Heinrich Heine Universität Düsseldorf

# How is Socially Responsible Academic Performance Prediction Possible? Insights from a Concept of Perceived AI Fairness.

Birte Keller, Marco Lünich, Frank Marcinkowski

Chapter - Postprint

# How is Socially Responsible Academic Performance Prediction Possible?

## Insights from a Concept of Perceived AI Fairness.

Birte Keller[1], Marco Lünich[1], Frank Marcinkowski[1]
[1]*Heinrich Heine University, Germany*

This is a manuscript of a chapter from the book "Strategy, Policy, Practice, and Governance for AI in Higher Education Institutions" (Almaraz-López et al., 2022) published by IGI Global. Please reference the final publication. The final proofed version might differ from the current document.



Corresponding author:

Birte Keller
Department for Social Sciences
Heinrich Heine University Düsseldorf
Universitätsstr. 1, 40225 Düsseldorf
Email: birte.keller@hhu.de

# How is Socially Responsible Academic Performance Prediction Possible?

## Insights from a Concept of Perceived AI Fairness.


### ABSTRACT

*The availability of big data at universities enables the use of artificial intelligence (AI) systems in almost all areas of the institution: from administration to research, to learning and teaching, the use of AI systems is seen as having great potential. One promising area is academic performance prediction (APP), which is expected to provide individual feedback for students, improve their academic performance, and ultimately increase graduation rates. However, using an APP system also entails certain risks of discrimination against individual groups of students. Thus, the fairness perceptions of affected students come into focus. To take a closer look at these perceptions, this chapter develops a framework of the "perceived fairness" of an ideal-typical APP system, which asks critical questions about input, throughput, and output and, based on the four-dimensional concept of organizational justice (Greenberg, 1993), sheds light on potential (un-) fairness perceptions from the students' point of view.*


Keywords: Organizational Justice, Fairness, Academic Performance Prediction, Dropout, Study Success, Students, Perceptions, Accountability, Input Data, Explainability, Interventions, Algorithmic Design

### INTRODUCTION

Students' academic learning and work are increasingly taking place online or on digital learning platforms. With the digitization of examinations and administrative processes, higher education institutions generate large amounts of data. The so-called big data enables artificial intelligence (AI) technologies, such as machine learning (ML), to predict academic performance in higher education (Alyahyan & Düştegör, 2020; Daniel, 2015). Some of the attested potentials are already being realized via automated admission systems, such as Parcoursup in France (Frouillou et al., 2020), automated grading (Kotsiantis, 2012), support for administrative or research tasks, and learning analytics (Daniel, 2015; Ekowo & Palmer, 2016). The latter includes a wide variety of applications that enable, for example, real-time performance feedback and advice or performance prediction of exam and study performance. Higher study success is expected to materialize through dropout and performance prediction systems, ultimately preventing student dropouts (Arnold & Pistilli, 2012; Attaran et al., 2018).

However, the use of big data and AI systems in higher education always carries several risks and may lead to potential damages of material (i.e., misallocation of resources) and social nature. For instance, one primary concern is that the data will inherently contain biases, that the algorithms themselves will thus perpetuate or even produce stereotypes, and therefore have discriminatory effects on particular students (Attaran et al., 2018; Ekowo & Palmer, 2017; Fazelpour & Danks, 2021). To take an example from higher education, prospective students could conceivably be disadvantaged in an automated admission process because they belong to a population that is statistically less likely to graduate (Muñoz et al., 2016). In this sense, the fairness aspects of algorithmic decision-making (ADM) are increasingly receiving attention in interdisciplinary research activities (e.g., Lee, 2018; Shin & Park, 2019; Starke et al., 2021). Many scholars have focused on the distribution of goods and their translation into different mathematical fairness notions (e.g., Verma & Rubin, 2018). Recently, however, the focus has shifted to individual perceptions of fairness, which are becoming more important in examining the public understanding and acceptance of AI systems and the legitimacy of AI-driven decision-making (Simmons, 2018; Wong, 2020). Therefore, we bring to the fore students' fairness perceptions of algorithmic decisions in higher education. Consequently, using an academic performance prediction (APP) system as an example, we analyze in this chapter the fairness challenges involved in the use of such systems in higher education that a) need to be considered in implementing APP within academic institutions and b) need to be investigated by conducting empirical research before and during the said implementation process. To this end, we refer to the four-dimensional concept of organizational justice (Greenberg, 1993), which is concerned with designing intra-organizational decision-making processes to achieve or maintain the highest possible satisfaction and commitment of organizational members.

Accordingly, the question arises of how universities can use data-driven technologies in an attempt to make internal decision-making processes more streamlined and effective while maintaining the commitment of their members (i.e., administrative staff, lecturers, and especially students). This chapter aims to tackle this puzzle by suggesting a conceptual framework of *perceived fairness* integrated within the implementation process of APP in higher education. In constructing our framework, we highlight intricate and controversial choices concerning the input, throughput, and output of systems of APP that relate particularly to the decisions underlying data, algorithms, outcomes, and communication of the individual choices within the process. Subsequently, we illustrate the potential pitfalls of the imprudent implementation of performance prediction, raise important research questions to be addressed, and offer recommendations for its socially acceptable usage. For this, we sketch an ideal-typical performance prediction system and argue from the perspective of distinct organizational stakeholders to identify the features (e.g., data, procedures, and communication) of a socially responsible APP system and how the characteristics of these elements should be optimally designed to optimize perceived AI fairness. The stakeholders we focus on are primarily students as they are directly affected by the APP system. Secondarily, we refer to the university administration, which has to decide on the actual implementation and application of the system. Lastly, we also reflect on the role of lecturers, although they are indirectly affected by the system.

With the four fairness dimensions in mind, this chapter contributes to the essential question of the social acceptability of using AI systems for performance prediction in higher education institutions. Thus, existing considerations on the perceived fairness of AI systems are extended to the scope of higher education. However, concerning oneself with fairness helps derive further insights for developing and implementing AI systems at universities, as such systems are expected to be used globally. Eventually, our analysis and the theoretical framework of the practice of fair academic performance prediction provide a foundation to inform useful future policies and research for policy-makers, administrators, and researchers in higher education.

## ACADEMIC PERFORMANCE PREDICTION

The use of digital learning and teaching processes on the one hand and the digitalization of performance data, on the other hand, are causing a rapid increase in the amount of data at universities (Daniel, 2017; Dede et al., 2016; Liebowitz, 2017). The potential applications thus extend to almost all university areas, as long as these data are available. Moreover, processes in administration, management, research, or learning analytics are automated or supported by algorithmic recommendations, predictions, or decision-making (Daniel, 2015; Keller et al., 2019). For example, using student data, predictive analytics predict performance and grades for individual exams, entire modules, or even entire courses of study (Alyahyan & Düştegör, 2020; Askinadze & Conrad, 2019). Instead of performance prediction, universities can also use dropout prediction. Here, the focus is not directly on the expected performance, but the aim is to predict the probability that a student will not complete their studies (Askinadze et al., 2018; Aulck et al., 2016). The performance or dropout prediction systems are based on historical data of previous academic performance and the actual dropouts of former students, which can be supplemented by socio-demographic data (Olaya et al., 2020). These data are then used to train ML algorithms to make more or less precise predictions about current students' success. Consequently, warnings or intervention recommendations based on this algorithmically derived information should improve student success and thus increase graduation rates (Attaran et al., 2018; Daniel, 2015; Yanosky & Arroway, 2015). Predictions make it possible to provide students with individualized feedback and targeted counseling services. Further, limited resources (financial, material, and human resources, e.g., in the form of support workshops) can be used more efficiently, especially for students who need them the most (Ekowo & Palmer, 2016; Muñoz et al., 2016; Yanosky & Arroway, 2015). The use of predictive analytics, in particular, should offer the opportunity to strengthen student retention in the university as well as drive a more robust social commitment of student behavior toward the university (Sclater et al., 2016). Corresponding systems are already in use worldwide: from the US at Georgia State University (Ekowo & Palmer, 2016) and Purdue University (Arnold & Pistilli, 2012) to Germany (Berens et al., 2019; Kemper et al., 2020), Bangladesh (Ahmed & Khan, 2019), and Australia (Adams Becker et al., 2017; Sclater & Mullan, 2017).

