Fair CRISP-DM: Embedding Fairness in Machine Learning (ML) Development Life Cycle

Vivek Kumar Singh
IST, College of Business Administration
University of Missouri, St. Louis
vsingh@umsl.edu

Anshuman Singh
IST, College of Business Administration
University of Missouri, St. Louis
a.singh@umsl.edu

Kailash Joshi
IST, College of Business Administration
University of Missouri, St. Louis
joshik@umsl.edu

Abstract

With the rapid adoption of machine learning (ML) technologies, organizations are constantly exploring efficient processes to develop such technologies. The cross-industry standard process for data mining (CRISP-DM) provides industry and technology-independent model for organizing ML project development. However, the model lacks fairness concerns related to ML technologies. To address this significant theoretical and practical gap in the literature, we propose a new model – Fair CRISP-DM, which groups and presents fairness concerns relevant to each phase of an ML project development. We contribute to the literature on ML development and fairness. Specifically, ML researchers and practitioners can use our model to check and mitigate fairness concerns in each phase of an ML project development.

Keywords: Machine Learning, Fairness, CRISP-DM

1. Introduction

Artificial Intelligence (AI) is growing exponentially and is expected to contribute $15.7 trillion to the global economy by 2030.[1] Machine Learning (ML), especially predictive analytics, is an integral part of AI. ML technologies discover patterns and learn from examples [1]. These technologies have helped achieve significant breakthroughs in many fields, such as image and speech recognition, health analytics, automobiles, e-commerce, and education [2][3]. Today, it is not easy to find an industry that has not been impacted by such technologies.

The cross-industry standard process for data mining (CRISP-DM) model is used as a comprehensive framework for machine learning (ML) project development across academia and industry [4][5]. Initially, the model was developed for managing data mining projects. However, with the growth of ML technologies, it is being applied in the development of ML projects. The model is independent of the underlying ML technology and context, and therefore is applied in multiple industries. It conceptualizes the ML development life cycle into six phases: (1) business understanding (2) data understanding (3) data preparation (4) modeling (5) evaluation and (6) deployment. We will discuss these phases in detail in the next section.

Algorithms and algorithmic decision making, which is the core of ML models, like any sophisticated technology, can benefit as well as harm individual and group interests. On the one hand, it can increase the productivity and profit of an organization, while on the other hand, it can reinforce societal stereotypes for different groups based on gender, race, minority status, etc [6][7]. Many fairness issues have been discovered post-deployment of ML systems that have led to financial losses for the implementing organizations as well as have negatively impacted their reputation and brand [8].

With their rapid adoption, fairness in ML projects is a growing area of concern and an emerging focus of research in Information Systems (IS) and its cognate disciplines such as computer science, statistics, and philosophy [9][10][11][12][13]. However, a focus on the aspects of algorithmic fairness issues and their mitigation is lacking in ML development models like CRISP-DM. When organizations encounter fairness issues in the implementation stage (or in a later stage of project development), it can be challenging to make amends and gain acceptance. Thus, it is essential to consider fairness issues right from the start of an ML project. In this paper, we present an improved ML development model – Fair CRISP-DM, which

[1] https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html

URI: https://hdl.handle.net/10125/79522
978-0-9981331-5-7

incorporates fairness in each phase of the ML project development cycle.

Recent research on fairness is mainly focused on model development. However, as mentioned above, model development or modeling is one of the six phases of the CRISP-DM model. We argue in this paper that fairness should be included at every phase of the model. Our proposed model – Fair CRISP-DM demonstrates why and how fairness should be included in each phase.

We conduct a comprehensive literature review of ML fairness research and map the fairness concerns and steps for their mitigation to the relevant phases in the CRISP-DM model. The list of fairness concerns in each phase will provide a checklist to academics and practitioners involved in machine learning projects. We acknowledge that not all fairness concerns may be relevant to each ML project, and some fairness issues may also span over multiple phases. Therefore, practitioners and researchers may choose relevant fairness concerns depending on the context.

