more coarse values. Every pattern of QIDs values, which is often common among a few records in the transformed microdata set, is called an Equivalence Class (EC).
- Sensitive Attributes (SATs) are those attributes that capture privacy-sensitive information about data subjects that we (possibly) do not want to disclose. These attributes are only present in the original microdata set and, thus, they are not present in other data sources which reside in other domains than that of the data controller organization. Examples of potential SATs are disease, salary, loans, disability status, and crime type. Determining SATs is a subjective and case-specific matter. Some legal regimes provide a list of sensitive attributes but note that not all of them are considered as SATs (i.e., a legally sensitive attribute, like gender, can be seen as a QID). SATs are typically important to data processors for data analytics purposes, and they may disclose personal data due to attribution. Therefore, SATs are usually protected via suppression.
- Non-sensitive Attributes (NATs) are all attributes that are (a) not EIDs, QIDs or SATs and (b) supposed to be published because they are needed for the purpose at hand. NATs are usually unprotected.
- The other attributes, unlike QIDs, SATs and NATs, are not needed to be shared, considering the data usage in mind. This can be due to their high privacy sensitivity (i.e., having a huge adverse impact on data subjects) and/or irrelevancy for the data sharing purpose in mind. If we decide not to share some attributes due to their (privacy) sensitivity, we must make sure that the published attributes (QIDs, SATs and NATs) do not leak information about such omitted privacy-sensitive attributes.

In summary, through attribute mapping, attributes $a_1, a_2, \dots, a_M$ of microdata set D_N are divided into 4 disjoint EID, QID, SAT and NAT subsets. To protect QIDs, the k-anonymity method is used via generalization of the values of QIDs and possibly suppression of some records. To generalize QID values, the appropriate taxonomy trees for QIDs are defined (Fung et al., 2010). The proportion of the number of suppressed records to the total number of records can be kept below a value to maintain the quality of the transformed microdata set.

Fairness

Fairness is a core concern of data-driven decision support systems, next to security/privacy and explainability concerns (Choraś et al., 2020). Currently, fairness

protection has gained importance in developing sociotechnical algorithmic systems (Starke et al., 2021). For sociotechnical systems see (Bargh & Troxler, 2020). The notion of fairness can be traced to (or found in) the fields of philosophy, sociology and legal sciences. These notions of human fairness are nowadays applied to algorithms (Starke et al., 2021) to align the corresponding systems with actual fairness values (Balayn et al., 2021). There is a vast body of literature on formalization of the concept of algorithmic fairness which can be categorized as pre-processing, in-processing and post-processing (Balayn et al., 2021; Berk, et al., 2021) as briefly explained in the following.

Pre-processing based methods aim at eliminating any sources of unfairness in the data before applying it to an algorithm. For example, all linear dependency between A and X attributes can be removed, some values of output attribute Y can relabeled to make per group base rates comparable. In-processing based methods build fairness adjustments into the algorithm. For example, uncertain risk forecasts (those that are close to 50% in case of binary outcomes) can be appropriately altered (e.g., by altering the forecast of high risk to low risk to serve some fairness goal), a penalty term can be added to the fitting procedure, or some constraints can be imposed to the optimization process. In post-processing based methods, the output of the algorithm is adjusted to make it fairer. For example, the value of an estimated outcome can be randomly reassigned so that the estimated outcome values (e.g., high risk or low risk) are independent of protected groups.

Note that algorithmic fairness measures concern group or individual fairness, depending on whether the measure indicates fairness per group (e.g., similar groups should have similar outcomes) or per individual, respectively (e.g., similar individuals should be treated similarly independently of their membership to one of the groups), see also (Balayn et al., 2021).

4. Adapting input data (a pre-processing case)

Adapting data before its processing is necessary when the data is shared with others, especially those from other organizations. Of course, for use and storage of data within own organizations, adapting data might be necessary when one knows that preserving (and accessing) the original data will not be necessary anymore. This is indeed an evidence of a due care practice. When sharing data with the public or other organizations, the purpose of data usage is often not specified beforehand and therefore the data should be protected against the worst-case scenario (i.e., protected against all privacy and fairness risks as much as possible and as much as acceptable).