However, the use of APP systems is not viewed exclusively positively. A typical problem with big data analyses is collecting, selecting, and processing personal data and ensuring privacy and data protection issues. Often, the persons affected are unaware of how and where their personal data are stored and for what purpose (Daniel, 2015; Hamoud et al., 2018). Another problem that can arise from the data basis is the reproduction of existing stereotypes. On the one hand, the historical data used may already have a bias, affecting the algorithmic performance prediction decision (Fazelpour & Danks, 2021; Muñoz et al., 2016). On the other hand, the algorithmic procedures can also discriminate themselves, for example, by using protected, respectively sensitive, attributes for the prediction (such as gender or ethnicity), which should not influence the algorithmic decision (Calders & Žliobaitė, 2013; Dwork et al., 2012).

## APP-FAIRNESS IN THE EYE OF THE BEHOLDER: A PROPOSAL

The risk of discrimination and the systematic disadvantage of particular groups of people speaks to the potential social damages of ML and is one of several critical issues from which perceptions of unfairness can arise. For instance, discriminatory systems have received public attention, such as the COMPAS system used in US courts to predict the likelihood of recidivism of offenders that systematically assigned black offenders a higher likelihood (Angwin et al., 2016). Such incidents have fueled public and academic debates about the fairness of AI, machine learning, and ADM. The scientific literature, especially the computer science literature, is mainly concerned with fair distributions and how these can be guaranteed through algorithmic decisions that translate different distributions into mathematical notions (Dwork et al., 2012; Verma & Rubin, 2018).

However, there is no universal definition of fairness. There are rather distinct ideas and concepts of fairness that differ from context to context (Wong, 2020), especially from individual to individual, depending on distinct normative conceptions and ethical considerations (Binns, 2018). Therefore, research on perceptions of fairness and their social consequences has gained attention in recent years (Grgić-Hlača et al., 2018; Lee et al., 2019; Saxena et al., 2020). To systematize and define fairness perceptions, these works mainly refer to the concept of organizational justice (Greenberg, 1993). The transfer of the fairness concept from human decision-making processes to AI-supported intra-organizational decisions makes it possible to look at the consequences for the institutions making the decisions, in our case, the institutions of higher education. Such an endeavor is critical in light of the hope mentioned above for improved student retention (Sclater et al., 2016). As Marcinkowski et al. (2020) suggest, (un)fairness perceptions concerning an AI system can have a negative impact on the support toward the higher education institutions. Such findings suggest that irrespective of conceptual considerations, formally fair systems (i.e., relying on conceptually fair formulas) may have unintended social consequences and may thus lead to fairness perceptions that deviate from formal fairness notions. Eventually, instead of the hoped-for higher academic success and higher graduation rates, in the worst case, this may lead to reputation losses of the higher education institution or student withdrawal, leading to fewer students and consequently to fewer financial resources. Such potential consequences are additional important reminders that speak to the possible damages of the imprudent implementation of APP within the academic context and thus require careful consideration and concomitant research (Data Ethics Commission, 2019; Zweig et al., 2018).

The framework proposed in this chapter for evaluating fairness perceptions of AI, elaborated for the use case of academic performance prediction in German higher education, is based on the concept of organizational justice (Greenberg, 1987, 1990, 1993) and the four dimensions of perceived fairness located therein: (1) distributive fairness, (2) procedural fairness, (3) interactional fairness, and (4) informational fairness.

## Distributive Fairness

The first dimension, distributive fairness, is concerned with the perception of the validity of outcomes (Cropanzano et al., 2001), that is, whether a person perceives a decision or consequence as correct. For instance, we assume that a decision is perceived as fair if it agrees with the self-assessment of the person affected and appears appropriate in this sense. In the context of APP, this refers to whether the AI's forecast of a student's performance is perceived as valid or whether a student feels systematically disadvantaged by the prediction and subsequent interventions. Thus, the perception of fair distribution can vary between those affected and depend on different normative assumptions concerning desirable distributions. For instance, according to *equity* theory (Adams, 1963, 1965), people may value the outcome of distribution in relation to their input provided (Tyler, 2000). Thus, if one considers different distributions of the interventions that can be distributed as a result of the APP (e.g., remedial offers or tutoring courses), it would consequently be conceivable that particularly high-performing students may (also) be expected to be entitled to additional support. In addition to this *equity* principle, Deutsch (1975) distinguished between two more distribution norms. When distributed according to the *equality* principle, every student would receive the same resources, regardless of the APP result. Under a *need*-based distribution, students who are particularly in need receive the most resources. In addition to the question of which distribution norm is perceived as fair by the majority of the student body, the question arises of how the respective distribution norm should subsequently be implemented. For instance, how is need defined, and are there different increments or even limits to students' needs?

Moreover, perceptions of unfairness may relate to the distribution of resources and the categorization of students by the APP system, which serves as a basis for the respective distribution. For example, AI methods can recognize patterns in data and then either divide the individual cases into categories, put them in a particular order (scoring system), or assign them a risk rating (Zweig & Krafft, 2018). An example of categorization is provided by the algorithm used by the Public Employment Service Austria (AMS), which classifies unemployed persons into three categories: (1) job seekers who are readily employable in the labor market without the need for additional support measures, (2) people who can be helped to find a job through support measures, and (3) people who are unlikely to find a job in the foreseeable future, even with further support. This categorization aims to use limited resources as efficiently as possible, meaning that people assigned to the second group receive all support resources, while the other two groups receive none (Allhutter et al., 2020). Transferring these categories to the decision of the APP system makes it clear that unfairness perceptions can also be evoked when students are denied support offers because they are certified as performing either "too well" or "too poorly."

The potential divergence between self-assessment, subjective expectations, and the results and consequences of APP hints at a potential conflict when the university's administration confronts students with unanticipated, error-prone, and potentially inexplicable decisions. Students may perceive decisions as valid and thus fair only when these decisions overlap with their self-assessment and as invalid and unfair otherwise, eventually questioning the entire idea of implementing algorithmic performance predictions.

## Procedural Fairness

Procedural fairness focuses not on the outcome itself but on the specific design and elements of the whole process involving AI applications that lead to a predictive result and a respective decision (Colquitt, 2001; Thibaut & Walker, 1975). Leventhal (1980) indicated various criteria for this, which are intended to ensure procedural fairness. These are *consistency*, *bias-suppression*, *accuracy*, *correctability*, *representativeness*, and *ethicality*. The criterion of *consistency* deals with a decision-making process that is always the same through the APP system. In this area, AI and computer technology, in general, are supposed to have advantages because, unlike human decision-makers, as long as the system runs, there are no issues of fatigue, reduced or misguided attention, and careless mistakes (Kaibel et al., 2019; Lee, 2018). Algorithms should also guarantee a higher degree of *neutrality* in the decision-making process (Araujo et al., 2020;

Logg et al., 2019). Therefore, the APP system judges presumably objectively and free of personal interests, interpersonal influence, or other biases that can usually be observed when humans interact. The *accuracy* of an APP system, must be guaranteed by ensuring that it functions technically correctly and that the incoming data are complete and error-free. This also includes that the data meet a *representativeness* requirement and that all relevant parties, especially minorities, can express their free opinions in the decision-making process and that their values and concerns (Lee et al., 2019) as well as *ethical* principles are taken into account (Slade & Prinsloo, 2013). Regarding issues of *correctability*, Thibaut and Walker (1975) also emphasized the possibility of participation in the decision-making process. Questions about the external human oversight of technical systems and the possibilities of human intervention in the decision-making process arise here (e.g., human-in-the-loop vs. out of the loop; Amershi et al., 2014; Holzinger, 2016; Starke & Lünich, 2020). In special cases, decisions derived from APP must be overturned, for instance, when it becomes evident that a mistake in the dataset leads to an unfavorable outcome for a student. According to the literature, it is imperative to ensure the accountability for decisions made by AI systems in this context (Busuioc, 2020; Diakopoulos, 2016), which becomes especially important in the context of responsible AI deployment that intervenes automatedly and determines the life trajectories of students in higher education.