The rest of the paper is organized as follows. In section 2, we present literature on fairness in ML and CRISP-DM model. Next, we present our proposed Fair CRISP-DM model in section 3, followed by a discussion in section 4, and conclude the paper in section 5.

2. Literature Review

In the literature review, we explore two streams of research: (1) fairness in machine learning and (2) machine learning project development life cycle and CRISP-DM. For a recent review on challenges in algorithmic fairness, see [14][9][15].

2.1 Fairness in Machine Learning

Fairness in machine learning or algorithmic fairness is a prominent cross-disciplinary topic connecting information systems [16], computer science [14][9][17], and social science [12][11].

In IS research, Kane et al. [16] present a new theoretical concept – "informania" which refers to an oppressive future resulting from extensive monitoring and control using ML systems. Among the adverse outcomes of ML systems described in the paper, fairness is one of them. The paper is futuristic, showing macro societal challenges arising from ML adoption; however, it does not provide guidelines to ML developers on how to develop fair ML systems. In another study, Kochling [18] examined algorithmic fairness in the recruitment context. The study finds uneven data representation over two dimensions – gender and ethnicity. One of the main limitations of studies like Kochling [18] is that the authors have considered only a few of the fairness concerns out of many such concerns documented in computer science research [14][9]. See Mehrabi et al. [9] for a comprehensive list of fairness concerns in ML projects.

In computer science, algorithmic fairness research started a decade ago [19]. However, we observe a sharp increase in the number of publications on this topic in the last 3-4 years. The majority of articles on this topic have documented a few of the many different types of fairness concerns; however, recent studies [10][9] have compiled a comprehensive list of fairness concerns and have also proposed new frameworks for mapping these concerns in ML development life cycle. However, this research is in its early stage. Our paper also contributes to this stream of research.

Mehrabi et al. [9] present 23 different fairness concerns and group them into three phases of ML project development – data, algorithm, and user interaction. Similarly, Suresh and Guttag [10] present six different biases. To the best of our knowledge, there is no study in the literature that presents a thorough list of fairness concerns and categorizes them based on a comprehensive ML or data mining process model such as CRISP-DM.

2.2 CRISM-DM Model

In this subsection, we describe the different phases of the CRISP-DM model. Although this model has been well studied in academia and industry [5][20][4], we provide a brief description of each phase of this model to motivate the connection between the phases and fairness concerns discussed in the next section.

One of the notable aspects of CRISP-DM is that it is independent of the industry in which it is being applied and the underlying technologies and algorithms used to solve different predictive analytics problems. Therefore, we believe that including fairness in CRISP-DM will be generalizable across industries and technologies.

As mentioned in the introduction section, there are six phases in CRISP-DM. In the business understanding (first phase), the project objectives and requirements are gathered, and this information is used

to define the machine learning problem and its objectives. Further, a plan is developed to achieve these objectives [21]. The key steps of business understanding include determining business objectives and data mining goals, situation assessment, and project planning [20].

The data understanding phase (second phase) focuses on data exploratory activities such as data collection for the model, data exploration and description, and data quality assessment. In the next phase, which is data preparation (third phase), a dataset is prepared for model training. This phase includes cleaning, missing data imputation, selection, merging, and/or formatting of data [20].

In the modeling phase (fourth phase), a model specification or algorithm is selected based on the nature of the ML problem. Generally, the models employed in ML can be grouped into three categories: regression models, forecasting models, and classification models.[2] Further, model training is conducted, and it is evaluated over validation data. Finally, model parameters are selected that provide the highest performance on the validation data.

In the evaluation phase (fifth phase), as the name suggests, the model is evaluated on the test or holdout data. Finally, the last phase (sixth phase), which is deployment, consists of deployment plan, monitoring, maintenance, reporting, and documentation.

In the next section, we will map different fairness concerns identified in the literature to phases of the CRISP-DM [20].