Hajian et al. (2014) use the generalization method to protect microdata against both privacy and fairness risks. Unlike the other existing pre-processing methods for discrimination prevention which are based on data perturbation (like those mention for noise-based approach in Section 3), the data generalization method transform data truthfully and consistently (Li et al., 2007). Via generalization of QID attributes one can realize the k-anonymity method (i.e., protecting against privacy risks) and via

generalization of Potentially Discriminatory (PD) attributes one can realize the so-called α-protection method (i.e., protecting against fairness risks).

PD attributes refer to those sensitive attributes in $\{a_1, a_2, \dots, a_M\}$ of a microdata set D_N that explicitly require protection against discrimination according to (privacy) laws, regulations, and social norms. For example, the special categories[1] of personal data in GDPR might be seen as PD attributes or U.S. federal laws (US EPA, 1963) prohibit discrimination based on race, color, religion, nationality, sex, marital status, age and pregnancy. The generalization of attributes is applied to the union of QID and PD such that k-anonymity is achieved among QID attributes and α-protection is established for PD attributes. It might be that some attributes are neither QID nor PD but are highly correlated with PD attributes. These so-called proxy PD attributes should also be generalized so that the union of PD and proxy PD attributes become α-protected. For a technical description of α-protection, the interested reader is referred to (Hajian et al., 2014). A simple strategy for generalizing PD and proxy PD is to generalize them to the root values in their taxonomy trees (e.g., for the gender attribute, the values male and female are generalized to *). This simple strategy resembles Fairness Through Unawareness (Balayn et al., 2021).

In summary, for protection against privacy risks, generalization is applied to only QIDs. For protection against bias and discrimination (i.e., applying α-protection), generalization should also be applied to those PD and proxy PD attributes that are not QIDs. This extra level of generalization reduces the overall data utility for the combined protections (i.e., against both privacy and fairness risks) relative to that of either protection against privacy risks or protection against fairness risks. This reduction of data utility can be the highest if the data is going to be made open (accessible to everybody) because the parameters k and α should be chosen conservatively.

5. Acting in the operation stage (based on outcome interpretation)

Without loss of generality, we assume that the data processing stage of a data-driven decision support system is realized by a machine learning system. The parameters of the resulting system are denoted by:

- A: The set of observable attributes protected[2] by laws, regulations or social norms,
- X: The set of the other observable attributes,
- U: The set of relevant latent attributes, which are not observed,
- Y: The observed attribute to be predicted (possibly contaminated with historical biases), and
- $\hat{Y}$: The predictor random variable of Y that depends on A, X and U.

[1] Article 9(1) of GDPR (2016) on processing of special categories of personal data: "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited."

[2] Note that these attributes can be protected against privacy or fairness risks.

A machine learning based extended data-driven decision support system uses predictions $\hat{Y}$ to create some impacts, denoted by $g(\hat{Y}, Context_g)$, $h(\hat{Y}, Context_h)$, …, in the real-world. Here variable "$Context_{impact}$" captures all the contextual parameters that, next to the predicted outcome, are used to create a social, business, or economic impact. For example, let $\hat{Y}$ be a forecast of the volume of crime per neighborhood in the near future. Let, as hypothetical cases, $g(\hat{Y}, Context_g)$ be the expected jail capacity in a country and $h(\hat{Y}, Context_h)$ be the forecasted police patrol capacity in a neighborhood. In the following, we use the *causal reasoning* theory to reason about how to approach and deal with the former case at the operation level. For the latter case, which can be seen as a potentially problematic case according to the causal reasoning theory, one needs to adopt an elaborated approach (which is out of the scope of this contribution).

Causal reasoning approach

Causal reasoning uses an acyclic directed graph called Structural Causal Model (SCM) to indicate the casual dependency/relationship of random variables (or attributes) in a machine learning system (Kilbertus et al. 2017; Kusner et al., 2017). As an example, the SCM of the example case of gender discrimination in Berkeley college admissions – for more information see (Kilbertus et al. 2017) – is shown in Figure 4. In this example, a lower college-wide admission rate is observed for women than for men. This is because women, compared to men, apply for more competitive departments (i.e., therefore A has impact on X as shown in the SCM model in Figure 4) and the department choice impacts the admission rates (thus, X impacts Y). According to this model, the admission rate is not just affected by the gender. Thus, a conclusion like being a woman automatically leads to a lower acceptance rate does not hold because there is another reason behind the lower acceptance rate of women (i.e., the department choice). In Figure 4, we have extended the SCM model of this example with the dashed parts to indicate a possible impact of outcome Y (particularly on women). This extension will be used in our arguments in the following.