## Interactional Fairness

The third dimension of interactional fairness relates to personal interactions between decision-makers and the perception of fair treatment of their subjects (Acikgoz et al., 2020; Bies & Moag, 1986). In higher education, this relates to the ethically appropriate handling of students and decisions that affect them in the APP process. The focus here is on the appropriate treatment (i.e., dignified and respectful) of those affected (Cropanzano & Greenberg, 1997; Tyler, 2000). Interactional fairness thus also deals with the ethical standards and moral challenges of the use of an APP system. First and foremost stands the question of whether and in what form machines should make moral-laden and consequential decisions at all (Awad et al., 2018). If the decision is to make use of APP, decision-makers must ensure that the students feel valued and appreciated as subjects. However, such decisions may bring questions of perceived data security and privacy violations to the fore. Accordingly, when developing an APP system, it is important to consider which data students consider most important and which data they are willing to disclose in terms of the expected benefits of the performance prediction system (Altman et al., 2018; Daniel, 2015; Ekowo & Palmer, 2017). Here, administrators and programmers must find a suitable way in each implementation case to exploit the potential of AI applications while taking into account the privacy demands and normative expectations of students who want to be treated as individuals with dignity.

### Informational Fairness

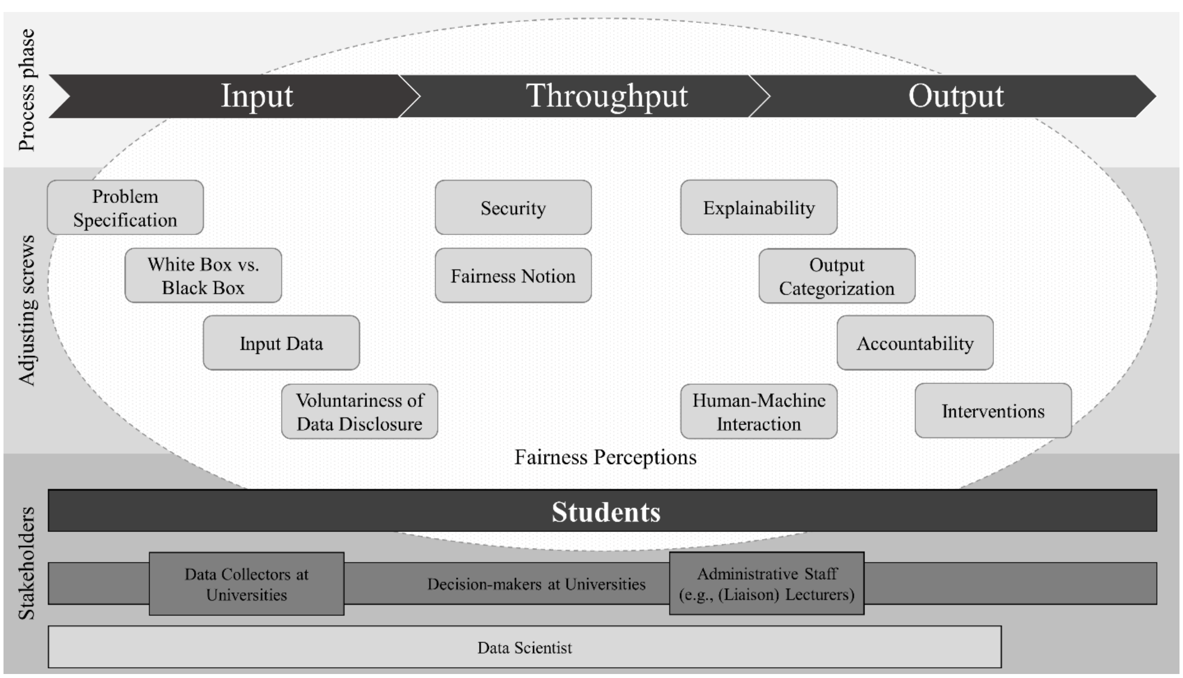
Lastly, informational fairness pertains to the requested levels of transparency and explainability of an APP system. In a system that lacks transparent and comprehensible information about predicted performance and respective decisions, the risk exists of creating another layer of the so-called black box of AI, whose processes and results may not be understandable to students, lecturers, and responsible university administrators (Prinsloo, 2020). Consequently, due to the process' impenetrability, there may not be any tangible starting points for complaints against perceived unfairness (Shin & Park, 2019). By contrast, explanations that explicate and justify a decision can help it be perceived as fair (Bies & Shapiro, 1988). Further, a good explainability of the APP system helps both the students and the university to show where the sticking points are in their study process (i.e., which factors have a significant positive or negative influence on their academic success) so that targeted interventions can be made (Arnold & Pistilli, 2012; Attaran et al., 2018; Daniel, 2017). Furthermore, only a transparent presentation of the functioning of the APP system enables a well-founded opinion to be formed about the underlying decision-making process as a whole. This requires information, for example, about the input data, the fairness criteria used, or even who has access to the results of the APP system. Although studies indicate that transparency can in principle lead to higher perceived fairness (Perez Vallejos et al., 2017; Wang, 2018), it is first necessary to determine how much transparency is desired by students in the first place, in light of the complex ML techniques underlying AI systems, which are generally not understood by ordinary citizens. Accordingly, it cannot necessarily be assumed that maximizing transparency and explainability is always possible and invariably leads to higher perceived informational fairness by default (de Fine Licht & de Fine Licht, 2020).

## ADJUSTING SCREWS IN THE DEVELOPMENT AND IMPLEMENTATION OF AN ACADEMIC PERFORMANCE PREDICTION SYSTEM

In developing a socially responsible APP system that takes particular account of students' interests, several adjustment screws need to be carefully set. For this purpose, it makes sense to look at the different levels of the development of an APP system individually. Therefore, in what follows, we consider the input, throughput, and output phases of an APP system and raise critical questions regarding the human-centered design of the socio-technical AI system (Lee et al., 2017; Shneiderman, 2020). In this regard, different stakeholders become relevant in each phase, as they influence the selection of methods and implementation, and thus interact with the APP system (Kitchin, 2017; Shin, 2019). In addition to the students, these also include faculty, computer scientists and decision-makers within the university. Unlike many scholars who argue from the perspective of computer science and highlight important issues and decisions at the level of technical development (Lepri et al., 2018; Veale et al., 2018; Zweig et al., 2018), we want to foreground students' perceptions. Therefore, in the input phase of our framework (see Figure 1), we do not focus on the mathematical or computer science development of the algorithm; instead, we examine the aspects on which students can form an opinion, such as the degree of complexity of the ML model or which input data should be taken into account (Zweig et al., 2018). During the throughput phase, the training data are merged with the selected methods, and quality criteria (Zweig et al., 2018), such as selecting an appropriate fairness notion and security precautions for the APP system, are defined. Lastly, considering the output phase, it is necessary to discuss interventions that could be derieved from the APP, how the decision concerning such interventions is communicated, who is accountable for possible failures of the system, and what can be done about any issues.