3. Fair CRISP-DM Model

We revise the CRISP-DM model to include fairness concerns presented in the extant research and propose a new model – Fair CRISP-DM. We group fairness concerns under different phases of the CRISP-DM model, as shown in Table 1. Further, we discuss each of the phases of the Fair CRISP-DM model along with the corresponding fairness concerns.

Table 1. Fair CRISP-DM Model

Fair CRISP-DM phases	Fairness concerns
Fairness in business understanding	▪ Defining (contextual) fairness ▪ Regulatory concerns regarding fairness ▪ Role of human in algorithmic decision making – fixing accountability & liability ▪ Cost and benefits of unfairness/fairness ▪ Funding bias ▪ Ethical and legal bias
Fairness in data understanding	▪ Historical bias ▪ Aggregation bias (clusters) ▪ Population bias ▪ Longitudinal data fallacy ▪ Behavioral bias ▪ Content production bias ▪ Linking bias ▪ Temporal bias ▪ Observer bias (data triangulation, multiple sources, inter-rater reliability) ▪ Simpson's paradox
Fairness in data preparation	▪ Representation bias ▪ Measurement bias ▪ Sampling bias ▪ Linking bias ▪ Self-selection bias
Fairness in modeling	▪ Temporal bias ▪ Algorithmic bias ▪ Omitted variable bias ▪ Cause-effect bias
Fairness in evaluation	▪ Evaluation bias
Fairness in deployment	▪ Popularity bias ▪ User interaction bias ▪ Social bias ▪ Emergent bias (concept drift when the underlying phenomenon changes)

3.1 Fairness in business understanding

The key fairness concerns in this phase are regulatory concerns regarding fairness [22][23][24][25][26], defining (contextual) fairness [14], role of human in algorithmic decision making – fixing, accountability & liability, cost [27], benefits of

[2] https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml

unfairness/fairness [28][3], funding bias [9], and ethical and legal bias [29].

Over time AI systems have advanced from a machine-oriented context (e.g., setting thermostats) [30] to a broader domain of human and social context applications (e.g., processing college admission applications), which pose unanticipated challenges and raises new questions. Some of these challenges are reflected in the emergence of the Society in the Loop (SIL) framework [31] to consider a more comprehensive set of issues in the fairness of algorithmic decisions. Specifically, the framework identifies the influences of the social milieu in which the AI systems operate and the ethical and fairness concerns that may arise in their adoption and usage for decision making. Many AI systems run into unforeseen difficulties and criticism when they fail to adequately identify, analyze, and address the fairness issues in the application domain. Often, it becomes hard to overcome the antagonism and loss of goodwill due to an unplanned, botched implementation attempt. Though a post hoc redesign and recoding of the system may attempt to address the issues, it becomes difficult to recover from the loss of time and reputational setbacks for the responsible parties.

It is critical to actively identify and address the likely ethical and fairness issues that can arise considering the application context. Although it is challenging to articulate a general definition of fairness [14], a contextual definition of fairness is required in the business understanding phase. In other words, the definition of fairness may change from one context to another. For example, the definition of fairness might differ in various contexts, such as human resource management, judiciary, and e-commerce.

Society in the loop (SIL) framework [31] provides a sound basis for ex-ante analysis and identification of fairness issues related to an AI application. It recognizes the inputs and interactions of multiple stakeholders in defining what is fair. Thus, it is crucial to identify the various stakeholders involved in the application domain and their fairness concerns in arriving at a consensus definition of fairness for the AI system being proposed that is acceptable to the responsible parties and can satisfy most stakeholders.

Information systems (IS) researchers have long recognized the importance of in-depth planning and feasibility analysis before jumping into coding and building a new system to minimize implementation failures [32]. Some of these lessons are also relevant for AI systems that face common systems development issues and unique challenges in the area of ethics and fairness due to their decision-making role. The cost and benefits of addressing fairness concerns should be evaluated in this phase of CRISP-DM.

Additionally, the context for the AI systems should be studied carefully to identify the role of humans in the final decision-making. Researchers have reported mixed results for purely AI algorithmic decisions. Human supervision and control [7] may help provide greater confidence in the fairness of decisions depending on the application context [7].