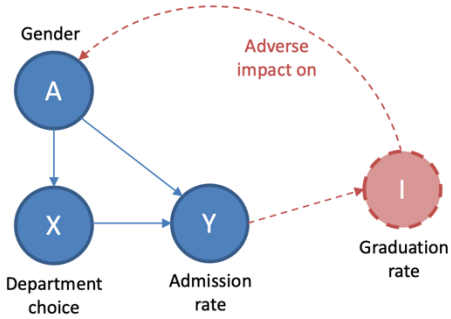

Figure 4: An example of a SCM from (Kilbertus et al. 2017), extended with the impact aspects (the dashed parts)

An SCM is based on some strong assumptions made by domain experts on the dependency/relationship of attributes. As such, they are not unique and many SCMs are possible. The advantage of using these models is to provide an intuitive explanation behind a certain machine learning based model. Providing intuitive explanation is, in turn, a key requirement of explainability for data driven systems (Selbst and Barocas, 2018).

Kilbertus et al. (2017) uses casual reasoning to frame the problem of fairness based on protected attributes. This viewpoint aims at shifting the attention from defining the right discrimination criteria (based on some observed attributes) to "[w]hat do we want to assume about our model of the causal data generating process?" They introduce some natural causal non-discrimination criteria and develop algorithms that satisfy them. One of these criteria is based on the concept of *resolving variables.* A resolving variable is "any variable in the causal graph that is influenced by protected variable A in a manner that we accept as non-discriminatory". A key characteristic of a resolving variable is that "all paths from the protected attribute A to Y are problematic, unless they are justified by a resolving variable" (Kilbertus et al., 2017). For example, in the SCM of Figure 4 there is no resolving variable on the path from A to Y, thus reasoning based on the SCM in Figure 4 (i.e., based on variables X, A and Y therein) should be done carefully as there is a chance of problematic reasoning (i.e., potentially being unfair).

Extending the causal reasoning approach

We propose to extend the causal model of the machine learning system (i.e., the decision support system) with the corresponding impacts – i.e., g($\hat{Y}$, Context$_g$), h($\hat{Y}$, Context$_h$), … – to have a causal model for the whole extended decision support system. An illustration of such an extension of the output attribute in Figure 4 is shown with the dashed line components.

Let's revisit the hypothetical example mentioned above with

- Case g($\hat{Y}$, Context$_g$) being the expected jail capacity in a country and
- Case h($\hat{Y}$, Context$_h$) being the forecasted police patrol in a neighborhood,
- Given that outcome $\hat{Y}$ is a forecast of the volume of crime per neighborhood in near future.

In Figure 5, we depict a high-level SCM of this example, extended with the proposed impacts which eventually (do not) affect the individuals/groups behind the protected attributes A. We consider these impacts g($\hat{Y}$, Context$_g$) and h($\hat{Y}$, Context$_h$) as some nodes on the path from protected attributes A to the end attributes A. As such here we model the fact that (how and whether) protected attributes A (are used to) impact those individuals characterized by those protected attributes.

Based on causal reasoning, we can argue and reason about whether the impact attributes g($\hat{Y}$, Context$_g$) and h($\hat{Y}$, Context$_h$) influence protected groups and individuals adversely (unacceptably) or not (i.e., acceptably). In our hypothetical example, the impact g($\hat{Y}$, Context$_g$) might be acceptable – so being seen as non-discriminatory – because, or let's assume that, the forecasted jail capacity in a country does not have adverse impact on protected groups. Therefore, it becomes acceptable to use the protected variables (and privacy-sensitive variables) to forecast $\hat{Y}$ (the value of Y) and use the result for forecasting prison capacity in the future. This is because g($\hat{Y}$, Context$_g$) can be seen as a resolving variable. On the contrary, foreseeably the impact h($\hat{Y}$, Context$_h$) can have adverse impacts on protected groups – recall the so-

called redlining attacks – and as such it is not resolving. Therefore, the latter case is a potentially problematic case and should be approached cautiously (maybe not to build any prediction model for it at all).