As we have already shown, uncovering and addressing ethical concerns is elementary from a university perspective to avoid decreasing student retention in the worst case or to optimize it in the best case. The concerns of those affected also allow essential conclusions to be drawn about the potential risks and harm that could come from an APP system that is not socially responsible. To better assess this risk potential, we raise critical questions about the process's input, throughput, and output phases and draw an ideal-typical APP system from a normative perspective. Furthermore, we show which fairness dimensions can influence relevant points from the students' perspectives. Against the background of the ideal-typical approach, it is important to keep in mind that the individual process phases and fairness dimensions are not always distinct. Further, various trade-offs imply that not all parameters of the APP system can be maximized simultaneously (Binns & Gallo, 2019).

*Figure 1. Framework of perceived APP-Fairness*



## Input

The first steps in the development and implementation of a socially responsible APP system address the question of the actual purpose and goal of the system. The foundation of the APP system must then be laid by selecting an appropriate ML model, operationalizing student performance, and determining the required input data.

### *Problem Specification of the Academic Performance Prediction System*

At the very beginning of the APP system development process, according to Fazelpour and Danks (2021), there is a need for “problem specification” (p. 4; see also Berendt, 2019; The Institute for Ethical AI in Education, 2021). From this specification, potential discrimination may arise immediately. Concerning the APP system, the central problem the system is supposed to solve and the goal it is pursuing must be identified. For example, an APP system can predict students' performance (e.g., Alyahyan & Düştegör, 2020), and closely related to this is the option of predicting dropouts (e.g., Askinadze & Conrad, 2019; Berens et al., 2019). From the students' point of view, dropout prediction could appear problematic, as this prediction implies a certain finality. However, a performance prediction does not necessarily imply a consequence but initially only offers information so that further reactions can follow. Besides the problem of defining student success, any performance level or likelihood of dropout then needs to be quantified. However, since these steps require specific knowledge and value judgments, conflicts may arise (Fazelpour & Danks, 2021). Thus, problem specification also connects to whether and in what form the APP should be taken by an automated system (Awad et al., 2018), whichis therefore tangential to interactional fairness perceptions. In addition to the perceptions and desires of students and university administrators, the views of faculty should also be included at this point to ensure that their efforts toward good teaching are supported and not undermined by the APP system.

### *Machine Learning Models - Black Box vs. White Box Approaches*

Once the use of the APP system is deemed appropriate and the goal is defined and quantified, a decision must be made on the ML model to be used. As already mentioned, different types of models can be distinguished. From the students' point of view, the distinction between white box and black box models is most decisive at this point. Although black box models can often achieve a higher level of accuracy, the decision-making process is no longer comprehensible (Gunning et al., 2019). All that is known is what input data are fed into the model and what the output looks like. The throughput level, however, remains hidden. Rudin (2019) questioned the feared trade-off between the model's accuracy and its explainability and argued instead for using models that can be interpreted from the outset (e.g., decision trees). The use of white box models enables students and the university to understand the factors of academic success or failure. Further, the transparency gained should positively affect the perception of the informational fairness of the APP system (Perez Vallejos et al., 2017; Wang, 2018). As de Fine Licht and de Fine Licht (2020) pointed out, what is particularly important here is the justification of the decision (i.e., the output of the APP system) rather than the disclosure of the code or the mathematical procedures behind the APP system. Accordingly, input features that are ultimately decisive for the APP should be made traceable, since the underlying mathematical calculations are usually too complex for people without computational science expertise (Dogruel et al., 2020).

### *Choice of Input Data*

Once the choice for a particular algorithm has been made, the question arises as to which input data to use in training and validating it. It is feared that the selection of data or, in particular, the omission of some data (e.g., for reasons of fairness or feared privacy violations) can lead to losses on the accuracy side of the prediction system (Machanavajjhala et al., 2011). In this context, Aggarwal et al. (2021) tested different algorithms for APP, which they fed either with academic input data with or without non-academic parameters (e.g., gender, age, location, and parents' income). They concluded that only a combination of both types of data enables the most effective prediction. Nevertheless, against the background of interactional fairness, concerns about potential privacy violations should be taken seriously. Such considerations also become important in the context of the risk of discrimination and the use of protected attributes. Consequently, some researchers call fordispensing with gender and race characteristics altogether to allow a higher degree of fairness (Barocas & Selbst, 2016). As a rule, however, this does not bring the desired success since correlations between different input variables can serve as a proxy and thus merely substitute for the protected attributes (Fazelpour & Danks, 2021; Johnson, 2021). To prevent systematic disadvantages through the APP system, high data quality must be ensured. Since discrimination can be reproduced by a bias in the historical training data as well as produced by the algorithm itself (Attaran et al., 2018; Ekowo & Palmer, 2017; Fazelpour & Danks, 2021), special care must be taken in the selection and preparation of the data.

### *Voluntariness of Data Disclosure*

To limit the risk of crossing personal privacy boundaries, an important factor may be the voluntary nature of participation in the APP. From this perspective, it would be advisable to leave it up to the students themselves whether they are willing to provide personal data for the APP system to benefit from the given predictions and possibly follow interventions. Thus, Zweig (2019) posited the possibility of escaping the system as a crucial influence on the potential harmfulness of the system. Although the author considered this approach in terms of society as a whole (e.g., the possibility of going to another university that does not deploy an automated APP), we argue that the voluntary nature of participation can have a decisive influence on perceptions of fairness without the need to change universities. This option allows students to make an individual cost-benefit calculation as to whether they are willing to be classified by the APP system, for example, at the price of offering personal data. Beyond this, however, it must be taken into account that ADM systems rely on large amounts of training data and that such data are preferably representative of the student body. Thus, should the voluntariness of participation lead to the majority or specific groups of students deciding against an automatic APP, this, of course, fundamentally questions the need for and benefit from such a system. This is especially true against the background of the already-mentioned consequences for the respective institution, which, in the worst case could, lead to students leaving the university (Marcinkowski et al., 2020).

## Throughput

To prevent these consequences, however, it is necessary to optimally adjust all parameters of the APP system so that the conviction of a socially responsible APP system may push the question of voluntariness into the background altogether. From the student's point of view, in the throughput phase, the adjusting screws of a socially responsible APP system deal with the security precautions of the system, the implemented idea of fairness, and the human–machine interaction between students, the university administration, and the APP system.

### *System and Data Security*

Ensuring security is a crucial requirement for AI applications in general (Jobin et al., 2019). On the one hand, this refers to the security of collecting, processing, and storing personal data and thus addresses some of the concerns regarding the privacy violations mentioned. On the other hand, it is important to protect the system from possible manipulation (Wirtz et al., 2018). Therefore, on a technical level, it is essential to store the data on secure servers and protect the system from hacker attacks, as is required in numerous ethical guidelines for AI development (e.g., AI HLEG, 2019; Jobin et al., 2019; OECD, 2021). First, it is important to prevent unauthorized persons from gaining access to both the input data (such as exam results) and the output of the APP system. This also applies to the requirement for neutrality, as a criterion of procedural fairness, to ensure the system's objectivity and prevent the influence of the personal interests of unauthorized persons. Second, the conditions of use need to determine the extent to which employees tasked with using and maintaining the APP system can access data and ML models, tinker with the data input and its processing, or even change its outcomes.