On contentious issues, it may require negotiation and adjustments to obtain a wider acceptance. Once the definition of fair outcomes is settled, additional planning may be needed to address the related operational issues. While consultation and negotiation may help overcome some of the procedural fairness concerns of the stakeholders, questions about informational fairness and distributive fairness may still linger and grow at the implementation stage. Therefore, AI implementation plan should also identify the information/statistics that should be released to the stakeholders on a periodic basis and consider the explainability issues in their design [33]. Following steps should be considered in this stage of fair CRISP-DM.

- Assessment of implementation context and likely impact of fairness concerns on AI systems implementation.
- Identification of direct stakeholders. Inclusion of other interested stakeholders in the planning and implementation process, including civil society organizations and public interest groups. The development plan should aim to establish common ethical standards and integrate them into AI projects [34][35].
- Analysis of stakeholder interests. Seeking inputs from stakeholder groups.
- Development of fairness definition for the context.
- Identification of algorithmic approaches to fairness assessment and correction.

- Consideration of competing objectives for the system and identifying an overall compromise between fairness, accuracy, transparency, accountability, explainability, privacy, and security [36].

- Development of a screening plan for variables inclusion/exclusion in line with the overall trade-offs.

- Educating stakeholders on the trade-offs inherent in meeting fairness goals and developing consensus for the compromise approach to meet fairness goals. It should be noted that such compromises or choices may be highly context-dependent based on the nature of decisions made.

- Plan for the extent and nature of human supervision and intervention in the final algorithmic decisions. Identification of training requirements for human participants, decision supervisors, and users of the system.

- Plan for addressing explainability issues in AI system design [33].

- Plan for post hoc fairness tests and audits [37].

- Design of the process for ensuring informational fairness and information disclosures to the stakeholders.

- Formal process for appeals and audits of the decisions for the stakeholders unsatisfied with the fairness of decisions [37].

- Addressing Ethical and legal fairness concerns [15] and funding bias concerns [9].

3.2 Fairness in data understanding

The first step in data understanding is the data collection process. The data used for machine learning may consist of a variety of formats such as transactional data, textual data, and multimedia data. Further, data might be available in archival storages, production databases, collected from surveys or experiments, or behavioral data from human interactions with the system. The unfairness in data collection may stem from data collection infrastructure, demographic disparities, and the type of data [38].

To study the impact of a demographic disparity, we need a cross-disciplinary approach including information systems, computer science, philosophy, and sociology [14]. There are societal decisions that impact individuals' opportunities and thus warrant critical examination of factors involved in automating and predicting such decisions.

The demographic disparities may be reflected in the training data [14][39]. For example, there are substantial gender differences across different occupations. Moreover, these demographic disparities may change over time. The data collection process should be aware of such disparities in different domains. Moreover, the demographic disparities may increase with the limitations of data collection infrastructure. A data collection process may systematically exclude certain demographic factors due to technical limitations. For example, data collected using smartphones on road conditions (e.g., potholes) exclude neighborhoods having low smartphone adoption [14].

The use of fair measurement is another concern in data collection as there may be subjectivity involved in measurement. Recent research by Jacobs and Wallach [40] states that the measurement techniques from social science research, including construct, its validity and reliability, are better at measuring fairness compared to direct measures used in computer science literature. This technique has been one of the core research methods in the Information Systems literature, and IS researchers are well-positioned for methodological contribution in developing fair measurements [41].

A target variable (also known as label in a classification problem) plays the most central role in ML, and a biased measurement of the target variable can directly bias the training model. Sometimes, demographic variables are used as a substitute for environmental factors. For example, race being used as a substitute measure of patients' environmental factors. Capturing the environmental factors directly might improve the predictive accuracy of the models [14].

The models which are deployed in real-time also generate training data. This training data is prone to feedback bias. The predictions of the model interact with the users' decisions, and users' choices are based on their intrinsic requirements and the outcome of the model. Thus, feedback bias may arise, which should be tested before deploying the model in the field.