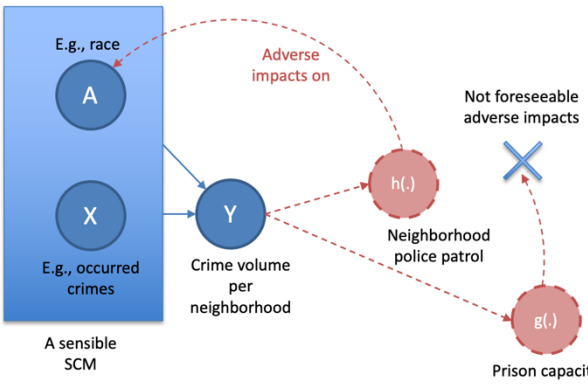

Figure 5: An extended SCM for the hypothetical example.

Using the causal reasoning rationale one can bring strong arguments forward for using all relevant attributes (including privacy sensitive and fairness protected attributes) to build a high-quality model and have a good forecast of crime volume in near future. This is due to knowing or arguing that the impact (e.g., the estimated prison capacity) does not have any privacy violating and unfair impacts on individuals and groups.

6. Discussion

As mentioned previously, moving from the case of data adaption in the pre-processing stage to the case of no adaption in the operation stage results in using the original data without modification (i.e., without protection against privacy and fairness risks). Often this delivers a better data utility and prediction quality. Nevertheless, one should keep in mind that the data processing outcome $\hat{Y}$ should only be used for the action in mind (i.e., $g(\hat{Y}, \text{Context}_g)$ being used only for the forecasted prison capacity). Therefore, there should be some (security) mechanisms in place to prevent illegitimate use of the outcomes like, monitoring, access control, and usage control (Bargh et al., 2017; 2016; 2014).

In practice, adapting data at both stages of pre-processing and operation is imaginable. This is the case, for example, when minor pre-processing is done to eliminate unnecessary (detailed) information and as such mitigate (to a limited level) the privacy and unfairness risks. Nevertheless, removing all such risks is unnecessary if the adverse impacts on individuals are justifiable/acceptable.

A key contextual parameter that justifies adoption of the first or the second case is the data environment within which the data is used. Adaption at the pre-processing stage is preferred when the data traverses an organization's boundaries and is shared with others or with the public. When the data is shared with the public, there is no mechanism to control the way that the data processing outcome is used. Therefore, the raw data should be adapted such that all risks are contained to an acceptable level. This would probably inflict harsh expenses on data utility. In case that the data is used

within one's own organization and for a specific purpose, adopting the second case (described in Section 5) can be considered.

The second case relates to the fair information use principle (being prominent in the Anglo-Saxon privacy law). This might be against the principle of informational self-determination (as realized in, e.g., the right to be forgotten and consent). On the other hand, the case seems to be in harmony with the spirit and principles of GDPR, which allows using (the special categories of) personal data for statistical and scientific purposes as well as for public interests. It is for future studies to research the legal grounds and implications of the second case.

7. Conclusions

In this contribution, we suggested looking at the whole process of data content adaption when protecting data utility, privacy and fairness in data-driven decision support systems. By means of two use cases, we argued that the quality and usage of data degrades substantially if we adapt the content of input data locally (especially when the data usage is not known at that time). Postponing data content adaption to a later stage, however, can lead to using high quality data without adapting data content for privacy and fairness protection in some practical situations. For identifying these situations, we used causal reasoning to argue whether using (privacy and fairness) protected attributes can be accepted as non-discriminatory. This approach leads to using high quality data for the prediction process in some cases; and may lead to improved predictions. Of course, we argued that there should be some mechanisms in place to guarantee the use of such prediction outcomes for the corresponding actions solely (and not using them for other purposes that may result in undesired/unjustifiable impacts on individuals and groups). We identified some directions for future research, which include specifying the actions that can be taken in pre-, in- and post-operation stages, investigating the ways to define fairness and privacy in a given context, based on societal, social, ethical, legal, personal norms, devising a method for mapping desired fairness concepts to formal measures concepts, and establishing a procedure to accommodate the residual privacy and fairness risks (i.e., those that cannot be addressed automatically).

8. References

Annas, G. J. (2003). HIPAA regulations: a new era of medical-record privacy? *New England Journal of Medicine*, 348, 1486.

Balayn, A., Lofi, C., & Houben, G. J. (2021). Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30(5), 739-768.

Bargh, M. S., Latenko, A., van den Braak, S., Vink, M., & Meijer, R. (2021). Personal data protection in the justice domain: Guidelines for statistical disclosure control. *Technical Report, Reeks Cahier 2021-10: PU-Tools 2.0 Project (nr. 3080a)*, Research and Documentation Center (WODC), The Hague, The Netherlands.

Bargh, M. S., Meijer, R., & Vink, M. (2018). On statistical disclosure control technologies: For enabling personal data protection in open data settings. *Technical Report, reeks Cahier 2018-20: PU-Tools Project (nr. 2889),* Research and Documentation Center (WODC), The Hague, The Netherlands.

Bargh, M. S., & Vink, M. (2017). On Usage Control in Relational Database Management Systems-Obligations and Their Enforcement in Joining Datasets. *In International Conference on Information Systems Security and Privacy*, vol. 2, February, 190-201. SCITEPRESS.

Bargh, M. S., Choenni, S., & Meijer, R. (2016). On design and deployment of two privacy-preserving procedures for judicial-data dissemination. *Government Information Quarterly*, 33(3), 481-493.

Bargh, M. S., Meijer, R., Choenni, S., & Conradie, P. (2014). Privacy protection in data sharing: towards feedback based solutions. *In Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance (ICEGOV)*, October 27-30, Guimaraes Portugal, 28-36.

Bargh, M. S., & Choenni, S. (2013). On preserving privacy whilst integrating data in connected information systems. *In Proceedings of the International Conference on Cloud Security Management (ICCSM'13)*, Seattle, US.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1), 3-44.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (Information Science and Statistics), New York, Springer.

Choenni, S., Netten, N., Bargh, M. S., & van den Braak, S. (2020). Exploiting big data for smart government: facing the challenges. *Handbook of Smart Cities*, 1-23.

Choenni, S., Netten, N., Bargh, M. S., & Choenni, R. (2018). On the usability of big (social) data. *In Proceedings of IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 1167-1174, IEEE.

Choraś, M., Pawlicki, M., Puchalski, D., & Kozik, R. (2020). Machine learning – the results are not the only thing that matters! what about security, explainability and fairness. *In International Conference on Computational Science*, 615-628, Springer, Cham.

Clifton, C., & Tassa, T. (2013). On syntactic anonymity and differential privacy. *In Proceedings of IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 88-93, IEEE.

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing. *ACM Computing Surveys*, 42(4), 1-53.

GDPR (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Hajian, S., Domingo-Ferrer, J., Farràs, O. (2014). Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *In Data Mining and Knowledge Discovery*, 28(5–6), 1158–1188.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., & Ullman, J. (2019). Differentially private fair learning. *In Proceedings of International Conference on Machine Learning*, 3000–3008.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *In proceedings of the 30th Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA.

Kusner, M., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. *In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA.

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. *In Proceedings of IEEE 23rd International Conference on Data Engineering,* 106-115, IEEE.

Li, N., Qardaji, W. H., & Su, D. (2011). Provably private data anonymization: Or, k-anonymity meets differential privacy. *CoRR, abs/1101.2604*, 49, 55.

Nissim, K., & Wood, A. (2018). Is privacy privacy? *In Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128), 20170358.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, 87, 1085.

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *arXiv preprint arXiv:2103*.12016.

Stoyanovich, J., Howe, B., Abiteboul, S., Miklau, G., Sahuguet, A., & Weikum, G. (2017). Fides: Towards a platform for responsible data science. *In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM'17*, 26:1–6. ACM, New York, NY, USA.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *In International Journal on Uncertainty*, 10(5), 557-570.

Sweeney, L. (2000). Uniqueness of simple demographics in the U.S. population: Laboratory for International Data Privacy, *Technical Report LIDAP-WP4*, Pittsburgh, PA, Carnegie Mellon University.

US EPA (1963). *United States Congress*, US Equal Pay Act.