### *Choice of Fairness Notion*

Furthermore, a decision must be made regarding which idea of fairness should be implemented in the APP system. At this point, above all, the perception of distributive fairness comes to the fore. Since the problem of discrimination by algorithms became apparent, numerous fairness notions have been developed (Dunkelau & Leuschel, 2019; Verma & Rubin, 2018) and deployed (Makhlouf et al., 2020; Srivastava et al., 2019; van Berkel et al., 2019). Although these are intended to guarantee that the algorithm itself does not systematically discriminate against individual groups of people, not every notion is equally suitable for every context and/or field of application (Lepri et al., 2018; Wong, 2020). As mentioned in the section on distributive fairness, the foremost decision must be which general idea of justice should be pursued. Since this objective must be implemented in the algorithm itself, a respective decision significantly affects the throughput of the APP system, although it also has an impact on the actual output and the perception of the same. In concrete terms, considering the limited resources available at universities, for example, a distribution according to *need* criteria could be considered fair. In this sense, students who need special support should be provided with appropriate measures and resources. Once the choice is made on the abstract distributional norm *need*, however, stakeholders sould then determine how best to operationalize this norm and which of the numerous fairness notions is best suited to implement this distributional norm without discriminating against individual student groups through the APP system.

### *Human–Machine Interaction*

Since we understand the APP system as a socio-technical system that is decisively shaped by interactions with developers, university administrators, and students, the element of human–machine interaction spans across all phases (input, throughput, output) and also becomes particularly relevant for interactional but also for informational fairness. Concerning human-computer interaction, a broad strand of literature (e.g., Lee & Rich, 2021; Liang & Lee, 2017) focuses mainly on reactions toward computer-based systems, robots, and AI within socio-technical contexts. In shaping the interaction between the APP system and the different stakeholders, respectful and appreciative interaction is of central importance, especially for the perception of interactional and procedural fairness. In addition to the need to consider the students' interests, one issue that needs to be addressed at this point is, for example, whether the students, as those affected by the APP, are in direct contact with the APP system (e.g., by directly receiving an automated email with the result of the APP) or whether the university administrators handle the transmission of the result. Based on two studies that address the perception of human and AI-based decisions in the hiring process, Acikgoz et al. (2020) concluded that human interaction is recommended when using AI-based decision-making to achieve a higher perception of fairness and make the applicants feel more valued. In this sense, the idea of human-in-the-loop should be revisited (Amershi et al., 2014; Holzinger, 2016; Starke & Lünich, 2020), and various considerations must be made that need to be empirically tested in the context of the APP system. For instance, it is unclear whether the involvement of lecturers or administrative staff in the process of the APP system can actually lead to the desired appreciation and hoped-for success or whether human intervention is somewhat undesirable from the students' point of view. A conceivable scenario would be, for example, a liaison lecturer who receives the result of the automated APP getting in personal contact with the students to discuss potential interventions (see below). Alternatively, all lecturers could have access to the output. Whereas in a negative interpretation, this could lead to (unconscious) prejudices in dealing with the students, a positive assessment would highlight the increased awareness of teachers regarding the challenges of a student and, consequently, the provision of supportive assistance. From the lecturer's perspective, their own autonomy also comes to the fore against the backdrop of human-in-the-loop considerations. Thus, excluding faculty from the process could lead to perceived threats to their own efforts in teaching and supporting students.

## Output

Concerning the output, there are again various attributes to consider for a socially responsible APP system. In addition to determining the concrete design of the output, this also includes the explainability and accountability of the output. Furthermore, possible interventions need to be discussed.

### *Output categorization*

First, we revisit the question of the categorization of the output, which is an essential component of the perceived distributive fairness of the APP system. On the one hand, this is closely related to the problem specification, as the definition of a concrete goal should also consider the desired output (i.e., how many and what kinds of performance levels students will ultimately be grouped into). On the other hand, these goals and the associated output must also be communicated appropriately. Discrepancies between the treatment expectations of the students and the actual prediction and its consequences, in turn, could lead to perceived informational unfairness and dissatisfaction (Fazelpour & Danks, 2021).

As noted above, the implementation of the APP system could conceivably follow a classification along the lines of the algorithm used by the Public Employment Service Austria (Allhutter et al., 2020). However, we have already pointed out that a division into (1) high-performing students, (2) low-performing students with the potential for intervention, (3) and low-performing students who are not seen as having promising potential for intervention has great potential for conflict, especially if this means that some students are denied support measures. This would be particularly unacceptable if an APP designed in this way created incentives for students to intentionally degrade their personal performance to receive additional support measures (Fazelpour & Danks, 2021).

### *Explainability of Predictions*

The demands for transparency and explainability of AI systems are central to the requirements of various ethical guidelines for human-centered AI (Jobin et al., 2019). Here, explainability refers to the output as well as to the entire process of decision-making. To minimize or even better avoid the previously mentioned discrepancies between student expectations and actual predictions, the goal of the APP system and the form of output to be expected should be clearly communicated. To enable an evaluation of the procedural fairness of the APP system, an explanation of the individual elements of the application is needed. In addition to the demand for clarification about the basis of the APP system, stakeholders should also disclose what would need to be changed to achive a different result next time (Wachter et al., 2017), or precisely what influencing factors lead to a good or bad individual APP. Thus, transparency is not exclusively relevant for the perception of informational fairness but is also a fundamental condition for assessing the remaining fairness dimensions. The focus is on the understandability of explaining the decision-making process and the results from the students' point of view. Accordingly, explanations must always be adapted to those affected so that they contain an appropriate amount of information from their point of view while reducing the technical complexity of the system and presenting it in an understandable way (Kasinidou et al., 2021). Again, it becomes clear that the use of interpretable white box models can provide a key advantage at this point, as methods such as decision trees directly reveal the factors that are responsible for the outcome (Rudin, 2019).

### *Accountability in the Case of Errors and Criticism*

The assumption of important decisions by algorithmic systems instead of human decision-makers makes responsibilities opaque. Further, various stakeholders are involved in the process of developing and deploying the APP system. Therefore, the question of who should take responsibility for possible wrong decisions is often discussed (Wirtz et al., 2018). Especially against the background of the risk of discrimination through faulty decisions, the issue of accountability comes to the fore (Busuioc, 2020; Diakopoulos, 2016). In terms of correctability, as a criterion of the procedural fairness dimension, having a say and appeal against the decision should be possible (Leventhal, 1980; Thibaut & Walker, 1975). Therefore, the questions of human-in-the-loop and the design of human–machine interactions are closely connected with the accountability of the decision of the APP system. Thus, the users of the system (i.e., the decision-makers of the universities) have an obligation to justify and account for both the use of the APP system and the resulting predictions and consequences. Furthermore, accountability must be made transparent and openly communicated to students about whom they can contact in case of potential criticism or problems (Diakopoulos, 2016).

### *Interventions Based on Academic Performance Predictions*

Once the APP results come into being, the question arises as to what interventions the university's administration deploys to achieve its ultimate goal of increasing study performance and student retention. Of course, the prediction of promising interventions is perceivably an additional area for applying ML and ADM that goes beyond the APP. However, we do not want to focus on predicting appropriate interventions.

More importantly, concerning the crucial perceptions of (distributive) fairness, the framework highlights the interplay of potential interventions and the institutional setup of how they are conveyed to individual students and what action is subsequently required from them. First, it is critical to discuss the kinds of interventions that may emerge from the APP. However, it is important to keep in mind that neither the decision of the APP system nor the assignment and implementation of the interventions are meant to replace the human efforts of the lecturers.

In the context of higher education, interventions can be differentiated between academic interventions and non-academic interventions. Academic interventions aim to improve academic performance by supporting learners in mastering essential scientific skills or content. Non-academic interventions, by contrast, do not directly aim to improve academic performance by supporting the learner in achieving academic goals. Examples of non-academic interventions are mentoring, helping the student cope with extra-curricular activities, and supporting the student in organizing their daily schedule.