Data visualization techniques can be used to detect demographic disparities. They can highlight disparities across demographic variables and help identify them for corrective action.

Simpson's paradox occurs when the relationships between dependent and independent variables differ at the population and group levels. To mitigate such bias, the data should be prepared for each group separately for modeling. Moreover, special models like multi-task learning can be used to address this bias as well.

3.3 Fairness in data preparation

The key fairness concerns related to this phase of CRISP-DM are (1) representation bias (2) measurement bias (3) sampling bias, (4) linking bias, and (5) self-selection bias.

The representation bias occurs when we do not take a representative sample of data from a population. The approach to mitigate this bias involves including data from underrepresented groups in the population.

The measurement bias occurs from two sources – (1) when an available proxy variable is used to measure a concept or construct; however, the variable does not completely or accurately capture the construct. Moreover, the second source of measurement bias stems from an erroneous measurement of the proxy variable. One of the ways to mitigate this bias is using measurement methods from social science involving construct validity and reliability [40].

Sampling bias occurs when the sample is not random, especially for the subgroups. The model trained from such a sample will be difficult to generalize. Taking a random sample will mitigate this bias.

In social networks, low-degree nodes may have different behavior compared to their links, and inferring about such nodes from network links leads to linking bias. This bias may be mitigated using an unbiased network sample.

Self-selection bias occurs when the participants or users self-select themselves in an experiment. This bias has been extensively studied in IS research. A random selection strategy in which participants are selected randomly into control and treatment group may mitigate this bias. Also, in archival data, propensity score matching and similar techniques can be applied to mitigate this bias [42].

3.4 Fairness in modeling

The model training phase is more effective in considering and selecting a fairness/accuracy trade-off point than post-training methods since the analyst has access to the training data in this phase. Also, fairness-aware models generated during model training can still be further improved using post-processing fairness mitigation methods [29]. The shift of incorporating fairness during the model training phase should increase confidence among model users based on procedural improvements in the development of the ML system. We can consider this as a step towards machine learning model assurance similar to software assurance during the software development lifecycle [5].

Model training typically involves feeding a learning algorithm with training data to generate a trained model with fitted parameters. Learning algorithms use an optimization procedure to minimize the error on training data with the error function depending on the type of model used [6]. Since learning algorithms focus on the minimization of error, the goal of error minimization may not align with bias reduction. The process of error minimization may result in model parameters that may lead to an increase in bias depending on the definition of bias/fairness used.

Fairness mitigation during modeling has been proposed for both classification and regression. Zafar et al. [7] proposed a quantitative measure of bias called decision boundary (un)fairness. They bound this measure using a covariance threshold and applied it as a constraint on the error minimization function. Hence, fair learning, in this case, involves a constrained optimization problem that turns out to be convex optimization and hence computationally tractable. The approach in [7] is applicable for logistic regression and SVM classifiers. The decision boundary fairness approach avoids both disparate treatment and disparate impact and allows one to formulate the learning problem as a fairness maximization problem subject to accuracy constraint instead of the traditional error minimization under fairness constraints.

The trade-off between accuracy and fairness can be expressed using Pareto optimality [8]. Under the Pareto optimality framework, fair learning requires solving a multi-objective optimization problem. We can choose the parameters of the model anywhere on the line (in general, hyperplane), joining the classifier

without bias correction and the classifier with maximal bias correction. This line is also called the Pareto front.

Fairness can also be incorporated in the modeling phase of regression. Agarwal et al. [9] proposed using regularization to penalize bias. The fairness penalty can be added to the loss function for both group fairness (statistical parity) and individual fairness (similarity-based) definitions. Again, the fairness penalties are convex resulting in efficient learners.

3.5 Fairness in evaluation

Evaluation bias occurs when inappropriate evaluation strategies and criteria are used for evaluating a model. To evaluate the fairness of ML algorithms, leading AI organizations and researchers are developing AI fairness tools such as IBM's AI Fairness 360 and Microsoft's Fairlearn projects [43][44].

3.6 Fairness in deployment

There are multiple fairness concerns such as popularity bias, user interaction bias, social bias, and emergent bias (e.g., concept drift wherein the underlying phenomenon changes) related to model deployment.

Popularity bias occurs when an item is recommended based on its popularity by a model. However, the popularity may be manipulated. For example, in e-commerce, reviews are used to measure the popularity of goods and services. Such measures of popularity might be manipulated using fake reviews. To mitigate this bias, other attributes apart from popularity should be considered for recommendation [9].

There are two types of user interaction bias – presentation bias and ranking bias [9]. The presentation bias occurs towards the content which is not presented by the model to the user. On the other hand, ranking bias happens when one item is ranked higher compared to other by a model. The mitigation approach must consider both these user interaction biases while using the data recorded from the model in its retraining.

Social bias occurs when an individual's action changes in the presence of others. For example, users may not interact with a model freely in the presence of others using the same model or platform simultaneously.

Finally, the emergent bias occurs when the underlying data population changes. To mitigate this bias, the model should be retrained based on changes in the population.

4. Discussion

Fairness has emerged as a critical factor in the deployment and acceptance of ML projects. However, it cannot be incorporated at the end of a development cycle as an afterthought. The Fair CRISP-DM model presented in this paper aims to highlight the need to focus on fairness issues right from the start of an ML project and incorporate consideration of the relevant biases in each phase of the development process. Our model contributes to two streams of research – fairness research and machine learning/data mining project development.

Fairness research is an interdisciplinary field. The interdisciplinary work provides different perspectives to define, understand, measure, and mitigate fairness. At the same time, it is challenging for IS and computer science research to comprehend these perspectives and apply them in their ML model. Only a few recent papers have connected the interdisciplinary research and presented actionable items for ML developers. Our paper contributes to this research stream by adapting a well-known data mining process model to embed fairness in each stage of the model.

Apart from bias and fairness, there are other challenges that inhibit trust in AI algorithms: (1) explainability, (2) privacy, (3) security [45]. Similar to fairness concerns, these challenges/concerns are missing in the CRISP-DM model. We plan to include these concerns in the CRISP-DM model as part of our future research.

The list of fairness concerns or biases is a contemporary area of research. We also plan to include new fairness concerns in our future research.

5. Conclusion

In this paper, we introduce a Fair CRISP-DM model for the development of ML projects. The model incorporates consideration of fairness issues in the development process, starting from the planning phase to the deployment phase for a successful implementation. We also attempt to map the relevant types of biases that should be in focus in each phase

and thereafter. Thus, the model also provides a good framework of analysis for practitioners and researchers in considering fairness issues related to a project. We plan to refine the model in future research. The model should be useful for practitioners to better plan and execute ML projects. We also list the tools and processes available to better identify and mitigate fairness concerns related to ML development, implementation, and usage.

6. References

[1] M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," *Calif. Manage. Rev.*, vol. 61, no. 4, pp. 5–14, 2019, doi: 10.1177/0008125619864925.

[2] H. Benbya and T. H. Davenport, "Special Issue Editorial : Artificial Intelligence in Organizations : Current State and Future Opportunities," *MIS Q. Exec.*, vol. 19, no. 4, 2020.

[3] J. Wawira Gichoya, L. G. McCoy, L. A. Celi, and M. Ghassemi, "Equity in essence: A call for operationalising fairness in machine learning for healthcare," *BMJ Heal. Care Informatics*, vol. 28, no. 1, 2021, doi: 10.1136/bmjhci-2020-100289.

[4] C. Shearer *et al.*, "The CRISP-DM model: The New Blueprint for Data Mining," *J. Data Warehous.*, 2000.

[5] J. Jackson, "Data Mining; A Conceptual Overview," *Commun. Assoc. Inf. Syst.*, vol. 8, 2002, doi: 10.17705/1cais.00819.