Further, interventions may be divided into generally supportive and sanctioning consequences for students. Supportive measures may – depending on the norm of *equity* or *need* – include support services and measures such as counseling (e.g., psycho-social or work-related guidance); additional courses and tutoring; a reward system for high-performing students or students who graduate in standard study time (e.g., certificates and honors for excellent students, a Dean's list); and the preferential admission to classes based on APP. Sanctioning measures may include debarring students from receiving certain resources or even pre-emptive de-registration. The listed measures are not exhaustive, and there is most often a fine line between whether a measure is eventually perceived as supportive or sanctioning by students. Thus, any measures depicted above may be perceived as supportive or sanctioning depending on their nature of being facultative or compulsory and with respect to the self-assessed performance level of a student.

Moreover, such perceptions may also hinge on how students receive information about their personal APP. As Fazelpour and Danks (2021) suggested, "Student support might be determined not directly by an algorithm, but rather by an administrator who receives algorithmic guidance. In such cases, the epistemic and ethical quality of the overall decision will depend not only on properties of the algorithm but also on how people understand and integrate its outputs into their deliberations" (p. 8). Thus, the involvement of human administrators needs to be taken into consideration, prompting the need to address the following questions: Who contacts students to discuss and assign interventions? When, where, and how do such discussions take place? Is the procedure of making contact and assigning interventions standardized or is there some leeway for administrators and students? For instance, should faculty members contact students (Arnold & Pistilli, 2012), or should they be contacted by specially trained liaison lecturers and social workers? Here, too, issues of fairness regarding privacy, explainability, and correctability of intervention decisions arise.

## FUTURE RESEARCH DIRECTIONS

Regarding APP in higher education, Fazelpour and Danks warned that "morally problematic decisions and unjust harms could result from a biased algorithm supporting an unbiased human, an unbiased algorithm supporting a biased human, or both being biased" (2021, p.8). Consequently, in this chapter, we raised many critical questions regarding the design and implementation of a socially responsible APP system at higher education institutions using a framework that highlights specific issues regarding the input, throughput, and output phases of the APP implementation process. Although we identified certain screws and suggested possible adjustments, the contextually optimal design parameters need further empirical support. As the research strand on the perceived fairness of ADM systems grow steadily, it has become apparent that perceptions of fairness are very context-specific (Lepri et al., 2018; Starke et al., 2021; Wong, 2020). So far, however, empirical studies have often referred to the application context of the COMPAS system (e.g., Grgić-Hlača et al., 2018) or hiring processes (e.g., Acikgoz et al., 2020). To the best of our knowledge, only the study by Marcinkowski et al. (2020) addresses the perceived fairness of AI systems in the higher education context. As perceptions can vary significantly from use case to use case and from affected person to another affected person in another context, experimental studies in particular can help clarify the questions raised above. As long as the students' perceptions are in the foreground, these studies will also offer the possibility of drawing conclusions about other potential AI applications at universities, such as algorithmic study admissions, automated grading, or virtual tutors.

Although we adopted an ideal-typical normative perspective in this chapter, not all attributes of the APP system can be optimized without making sacrifices at other levels. Thus, the examination of trade-offs should also be considered and enlightened by empirical evidence (Binns & Gallo, 2019; Kieslich et al., 2021, pre-print). For instance, we have already discussed feared trade-offs between accuracy and explainability or between accuracy and privacy. Also conceivable are trade-offs due to the suggested voluntary nature of participation in the APP in connection with the system's accuracy since the system depends on training with large amounts of representative data. The proposed framework proves fruitful for the necessary empirical research that needs to be conducted to adjust the screws in the implementation process of APP systems in higher education.

## CONCLUSION

Against the background of the risk of discrimination by ADM systems and the demands for human-centered AI (Jobin et al., 2019; Lee et al., 2017; Shneiderman, 2020), the socially responsible development and implementation of an APP system must take into account the interests of those affected. In institutions of higher education, this means the perceptions of students. Therefore, this chapter aimed to outline an ideal-typical APP system that, on the one hand, points out possible risks of the system by raising critical questions about the design of the APP system, and, on the other hand, shows which perceptions of fairness or unfairness can be triggered by the adjustment of certain elements in the phases of input, throughput, and output within proposed systems. To this end, a conceptual framework of *perceived APP-fairness* was presented, drawing on the established four-dimensional concept of organizational justice (Greenberg, 1993). Of particular importance is considering the system to be a socio-technical system (Kitchin, 2017; Shin, 2019) that continually involves various stakeholders, such as university administrators or students. A socially responsible system can only be achieved through the mutual interaction between stakeholders and the APP system. For this purpose, we provided a broad overview from the specific objective to the choice of ML model to the input data in the first phase. In the throughput phase, security-related aspects of the APP system were addressed, and the careful selection of an appropriate fairness notion was emphasized. Finally, accountability and explainability were emphasized, and considerations for output categorization and potential interventions were provided. We also highlighted avenues for future empirical research. Such research should focus on the fine-tuning of individual elements from students' perceptions. In this sense, the framework presented here is intended to provide a decisive contribution for decision-makers at universities who will sooner or later have to make far-reaching decisions for or against the implementation of a wide variety of AI systems at universities.

## ACKNOWLEDGMENT

This research was supported by the German Federal Ministry of Education and Research [grant number 16DHB4020]. We also thank Albina Maxhuni and Cara Schönmüller for their support as well as the reviewers for their much-appreciated feedback on the first version of this chapter.

### *References*

Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment*, *28*(4), 399–416. https://doi.org/10.1111/ijsa.12306

Adams, J. S. (1963). Toward an Understanding of Inequity. *Journal of Abnormal and Social Psychology*, *67*(5), 422–436.

Adams, J. S. (1965). Inequity In Social Exchange. *Advances in Experimental Social Psychology*, *2*, 267–299. https://doi.org/10.1016/S0065-2601(08)60108-2

Adams Becker, S., Cummins, M., Davis, A., Freeman, A., Hall Giesinger, C., & Ananthanarayanan, V. (2017). *NMC Horizon Report: 2017 Higher Education Edition*. Austin, Texas. The New Media Consortium. https://www.sconul.ac.uk/sites/default/files/documents/2017-nmc-horizon-report-he-EN.pdf

Aggarwal, D., Mittal, S., & Bali, V. (2021). Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques. *International Journal of System Dynamics Applications*, *10*(3), 38–49. https://doi.org/10.4018/IJSDA.2021070103

Ahmed, S. A., & Khan, S. I. (2019). A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective. In *Proceedings of the 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCCNT45670.2019.8944511

AI HLEG. (2019). *Ethics Guidelines for Trustworthy AI*. High-Level Expert Group on Artificial Intelligence set up by the European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020). Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data*, *3*(5), 1–17. https://doi.org/10.3389/fdata.2020.00005

Altman, M., Wood, A., & Vayena, E. (2018). A Harm-Reduction Framework for Algorithmic Fairness. *IEEE Security & Privacy*, *16*(3), 34–45. https://doi.org/10.1109/MSP.2018.2701149

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(3), 1–21. https://doi.org/10.1186/s41239-020-0177-7

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, *35*(4), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Araujo, T., Helberger, N., Kruikemeier, S., & Vreese, C. H. de (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Arnold, K. E., & Pistilli, M. D. (2012). *Course Signals at Purdue: Using Learning Analytics to Increase Student Success*. Vancouver, BC, Canada, 29th-May 2nd. 2nd International Conference on Learning Analytics and Knowledge. http://dl.acm.org/citation.cfm?id=2330601

Askinadze, A., & Conrad, S. (2019). Predicting Student Dropout in Higher Education Based on Previous Exam Results. In M. C. Desmarais, C. F. Lynch, A. Merceron, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining: EDM, Montréal, Canada, July 2-5, 2019. International Educational Data Mining Society (IEDMS)* (pp. 500–503).

Askinadze, A., Liebeck, M., & Conrad, S. (2018). Predicting Student Test Performance based on Time Series Data of eBook Reader Behavior Using the Cluster-Distance Space Transformation. In Ogata, Hiroaki, Song, Yanjie, Yang, Jie-Chi, Chang, Maiga, Wong, Lung-Hsiang & M. M. T. Rodrigo (Eds.), *ICCE 2018: 26th International Conference on Computers in Education, extended summary proceedings.* Taiwan: Asia-Pacific Society for Computers in Education.

Attaran, M., Stark, J., & Stotler, D. (2018). Opportunities and challenges for big data analytics in US higher education: A conceptual model for implementation. *Industry and Higher Education*, *32*(3), 169–182. https://doi.org/10.1177/0950422218770937

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). *Predicting Student Dropout in Higher Education.* ICML Workshop on #Data4Good: Machine Learning in Social Good Applications. http://arxiv.org/pdf/1606.06364v4

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*, 59–64. https://doi.org/10.1038/s41586-018-0637-6

Barocas, S., & Selbst, A. D. (2016). Big Datas Disparate Impact. *California Law Review*, *104*(3). https://doi.org/10.15779/Z38BG31

Berendt, B. (2019). AI for the Common Good?! Pitfalls, challenges, and ethics pen-testing. *Paladyn, Journal of Behavioral Robotics*, *10*(1), 44–65. https://doi.org/10.1515/pjbr-2019-0004

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk: Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, *11*(3), 1–41. https://doi.org/10.5281/zenodo.3594771

Bies, R. J., & Moag, J. S. (1986). Interactional justice: Communication criteria for fairness. *Research on Negotiation in Organizations*, *1*, 43–55.

Bies, R. J., & Shapiro, D. L. (1988). Voice and Justification: Their Influence on Procedural Fairness Judgments. *Academy of Management Journal*, *31*(3), 676–685. https://doi.org/10.5465/256465

Binns, R. (2018). What Can Political Philosophy Teach Us about Algorithmic Fairness? *IEEE Security & Privacy*, *16*(3), 73–80. https://doi.org/10.1109/MSP.2018.2701147

Binns, R., & Gallo, V. (2019). *Trade-offs*. Information Commissioner's Office. https://ico.org.uk/about-the-ico/news-and-events/ai-blog-trade-offs/

Busuioc, M. (2020). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, 1–12. https://doi.org/10.1111/puar.13293

Calders, T., & Žliobaitė, I. (2013). Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and Privacy in the Information Society: Data Mining and Pofiling in Large Databases* (pp. 43–57). Springer Berlin Heidelberg.

Colquitt, J. A. (2001). On the Dimensionality of Organizational Justice: A Construct Validation of a Measure. *Journal of Applied Psychology*, *86*(3), 386–400. https://doi.org/10.1037//0021-9010.86.3.386

Cropanzano, R., & Greenberg, J. (1997). Progress in Organizational Justice: Tunneling Through the Maze. *International Review of Industrial and Organizational Psychology*, *12*, 317–372.

Cropanzano, R., Rupp, D. E., Mohler, C. J., & Schminke, M. (2001). Three Roads to Organizational Justice. In G. R. Ferris (Ed.), *Research in Personnel and Human Resources Management, Vol. 20* (pp. 1–113). Emerald Group Publishing Limited. https://doi.org/10.1016/S0742-7301(01)20001-2

Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, *46*(5), 904–920. https://doi.org/10.1111/bjet.12230

Daniel, B. (2017). Overview of Big Data and Analytics in Higher Education. In B. Daniel (Ed.), *Big Data and Learning Analytics in Higher Education: Current Theory and Practice* (pp. 1–4). Springer International Publishing.

Data Ethics Commission. (2019). *Opinion of the Data Ethics Commission: Executive Summery*. Data Ethics Commission of the Federal Government. Federal Ministry of the Interior, Building and Community, Germany. https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & SOCIETY*, *35*(4), 917–926. https://doi.org/10.1007/s00146-020-00960-w

Dede, C., Ho, A., & Mitros, P. (2016). Big Data Analytics in Higher Education: Promises and Pitfalls. *EDUCAUSE Review*, *51*(5).

Deutsch, M. (1975). Equity, Equality, and Need: What Determines Which Value Will Be Used as the Basis of Distributive Justice? *Journal of Social Issues*, *31*(3), 137–149. https://doi.org/10.1111/j.1540-4560.1975.tb01000.x

Diakopoulos, N. (2016). Accountability in Algorithmic Decision Making. *Communications of the ACM*, *59*(2), 56–62. https://doi.org/10.1145/2844110

Dogruel, L., Facciorusso, D., & Stark, B. (2020). 'I'm still the master of the machine.' Internet users' awareness of algorithmic decision-making and their perception of its effect on their autonomy. *Information, Communication & Society*, 1–22. https://doi.org/10.1080/1369118X.2020.1863999

Dunkelau, J., & Leuschel, M. (2019). *Fairness-Aware Machine Learning: An Extensive Overview*. Working Paper No. 2. Düsseldorf. Heinrich-Heine-University. https://www.sozwiss.hhu.de/fileadmin/redaktion/Fakultaeten/Philosophische_Fakultaet/Sozialwissenschaften/Kommunikations-_und_Medienwissenschaft_I/Dateien/Dunkelau___Leuschel__2019__Fairness-Aware_Machine_Learning.pdf

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness. In S. Goldwasser (Ed.), *ACM Conferences. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). ACM.

Ekowo, M., & Palmer, I. (2016). *The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis*. https://www.newamerica.org/education-policy/policy-papers/promise-and-peril-predictive-analytics-higher-education/

Ekowo, M., & Palmer, I. (2017). *Predictive Analytics in Higher Education: Five Guiding Practices for Ethical Use*. https://na-production.s3.amazonaws.com/documents/Predictive-Analytics-GuidingPractices.pdf

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*(8), e12760. https://doi.org/10.1111/phc3.12760

Frouillou, L., Pin, C., & van Zanten, A. (2020). APB and Parcoursup platforms to provide equal opportunities? Evolution of procedures and norms in access to higher education in France. *L'année Sociologique*, *70*(2), 337–363. https://doi.org/10.3917/anso.202.0337

Greenberg, J. (1987). A Taxonomy of Organizational Justice Theories. *Academy of Management Review*, *12*(1), 9–22. https://doi.org/10.5465/amr.1987.4306437

Greenberg, J. (1990). Organizational Justice: Yesterday, Today, and Tomorrow. *Journal of Management*, *16*(2), 399–432. https://doi.org/10.1177/014920639001600208

Greenberg, J. (1993). The Social Side of Fairness: Interpersonal and Informational Classes of Organizational Justice. In R. Cropanzano (Ed.), *Justice in the workplace: Approaching fairness in human resource management* (pp. 79–103). Lawrence Erlbaum Associates, Inc.