[6] G. Satell and Y. Abdel-Magied, "AI Fairness Isn't Just an Ethical Issue," *Harv. Bus. Rev.*, 2020.

[7] C. Starke, J. Baleis, B. Keller, and F. Marcinkowski, "Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature," 2021, [Online]. Available: http://arxiv.org/abs/2103.12016.

[8] R. Burkhardt, N. Hohn, and C. Wigley, "Leading your organization to responsible AI," *McKinsey Anal.*, no. May, 2019.

[9] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," 2019, [Online]. Available: http://arxiv.org/abs/1908.09635.

[10] H. Suresh and J. V. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning," 2019, [Online]. Available: http://arxiv.org/abs/1901.10002.

[11] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annu. Rev. Stat. Its Appl.*, vol. 8, pp. 141–163, 2021, doi: 10.1146/annurev-statistics-042720-125902.

[12] S. Fazelpour and D. Danks, "Algorithmic bias: Senses, sources, solutions," *Philos. Compass*, vol. 16, no. 8, 2021, doi: 10.1111/phc3.12760.

[13] S. Passi and S. Barocas, "Problem formulation and fairness," *FAT* 2019 - Proc. 2019 Conf. Fairness, Accountability, Transpar.*, pp. 39–48, 2019, doi: 10.1145/3287560.3287567.

[14] S. Barocas, M. Hardt, and A. Narayanan, "Fairness in machine learning," *Nips Tutor.*, vol. 1, p. 2, 2017.

[15] E. Ntoutsi *et al.*, "Bias in data-driven artificial intelligence systems—An introductory survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, 2020, doi: 10.1002/widm.1356.

[16] G. C. Kane, A. G. Young, A. Majchrzak, and S. Ransbotham, "Avoiding an Oppressive Future of Machine Learning: A Design Theory for Emancipatory Assistants," *MIS Q.*, vol. 45, no. 1, pp. 371–396, 2021, doi: 10.25300/misq/2021/1578.

[17] L. Zhang, Y. Wu, and X. Wu, "A causal framework for discovering and removing direct and indirect discrimination," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 0, pp. 3929–3935, 2017, doi: 10.24963/ijcai.2017/549.

[18] A. Köchling, S. Riazy, M. C. Wehner, and K. Simbeck, "Highly Accurate, But Still Discriminatory: A Fairness Evaluation of Algorithmic Video Analysis in the Recruitment Context," *Bus. Inf. Syst. Eng.*, vol. 63, no. 1, pp. 39–54, 2021, doi: 10.1007/s12599-020-00673-w.

[19] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," *ITCS 2012 - Innov. Theor. Comput. Sci.*

Conf., pp. 214–226, 2012, doi: 10.1145/2090236.2090255.

[20] P. Chapman, "The CRISP-DM User Guide," *Cris. User Guid.*, p. 14, 1999.

[21] R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000.

[22] P. Almeida, C. Santos, and J. S. Farias, "Artificial Intelligence Regulation: A Meta-Framework for Formulation and Governance," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020, doi: 10.24251/hicss.2020.647.

[23] M. C. Buiten, "Towards intelligent regulation of artificial intelligence," *Eur. J. Risk Regul.*, vol. 10, no. 1, pp. 41–59, 2019, doi: 10.1017/err.2019.8.

[24] S. Chatterjee, "Personal Data Sharing and Legal Issues of Human Rights in the Era of Artificial Intelligence: Moderating Effect of Government Regulation," *Int. J. Electron. Gov. Res.*, vol. 15, no. 3, pp. 21–36, 2019, doi: 10.4018/IJEGR.2019070102.

[25] M. Fenwick, E. P. M. Vermeulen, and M. Corrales, "Business and regulatory responses to artificial intelligence: Dynamic regulation, innovation ecosystems and the strategic management of disruptive technology," *Perspect. Law, Bus. Innov.*, pp. 81–103, 2018, doi: 10.1007/978-981-13-2874-9_4.

[26] E. Kurshan, H. Shen, and J. Chen, "Towards Self-Regulating AI: Challenges and Opportunities of AI Model Governance in Financial Services," 2020, doi: 10.1145/3383455.3422564.

[27] B. Deng, "Machine ethics: The robot's dilemma," *Nature*, vol. 523, no. 7558, pp. 24–26, 2015, doi: 10.1038/523024a.

[28] C. Haas, "The price of fairness - A framework to explore trade-offs in algorithmic fairness," *40th Int. Conf. Inf. Syst. ICIS 2019*, 2020.

[29] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," 2020, [Online]. Available: http://arxiv.org/abs/2010.04053.

[30] S. G. Tzafestas and H. B. Verbruggen, *Artificial intelligence in industrial decision making, control and automation*, vol. 14. Springer Science \& Business Media, 2012.

[31] I. Rahwan, "Society-in-the-loop: programming the algorithmic social contract," *Ethics Inf. Technol.*, vol. 20, no. 1, pp. 5–14, 2018, doi: 10.1007/s10676-017-9430-8.

[32] A. Dennis, B. Wixom, and D. Tegarden, *Systems analysis and design: An object-oriented approach with UML*. John wiley \& sons, 2015.

[33] H. Liu *et al.*, "Trustworthy AI: A Computational Perspective," 2021, [Online]. Available: http://arxiv.org/abs/2107.06641.

[34] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, Transparent, and Accountable Algorithmic Decision-making Processes," *Philos. Technol.*, vol. 31, no. 4, pp. 611–627, 2018, doi: 10.1007/s13347-017-0279-x.

[35] I. Zliobaite, "Measuring discrimination in algorithmic decision making," *Data Min. Knowl. Discov.*, vol. 31, no. 4, pp. 1060–1089, 2017.

[36] N. Köbis, C. Starke, and I. Rahwan, "Artificial Intelligence as an Anti-Corruption Tool (AI-ACT) -- Potentials and Pitfalls for Top-down and Bottom-up Approaches," 2021, [Online]. Available: http://arxiv.org/abs/2102.11567.

[37] S. Brown, J. Davidovic, and A. Hasan, "The algorithm audit: Scoring the algorithms that score us," *Big Data Soc.*, vol. 8, no. 1, 2021, doi: 10.1177/2053951720983865.

[38] M. Andrus, E. Spitzer, J. Brown, and A. Xiang, "What we can't measure, We can't understand: Challenges to demographic data procurement in the pursuit of fairness," in *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 249–260, doi: 10.1145/3442188.3445888.

[39] S. Rančić, S. Radovanović, and B. Delibašić, "Investigating Oversampling Techniques for Fair Machine Learning Models," in *International Conference on Decision Support System Technology*, 2021.

[40] A. Z. Jacobs, "Measurement and fairness,"

FAccT 2021 - Proc. 2021 ACM Conf. Fairness, Accountability, Transpar., pp. 375–385, 2021, doi: 10.1145/3442188.3445901.

[41] D. Gefen, E. E. Rigdon, and D. Straub, "An update and extension to SEM guidelines for administrative and social science research," *MIS Q.*, vol. 35, no. 2, 2011, doi: 10.2307/23044042.

[42] M. DeFond, D. H. Erkens, and J. Zhang, "Do client characteristics really drive the big N audit quality effect? New evidence from propensity score matching," *Manage. Sci.*, vol. 63, no. 11, pp. 3628–3649, 2017, doi: 10.1287/mnsc.2016.2528.

[43] R. K. E. Bellamy *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM J. Res. Dev.*, vol. 63, no. 4–5, 2019, doi: 10.1147/JRD.2019.2942287.

[44] S. Bird *et al.*, "Fairlearn: A toolkit for assessing and improving fairness in AI," *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020, [Online]. Available: https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf.

[45] S. Alfeld, X. Zhu, and P. Barford, "Data poisoning attacks against autoregressive models," in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016, pp. 1452–1458.