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human Perceptions of Fairness in Algorithmic Decision Making. In P.-A. Champin, F. Gandon, M. Lalmas, & P. G. Ipeirotis

(Eds.), *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18* (pp. 903–912). ACM Press. https://doi.org/10.1145/3178876.3186138

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai - Explainable Artificial Intelligence. *Science Robotics*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

Hamoud, A. K., Hashim, A. S., & Awadh, W. A. (2018). Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(2), 26–31. https://doi.org/10.9781/ijimai.2018.02.004

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131. https://doi.org/10.1007/s40708-016-0042-6

The Institute for Ethical AI in Education. (2021). *The Ethical Framework for AI in Education*. https://www.buckingham.ac.uk/research-the-institute-for-ethical-ai-in-education/

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, G. M. (2021). Algorithmic bias: on the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y

Kaibel, C., Koch-Bayram, I., Biemann, T., & Mühlenbock, M. (2019). Applicant Perceptions of Hiring Algorithms - Uniqueness and Discrimination Experiences as Moderators. *Academy of Management Proceedings.* Advance online publication. https://doi.org/10.5465/AMBPP.2019.210

Kasinidou, M., Kleanthous, S., Barlas, P., & Otterbacher, J. (2021). "I agree with the decision, but they didn't deserve this";: Future Developers' Perception of Fairness in Algorithmic Decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 690–700). ACM. https://doi.org/10.1145/3442188.3445931

Keller, B., Baleis, J., Starke, C., & Marcinkowski, F. (2019). *Machine Learning and Artificial Intelligence in Higher Education: A State-of-the-Art Report on the German University Landscape*. Working Paper No. 1. Düsseldorf. Heinrich-Heine-University. https://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Sozialwissenschaften/Kommunikations-_und_Medienwissenschaft/KMW_I/Working_Paper/Keller_et_al.__2019__-_AI_in_Higher_Education.pdf

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Kieslich, K., Keller, B., & Starke, C. (2021, pre-print). *AI-Ethics by Design. Evaluating Public Perception on the Importance of Ethical Design Principles of AI*. http://arxiv.org/pdf/2106.00326v1

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, *37*(4), 331–344. https://doi.org/10.1007/s10462-011-9234-x

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 1-16. https://doi.org/10.1177/2053951718756684

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural Justice in Algorithmic Fairness:: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1–26. https://doi.org/10.1145/3359284

Lee, M. K., Kim, J. T., & Lizarondo, L. (2017). A Human-Centered Approach to Algorithmic Services: Considerations for fair and motivating smart community service management that allocates donations to non profit organizations. In G. Mark, S. Fussell, C. Lampe, m. schraefel, J. P. Hourcade, C. Appert, & D. Wigdor (Eds.), *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3365–3376). ACM. https://doi.org/10.1145/3025453.3025884

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Leventhal, G. S. (1980). What Should Be Done with Equity Theory? New Approaches to the Study of Fairness in Social Relationships. In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social Exchange* (pp. 27–55). Springer US.

Liebowitz, J. (2017). Thoughts on Recent Trends and Future Research Perspectives in Big Data and Analytics in Higher Education. In B. Daniel (Ed.), *Big Data and Learning Analytics in Higher Education: Current Theory and Practice* (pp. 7–17). Springer International Publishing.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation:: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Machanavajjhala, A., Korolova, A., & Das Sarma, A. (2011). Personalized Social Recommendations - Accurate or Private? *Proceedings of the VLDB Endowment*, *4*(7). https://arxiv.org/ftp/arxiv/papers/1105/1105.4254.pdf

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). *On the Applicability of ML Fairness Notions.* ArXiv E-Prints. http://arxiv.org/pdf/2006.16745v2

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (Un-)Fairness in Higher Education Admissions. In M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor, & G. Zanfir-Fortuna (Eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 122–130). ACM. https://doi.org/10.1145/3351095.3372867

Muñoz, C., Smith, M., & Patil, D. J. (2016). *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.* Executive Office of the President of USA. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

OECD. (2021). *Recommendation of the Council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift Modeling for preventing student dropout in higher education. *Decision Support Systems, 134,* 1–11. https://doi.org/10.1016/j.dss.2020.113320

Perez Vallejos, E., Koene, A., Portillo, V., Dowthwaite, L., & Cano, M. (2017). Young People's Policy Recommendations on Algorithm Fairness. In P. Fox, D. McGuinness, L. Poirer, P. Boldi, & K. Kinder-Kurlanda (Eds.), *Proceedings of the 2017 ACM on Web Science Conference* (pp. 247–251). ACM. https://doi.org/10.1145/3091478.3091512

Prinsloo, P. (2020). Of 'black boxes' and algorithmic decision-making in (higher) education – A commentary. *Big Data & Society*, *7*(1), 1-6. https://doi.org/10.1177/2053951720933994

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2020). How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, *283*, 1–15. https://doi.org/10.1016/j.artint.2020.103238

Sclater, N., & Mullan, J. (2017). *Jisc briefing: Learning analytics and student success - assessing the evidence.* Jisc. https://repository.jisc.ac.uk/6560/1/learning-analytics_and_student_success.pdf

Sclater, N., Peasgood, A., & Mullan, J. (2016). *Learning Analytics in Higher Education: A review of UK and international practice: Full report.* Jisc. https://www.jisc.ac.uk/sites/default/files/learning-analytics-in-he-v3.pdf https://doi.org/10.1002/aehe.20121

Shin, D. (2019). Toward Fair, Accountable, and Transparent Algorithms: Case Studies on Algorithm Initiatives in Korea and China. *Javnost - the Public*, *26*(3), 274–290. https://doi.org/10.1080/13183222.2019.1589249

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Simmons, R. (2018). Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System. *University of California Davis Law Review*, *52*, 1–39. https://ssrn.com/abstract=3156510

Slade, S., & Prinsloo, P. (2013). Learning Analytics. *American Behavioral Scientist*, *57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2459–2468. https://doi.org/10.1145/3292500.3330664

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). *Fairness Perceptions of Algorithmic Decision Making: A Systematic Review of the Empirical Literature*. https://arxiv.org/ftp/arxiv/papers/2103/2103.12016.pdf

Starke, C., & Lünich, M. (2020). Artificial intelligence for political decision-making in the European Union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data & Policy*, *2*, 1–17. https://doi.org/10.1017/dap.2020.19

Thibaut, J. W., & Walker, L. (1975). *Procedural Justice: A Psychological Analysis*. Lawrence Erlbaum Associates, Inc.

Tyler, T. R. (2000). Social Justice: Outcome and Procedure. *International Journal of Psychology*, *35*(2), 117–125. https://doi.org/10.1080/002075900399411

van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M., & Kostakos, V. (2019). Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–21. https://doi.org/10.1145/3359130

Veale, M., van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM. https://doi.org/10.1145/3173574.3174014

Verma, S., & Rubin, J. (2018). Fairness Definitions Explained. *Proceedings of the ACM/IEEE International Workshop on Software Fairness, FairWare '18*, 1–7. https://doi.org/10.1145/3194770.3194776

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wang, A. J. (2018). Procedural Justice and Risk-Assessment Algorithms. *SSRN Electronic Journal*, 1–31. https://doi.org/10.2139/ssrn.3170136

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2018). Artificial Intelligence and the Public Sector: Applications and Challenges. *International Journal of Public Administration*, *42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Wong, P.-H. (2020). Democratizing Algorithmic Fairness. *Philosophy & Technology*, *33*(2), 225–244. https://doi.org/10.1007/s13347-019-00355-w

Yanosky, R., & Arroway, P. (2015). *The Analytics Landscape in Higher Education, 2015*. Educause Center for Analysis and Research (ECAR). https://library.educause.edu/-/media/files/library/2015/5/ers1504cl.pdf

Zweig, K. A. (2019). Algorithmische Entscheidungen: Transparenz und Kontrolle. *Analysen & Argumente. Digitale Gesellschaft.*, *338*, 1–16. https://www.kas.de/documents/252038/4521287/AA338+Algorithmische+Entscheidungen.pdf/533ef913-e567-987d-54c3-1906395cdb81?version=1.0&t=1548228380797 (Konrad-Adenauer-Stiftung e. V.).

Zweig, K. A., Fischer, S., & Lischka, K. (2018). *Wo Maschinen irren können: Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung.* Bertelsmann Stiftung. https://doi.org/10.11586/2018006

Zweig, K. A., & Krafft, T. D. (2018). Fairness und Qualität algorithmischer Entscheidungen. In R. Mohabbat-Kar, B. E. P. Thapa, & P. Parycek (Eds.), *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft* (pp. 204–227). Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT.