Algorithmic fairness as sociotechnical system: A typology of the information construct.

Mortaza S. Bargh[a*], *Sunil* Choenni[b], *Floris* ter Braak[c]

[a]Research and Data Centre, Ministry of Justice and Security, The Hague, The Netherlands, email address m.shoae.bargh@wodc.nl, ORCID number 0000-0001-5395-456X.

[b]Affiliation 1: Research and Data Centre, Ministry of Justice and Security, The Hague, The Netherlands, email address r.choenni@wodc.nl, Affiliation 2: Creating 010 Research Centre, Rotterdam Univeristy of Applied Sciences, Rotterdam, The Netherlands, email address r.choenni@hr.nl, ORCID number 0000-0003-2772-6330.

[c]Research and Data Centre, Ministry of Justice and Security, The Hague, The Netherlands, email address f.ter.braak@wodc.nl, ORCID number 0000-0001-9966-3514.

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 19 May 2025

Abstract. Organizations and enterprises search for ways to exploit the vast amount of data that is produced by citizens, sensors, devices and administrative processes. Capitalizing on the produced data should be done responsibly by preventing, mitigating and managing undesired side effects such as violation of rules and regulations, human rights, ethical principles as well as privacy and security requirements. A key challenge in employing data, algorithms and data-driven systems is to adhere to the principle of fairness and justice. In this contribution we focus on the issue of algorithmic fairness, which itself can be framed as a sociotechnical system with interacting social and technical/formal subsystems. Information is a key construct of any sociotechnical system, where information creation and exchange can ease the opacity of interactions between the social and formal subsystems, and of interactions between the subsystems and the environment in which they operate. Based on literature, we categorize the types and flows of the information construct within the sociotechnical systems of algorithmic fairness in 7 categories. As such, the presented insights about the 7 categories of the information construct can form a common mental model whereby social and technical disciplines can inform each other systematically and align their views on algorithmic fairness.

Keywords. Algorithmic fairness, fairness, information construct, justice, sociotechnical systems.
Research paper, DOI: https://doi.org/10.59490/dgo.2025.952

1. Introduction

Data are currently being generated, collected, analyzed, and distributed at a fast-growing pace. This growth is due to proliferation of connected devices (such as cameras, smartphones, sensors, and smart household appliances), widespread and intensive use of social networks, and fast-paced digitalization of business and organizational processes/services, among others. Public organizations and private enterprises collect a vast amount of data directly as a necessary input for provisioning their services or as a byproduct of their services. As a result of this growth, there is a rising interest (and demand) to harvest the available data by using Artificial Intelligence (AI) and Machine Learning (ML) algorithms to develop advanced data-driven systems, like decision-support systems, that ease our daily lives, create additional value for businesses, provide insight into societal phenomena, and guide policymaking processes. These data-driven systems are applied to various domains of society like justice, healthcare, education, public safety and security, public administration, transportation and logistics. Specifically within public organizations, AI and ML have gained a foothold in daily

practice for policy making (e.g., to detect social issues quickly, improve policy decisions, and monitor policy implementation), for provisioning public services (e.g., to improve service delivery and develop innovative services), and for managing internal affairs (e.g., to develop management innovations, human resources, procurement and finances), to name a few (Leeuw, 2025). All these have presently resulted in various forms of smart environment paradigms like smart cities, smart healthcare, smart logistics, and smart government.

The current trend of digitization, digitalization and digital transformation worldwide, which relies on digital technology and data-driven systems, has a huge impact on several aspects of our lives. In addition to its potentials, this trend inflicts a wide range of disorders and problems, ranging from sustainability-oriented ones to efficiency-oriented ones (Toli & Murtagh, 2020). Sustainability-oriented problems are related to social values (e.g., equity, community autonomy, citizen well-being, quality of life and gratification of fundamental human needs), economy vitality and diversity, and environment conservation (e.g., flora, fauna, and natural resources). Efficiency-oriented problems are related to efficient management of public and commercial services in various domains of society.

However, as data and data-driven systems are transforming our society on every scale, the so-called smart systems that capitalize on data should not be perceived as pure technological systems. The way that data are collected, which are often blended with sensitive and stigmatizing information about individuals, groups, and businesses, and the way that algorithms and data-driven smart systems are devised, designed, implemented, deployed, and (mis)used (are going to) impact us deeply at individual, group and societal levels. As such, all societal, environmental, technological transformations affected by or derived from data-driven systems must be attentive and respectful of legal, ethical and human rights principles.

Despite the success that data-driven smart systems are experiencing, there are still many key stumbling blocks in their large-scale and real-life deployments. Realizing smart systems, on the one hand, entails several technological challenges like establishing efficient mechanisms for integration, storage, and retrieval of (large volumes of) data, and processing (different types of) data in almost real-time. On the other hand, there are many non-functional challenges for using data and algorithms in practice, also known as soft challenges. Some examples of these challenges are mitigating the security issues of largely distributed data-reliant systems, managing the quality issues of loosely coupled data sources, and dealing with biased and/or wrongful discrimination of individuals and groups (i.e., unfairness) as embedded in collected data or induced due to inattentive design or use of algorithms. Not handling these soft challenges appropriately and adequately would harm individuals, groups and society, adversely affecting basic human rights like privacy, liberty, autonomy and dignity.

In this contribution, we consider decision-support systems, which are a specific case of data-driven smart systems. Often in such systems, AI and ML algorithms are used to train classification models, the trained models are used to predict some outcomes, and the predicted outcomes are used as (partial) evidence for making decisions about individuals (e.g., in credit granting) and/or groups (e.g., in policy-making). A publication by the European Union (EU) office (see Misuraca & van Noordt, 2020) has extensively analyzed the landscape of AI usage in public services across Europe. During the period of May 2019 – February 2020, the report gives an account of 230 initiatives for using AI in public services across Europe. Of these 230 initiatives, there are 29 initiatives (about 13%) categorized as "Expert and Rule-based Systems, Algorithmic Decision Making", which can give a good estimate of the mentioned AI/ML based-decision support-systems for public services. A well known example of such systems in the justice domain is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool. The tool was used for measuring the risk of a defender to recommit crime, i.e., recidivism. Although the tool was designed to be fair from some perspective (Dieterich et al., 2016), a follow up study showed that the COMPAS tool was unfair from another perspective (Larson et al., 2016). Specifically, African American defendants were more likely than Caucasian defendants to be incorrectly flagged as higher risk of recidivism (false positive rate), while Caucasian defendants were more likely than African American defendants to be incorrectly flagged as low risk of recidivism (false negative rate). Note the results of this paper can be applied to other usage areas of algorithms.

As the outcome of such decision-support systems can be biased against social groups and individuals unjustifiably (Choenni et al., 2018; Dolata et al., 2022; Netten et al., 2018; Starke et al., 2022), we aim at addressing the soft-challenge of social biases and wrongful discrimination (i.e., unfairness) when realizing and using data-driven decision-support systems. This unfairness stems from, for example, the biased data used for training algorithms, the bias introduced in such algorithms, and/or the way that the outcomes are interpreted and applied to the practice (Choenni et al., 2021). In other words, we focus on the field of *algorithmic fairness* which

aims at addressing these biases and discrimination when the outcomes of an AI/ML algorithm is used within (i.e., as a part of) social service provisioning. Not only are decision-support systems used within social services sociotechnical systems, but also their algorithmic fairness component can be regarded as a sociotechnical system with two interacting subsystems: a formal/technical subsystem and a social subsystem (Dolata et al., 2022). Sociotechnical systems are generally characterized by some generic constructs in literature. Information is one of these constructs that shapes and is shaped by how the components of a sociotechnical systems interact and it paves the way to achieve the desired goals of a system by providing some sort of order (i.e., by reducing uncertainty/entropy via imparting dialog, meaning, and utility to sociotechnical interactions) (Chatterjee et al., 2021). Towards the research objective described above, we deliver the following contributions in this paper.

- We present a model of the sociotechnical systems corresponding to algorithmic fairness, which conceptualizes the development stages of these systems. The model encompasses fairness conceptualization, design, and operation stages, where each of these stages has footholds in both technical and social subsystems.
- We delve into the information construct of sociotechnical systems and, based on literature, identify and specify 7 information artifacts that may arise due to interactions among the components of the sociotechnical systems associated with algorithmic fairness.

For this study we have conducted extensive content analysis of the literature to derive the insights and the model presented. The conducted literature review can be characterized as a theoretical review that "draws on existing conceptual and empirical studies to provide a context for identifying, describing, and transforming into a higher order of theoretical structure and various concepts, constructs or relationships" (Paré et al., 2015). Our primary goal was to develop a conceptual model that explains the existing structure and the information construct of algorithmic fairness. To this end, we studied and analyzed a number of selected papers that have been published in leading journals and conferences in fields of data science, law, and ethics. The analysis is performed through the lens of applying AI/ML classification outcomes to social context where, due to uncertainty, the impact on individuals and social groups can be unfair and can thus lead to (social) injustice.

The outline of the paper is follows. In Section 2 we provide some background information about algorithmic fairness and present our conceptual model of algorithmic fairness sociotechnical systems. In Section 3 we review the related literature. In Section 4, we present our typology of the information construct for algorithmic fairness, supported by evidences from the literature. Finally, in Section 5 we discuss the results, draw our conclusions, and mention several directions for future research.

2. Algorithmic fairness as a sociotechnical system

In this section we provide some background information and present our first contribution. We start with explaining relevant concepts in Section 2.1. In Section 2.2 we explain the social and formal views on fairness and algorithmic fairness, as well as their pitfalls. Subsequently, in Sections 2.3 and 2.4 we explain the need for adopting a sociotechnical viewpoint and approach, respectively, to address the issues of algorithmic fairness in complex social settings. In Section 2.5, we summarize the outcome of the section.

2.1. Fairness and algorithmic fairness

Fairness in the context of decision-making refers to the "absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" (Mehrabi et al., 2021). Two sources of unfairness (lack of fairness) are bias and wrongful discrimination. (NB., from this point on we will use the term discrimination instead of wrongful discrimination for brevity as well as for conformity to the literature.) In data-driven decision support systems bias can exist in data, algorithms, and/or the way that the outcomes of algorithms are applied to practice (i.e., in interaction between algorithms and users), see (Mehrabi et al., 2021). Some forms of bias can lead to unfairness in different down-stream systems that use the outcomes of data-driven decision support systems. Discrimination is attributed to human prejudice and stereotyping that is based on the sensitive attributes (like gender, race and religion).

Algorithmic fairness refers to those technological solutions that are devised for detecting and preventing systematic harms or benefits to different subgroups and individuals in data-driven decision-making (Dolata et al.,

2022). In other words, algorithmic fairness aims at (a) mathematically quantifying biases in AI and ML based systems using various metrics (or indicators) and (b) devising measures (i.e., solutions) to mitigate unjustified biases and discrimination. We note that protecting both individuals and subgroups is at the focus of such solutions. Further, as will be explained in the following sections, the scope of algorithmic fairness spans beyond technical solutions and intertwines with social solutions (i.e., social measures or interventions) for mitigating undesired biases and discrimination in AI and ML based systems. Therefore, we use *algorithmic-based fairness* to refer to those technical and non-technical solutions in data-driven decision-making processes that can be devised for preventing systematic harms or benefits to individuals or different subgroups (one may refer to this holistic view as algorithmic justice).

Note that algorithmic fairness together with algorithmic transparency (and the related concepts of explainability and interpretability) and cybersecurity (which includes privacy protection) constitute the core components of the field of responsible AI (Chorąś et al., 2020).

2.2. Social and formal views on (algorithmic) fairness

Figure 1 shows our model of the sociotechnical systems corresponding to algorithmic fairness. It actually conceptualizes the development stages of these systems vertically; and the technical and social subsystems of sociotechnical systems horizontally. Conceptually, as shown by conceptualization space in Figure 1, there are two main views on fairness that originate from the social discourse and formal discourse on fairness (see Starke et al., 2022 and Dolata et al., 2022). The social view on fairness can be traced back to disciplines like philosophy, political science, legal science, organization science, criminology, sociology, anthropology, neuroscience, and psychology. The formal view on fairness encompasses attempts to model (aspects of) the social concepts of fairness, especially those borrowed from anthropological and organizational viewpoints. To this end, based on the derived models, one builds formal (i.e., technical) fairness measures and mechanisms in Information System (ISs). While the social discourse inspires the formal discourse (see the link drawn in Figure 1) – thus resulting in mathematical or formal definitions of fairness or the so-called fairness metrics – the formal discourse lays down the ground for the formal aspects of algorithmic fairness.

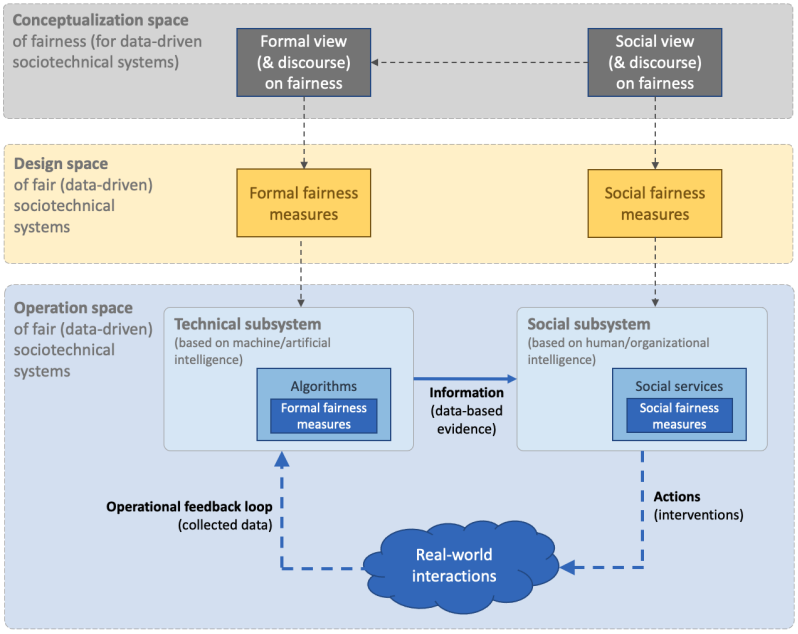

Fig. 1 – An illustration of social and formal aspects of algorithmic fairness in three system development spaces.

In practice, the design space of algorithmic fairness encompasses technical metrics as well as both formal and social measures (solutions), as indicated in the middle of Figure 1. Through design, the formal and social measures of fairness should be combined appropriately to deliver algorithmic-based fairness in a given context. In the operation phase of a data-driven decision support system, as indicated by operation space in Figure 1, the formal and social subsystems, which include formal and social fairness measures, are applied into practice to operationalize the intended intervention. Depending on the operation context, the operation space may encompass group related actions (e.g., making policies) and/or individual related actions (making decisions

about someone). Through interacting with the real-world phenomena and environment, the interaction between technical and social subsystems may lead to emerging behaviors that were not foreseen in the design phase.

2.3. Algorithmic fairness as a sociotechnical system

Investigating algorithmic fairness solutions from the literature, the authors of (Dolata et al., 2022) identify several wrong assumptions that are typically made about algorithmic fairness within the social and formal discourses. For example, in the social discourse one may ignore the differences among various formal fairness measures and treat them equally and/or assume that only human reasoning yields fair decision making (i.e., the so-called black box and purity assumptions). Similarly, in the formal discourse one may assume that there is a technical solution that captures all complex notions of fairness completely (i.e., the so-called equivalence and translation assumptions). Although these assumptions might be stereotypical to some degrees, we think reviewing them is useful for informing about and preventing the possible pitfalls of algorithmic fairness. We argue that system designers may cause harm if they either totally ignore or naively adopt the existing formal solutions.

Each of the social and formal views on fairness puts forward several, to some degree, valid points about algorithm fairness. Nevertheless, there are also interactions between social and formal aspects (and both with environment) that should be considered. In other words, not only is the data driven decision support system a sociotechnical system, so is the algorithmic fairness (aspect of the) system (Dolata et al., 2022). One can characterize algorithmic fairness as a sociotechnical phenomenon because, as illustrated in Figure 1, the process of algorithm design and development is a social practice, since the outcomes of algorithms impact individuals, groups, and society; and algorithms and algorithmic fairness are a matter of public debate (Dolata et al., 2022). As such, the social and formal aspects of algorithmic fairness interact directly or indirectly (i.e., via feedback loops as shown in Figure 1). Further, we note that the nature of these interactions is dynamic and changes in time due to gradual impacts of the interventions triggered by the models, which are a result of the algorithms. Part of this dynamicity can be attributed to the changes that occur in the environment in which the algorithms are deployed due to, for example, introducing new policies, laws, and technologies.

2.4. Adopting a sociotechnical approach

Looking at formal and social aspects individually and optimizing fairness in the formal or social domain separately does not consider the existing interactions and inter-dependencies between them nor does it account for the emerging interactions among social, formal and environmental factors. These interactions and inter-dependencies could be noticed and dealt with if one would consider them together. As the overall outcome of the sociotechnical system should be fair, unfairness threats must be sought in the combined system (like undesired emerging behavior when the formal and social subsystems interact with each other and with the environment). A careful consideration, analysis, or design of machine and human collaborations and engagements in a sociotechnical system relies on the following constructs (Chatterjee et al., 2021; Lee, 2004; Sarker et al., 2019):

- having social and formal subsystems,
- having reciprocal interactions between the social and formal subsystems (and the environment), and
- information arising from the interaction between social and formal subsystems.

In this paper, as to be explained in the following section, we contribute to the basic construct of information within the sociotechnical system of algorithmic fairness (i.e., the last construct mentioned above). The *information* construct is not a subsystem but an emergent property of the overall system that shapes and is shaped by how its social and technical subsystems interact (Chatterjee et al., 2021). "[T]he role of information is to help realize the path (by being well-formed and meaningful) of achieving the desired goals of a system. In other words, information provides some sort of order to a goal-seeking system in its effort to realize those goals, thus reducing entropy" via imparting dialog, meaning, and utility to sociotechnical interactions (Chatterjee et al., 2021). Thus, creating and exchanging information in a well-formed, meaningful, and order-giving way, is a key element/construct of sociotechnical systems. Through contributing to the basic construct of information exchange for the sociotechnical system of algorithmic fairness, we intend to move one step away from the traditional formal notion of algorithmic fairness toward the *sociotechnical* notion of fairness (or *algorithmic based fairness* as we defined in Section 2.1).

2.5. Recapturing on the proposed (mental) model

The model shown in Figure 1 is the a conceptual mental model of the sociotechnical systems associated with algorithmic fairness that we introduce in this contribution. The model aims at conceptualizing the development stages of these systems in three layers (or spaces) of fairness conceptualization, fairness design, and fair algorithm operation. The proposed model of the sociotechnical system of algorithmic fairness captures also the footholds of these stages in both technical and social subsystems.

3. Related work

Algorithmic fairness is complex as it encompasses social and formal components as well as interactions between them and the environment. There is a lot of literature describing the complex landscape of algorithmic fairness from various perspectives. Although there are valuable contributions to various aspects of algorithmic fairness, there is still room to provide more in depth understanding about how these aspects influence each other and how the disciplines involved can inform each other about the characteristics (e.g., capabilities and limitations) of fairness concepts and notions in their fields. In brief, an overarching and unifying framework for algorithmic fairness that streamlines the different perspectives and disciplines is missing.

A lot of the literature emphasizes the need for adopting a cross-disciplinary approach, particularly from a sociotechnical perspective. The authors of (Altman et al., 2018) argue that for algorithmic fairness one should consider the foreseeable effects or harms that algorithmic design, implementation, and use have on the well-being of individuals. They propose a harm-reduction framework for algorithmic fairness, which aims at identifying and mitigating harms during all stages of data lifecycle. Based on a systematic literature review, Starke et al. (Starke et al., 2022) argue about the necessity of a human-centric approach that takes individuals' perceptions on fairness into account when designing and implementing algorithmic decision-making systems. They advocate for more interdisciplinary research that adopts a society-in-the-loop framework. Dolata et al. (Dolata et al., 2022) argue that since fairness is an inherently social concept, using algorithmic fairness technologies should be based on a sociotechnical approach. They aim at embedding algorithmic fairness in the sociotechnical view of ISs and call for undertaking a holistic approach to algorithmic fairness. We build on these ideas behind the sociotechnical approach for algorithmic fairness and aim at providing a systemic view on algorithmic fairness and an informative overview of the main information flows within that systemic view.

4. A typology of the information construct for algorithmic fairness

Realizing an effective sociotechnical approach to algorithmic fairness requires, among others, information harvesting and exchange across the formal and social disciplines involved. In addition to creating transparency, the basic construct of information for sociotechnical systems, as mentioned in Section 2.2, can impart meaning and utility to sociotechnical interactions, can ensure alignment of the social and formal subsystems, and thereby can play a major role in realizing the overall goals of the system. Providing appropriate information about, for example, system components, their dynamic interactions, and the emerging behaviors, can reduce randomness and uncertainty when realizing and operationalizing the system. According to (Beynon-Davies & Wang, 2019), reducing entropy (or in other words uncertainty) is a unifying conception of information that can be related to environmental structures, organization of signals, coding and decoding of signals, shared intentionality of actors, and coordination of joint actions.

Based on literature we distinguish several flows and types of the information construct that may arise due to the interactions between the components of the sociotechnical system of algorithmic fairness shown in Figure 1. The resulting information types and flows are shown in Figure 2, which are inspired by various papers that contribute to each or a subset of therein identified elements. Without intending to be exhaustive, we motivate these information flows/types by reviewing a couple of example works per flow/type shortly in the following.

4.1. Informing about social justice norms

As illustrated by link (a) in Figure 2, the social view informs the formal discourse about social justice and fairness norms. As a result, social view inspires formal view on (aspects of) fairness. The social definitions of fairness originate from various disciplines like philosophy, political science, legal science, criminology, sociology, anthropology, neuroscience, and psychology (Dolata et al., 2022). In (Binns, 2018) the author explains

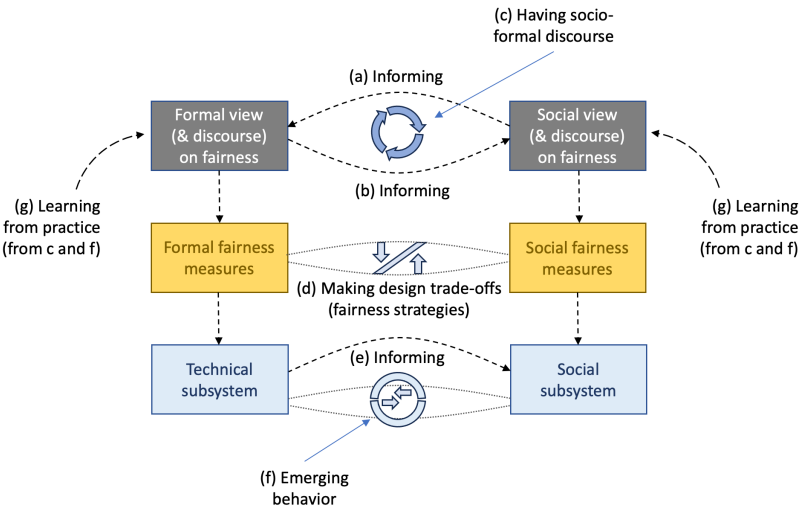

Fig. 2 – An illustration of the informational interactions between social and formal aspects of algorithmic fairness at three levels, which forms the information construct.

the terms fairness and non-discrimination as have been laid down by moral and political philosophers. Binns aims at providing a rigorous understanding about these concepts from political philosophy and elucidating the emerging debates about fair algorithms. This work can be regarded as a typical example of information link (a) shown in Figure 2.

It appears that there are inherent trade-offs between some formal metrics that quantify group and individual fairness. In (Dwork et al., 2012) the authors raise the concern that simple statistical parity metrics between protected groups could be intuitively unfair at the individual level. (In several examples they show the outcome is blatantly unfair for individuals, while statistical parity is maintained for the group the individuals belong.) The work marks another example for information link (a) in Figure 2.

4.2. Informing about algorithmic fairness measures

As illustrated by link (b) in Figure 2, the formal view uses data, simulation and formalism to quantitatively inform the social discourse about some aspects of social fairness (and justice) norms. Often inspired by social definitions of social biases, IS developers define and realize technical measures, i.e., the so-called algorithmic fairness solutions, to deal with the biases, especially when using algorithms in social domains. Existing technical definitions or measures of social bias are of four types: statistical-based, similarity-based, performance-based, and causal reasoning-based (Starke et al., 2022). In practice, there are often discrepancies between the fairness definition(s) as realized in the technical domain, i.e., those offered by algorithmic fairness solutions, and the fairness definition(s) as desired in non-technical domains. In addition to discrepancies between technical and non-technical definitions of fairness, there are discrepancies between the ways that algorithmic fairness solutions are viewed in the technical domain and in non-technical domains. Both sorts of discrepancies (i.e., those between fairness definitions and between views on algorithmic fairness) can contribute to unjustified discrimination of social groups and individuals, especially when applying the outcomes of decision-support systems into practice in social contexts (i.e., where humans, social groups and society are affected directly or indirectly).

To elaborate further on information flow (b), we explain a well-known inconsistency issue among three statistical measures, which is identified in the literature using formal and analytical methods. Specifically, the inconsistency is concerned with statistical measures, on the one hand, the Positive Predictive Value (PPV) and, on the other hand, the False Negative Ratio (FNR) and False Positive Ratio (FPR). Chouldechova (Chouldechova, 2017) gives an analysis of the relation between the PPV, FNR and FPR, assuming that the base rate for protected and unprotected groups differ. Base rate, also called prevalence or prior probability, is the percentage of a social group (like men or women) in a dataset used for training a ML model. Assuming that the base rates for the protected group and the unprotected group are different implies that it is impossible to make both PPV and FNR/FPR the same for both sensitive groups (Chouldechova, 2017). This inconsistency necessitates making trade-offs if and when applying them together into practice is required.

These achievability conditions are also derived in (Kleinberg et al., 2016) in a different and more generic

setting. Both results of (Chouldechova, 2017) and (Kleinberg et al., 2016) show the need for making trade-offs between false positive (idem, false negative) ratios vs predicted parity (or well-calibration in the generic case of Kleinberg et al., 2016). The arguments of (Chouldechova, 2017) and (Kleinberg et al., 2016) are important in investigating the famous COMPAS like cases, where the critique focused on the violation of the false positive (idem, false negative) ratios and the counterarguments established the satisfaction of predictive parity (and well-calibration) condition, see (Corbett-Davies et al., 2016; Flores et al., 2016).

It is important to realize that these results, which are derived with formal methods (and data analytics), hold for all decision-makings; including those made by human decision makers who carry out structured decision rules (Corbett-Davies et al., 2016).

4.3. Information derived from socio-formal discourse

As illustrated by link (c) in Figure 2, the exchange of information flows (a) and (b) paves the way for a socio-formal discourse, which leads to enriching the social and/or formal concepts of fairness.

In (Binns, 2020), the author reconsiders the conflict or inconsistency between group and individual fairness metrics as was already raised in (Dwork et al., 2012). This conflict exists in cases where two similar individuals who belong to different social groups (i.e., they are otherwise similar but differ in a protected characteristic) are assigned different outcomes in order to satisfy group fairness. In this case group fairness leads to individual unfairness. The authors of (Binns, 2020) argue that the conflict is a misconception based on theoretical discussions from formal and social discourses. As a result, the paper concludes that individual and group fairness are not fundamentally in conflict and that "the apparent conflict ... is more of an artefact of the blunt application of fairness measures, rather than a matter of conflicting principles". Similarly, in (Hellman, 2020) the author reconsiders the legal and philosophical rationale behind the fairness metrics that are shown to be in conflict in the case of the COMPAS tool. Both papers (Binns, 2020) and (Hellman, 2020) can be seen as a testimony of the information creation shown by link (c) in Figure 2.

4.4. Information about design trade-offs

The link (d) in Figure 2 represents the information (e.g., the design guidelines) needed for dealing with the inconsistencies that might exist among formal (and social) fairness metrics and/or measures. In the design space, this information is needed for making trade offs within and between contending social and formal fairness measures (i.e., among the partial solutions of the algorithmic-based fairness).

Both results of (Chouldechova, 2017) and (Kleinberg et al., 2016) show the need for making trade-offs between the contending statistical measures of PPV, FNR and FPR, as mentioned for link (b). The author in (Chouldechova, 2017) propose three strategies for making such trade-offs namely:

- "Allow unequal false negative rates to retain equal PPV's and achieve equal false positive rates,
- Allow unequal false positive rates to retain equal PPV's and achieve equal false negative rates,
- Allow unequal PPV's to achieve equal false positive and false negative rates".

Other options that in theory might work would be making perfect prediction (i.e., making FPR and FNR) zero (or negligible) for both protected and unprotected groups) or making the base rates for social groups equal. Although the last two options are impractical, but they could be seen as a justification of striving for perfect solutions (like eradication of social inequalities).

The trade-offs can not only be made between technical algorithmic measures as mentioned above, but also between technical and social and/or less technical ones. For the latter category one can think of trending solutions of algorithmic recourse (Karimi et al., 2022) and algorithmic contestability (Alfrink et al., 2023). Algorithmic recourse is concerned with providing explanations and recommendations to individuals who have received unfavorable outcomes from automatic decision-making systems. The explanations and recommendations should provide end-users with actionable measures whereby the outcome of an AI system can be changed to the favorable one. Making AI systems contestable by design is another way to mitigate these concerns. Via algorithmic contestability one aims at making AI systems responsive to human intervention throughout the system lifecycle. Such a human intervention can ask an AI system, for example, to explain and

interpret how outcomes are derived (including, from which input and training data). Unlike in algorithmic recourse, in algorithmic contestability the objective is not to change the algorithm outcome, but to change the outcome of the whole process from which the algorithm is part of (i.e., the term algorithmic-based fairness we introduced above). Answering these questions in both algorithmic recourse and algorithmic contestability can be facilitated by having data lineage in place (Bargh, 2024).

These works can be seen as examples for link (d) in Figure 2, representing the information (or guidelines) needed for making design trade offs among various formal metrics and social interventions.

4.5. Information about the applied fairness measures

As illustrated by link (e) in Figure 2, during system operation phase, there is a need for the AI/ML system to explain and inform social domain about the meaning and certainty of the outcomes of the technical subsystem. This information may help social domain practitioners to interpret the outcomes appropriately within the operation context.

Several works give overviews about the formal (i.e., technical) metrics and measures to algorithmic fairness from different perspectives, which can be characterized as examples of information flows and types denoted by link (e) - and to some degree of types (a) and (b) - in Figure 2. In (Verma & Rubin, 2018; Zafar et al., 2017) the authors present statistical fairness metrics according to three fairness notions of disparate treatment, disparate impact and disparate mistreatment. (Note that the scope of technical solutions covered in Verma & Rubin, 2018 is broader than statistical fairness metrics.) Carey and Wu (Carey & Wu, 2023) survey statistical fairness metrics and explain the philosophical views on fairness (like equality of opportunity and luck-egalitarian) that support or inspire those metrics.

Similarly, Verma and Rubin (Verma & Rubin, 2018) provide an overview of the most prominent formal metrics and measures of fairness which can be applicable for the algorithmic classification problem. Further, they elaborate on inconsistencies among these definitions. Balayn et al. (Balayn et al., 2021) and Mehrabi et al. (Mehrabi et al., 2021) survey the literature about bias and unfairness, the metrics and methods to identify these biases, and the measures to deal with these biases. While the former focuses on bias in training data (thus on fairness metrics, fairness identification, and unfairness mitigation methods in the field of data management), the latter considers these aspects also within algorithms and at the interaction point between end-users and algorithms.

4.6. Information about unforeseen behavior

As illustrated by link (f) in Figure 2, due to interactions of the technical and social subsystems in the operation phase, a new behavior may emerge that was not known beforehand (i.e., during the conceptualization and design phases). This emerging behavior should be accounted for and documented for further use.

Considering the feedback loops shown in Figure 1, context dependency, and the growing use of AI algorithms, algorithmic fairness sociotechnical systems can be perceived as complex systems. Scholars consider the notion of complex systems versus simple systems, although the boundary between them cannot be identified precisely (Manning & Ravi, 2013). In narrative terms, complex systems encompass, among others, contextuality, reflexivity, temporal sensitivity; allowing for reproduction, self-organization, and adaptation. In such systems, interaction among localized mechanisms within sybsystems may cause undesired emergent (mis)behaviors at a more aggregated scale (Manning & Ravi, 2013; Mogul, 2006).

When an algorithmic fairness solution is designed and implemented, there might arise new insights that were not foreseen beforehand during conceptualization or design phases. A testimonial example for this is the COMPAS tool. As mentioned in the previous sections, the tool was designed to be fair from the viewpoint of some formal fairness metrics (Dieterich et al., 2016) but, as shown by a follow up study (Larson et al., 2016), the tool was unfair according to another metric. This case can be seen as an example of link (f) shown in Figure 2.

4.7. Informing formal and social definitions

As illustrated by link (g) in Figure 2, the emerging behavior can be used for enriching social and formal views and their discourse. Current research communities are predominantly reactive in addressing the emerging "issues" from the growing use of algorithms in society. Therefore, there is an urgent need for the unity of intellectual frameworks beyond disciplinary perspectives and research practices, to offer holistic thought leadership in this space (Engin et al., 2024). A key aspect of this holistic approach is to learn from and adopt to the emerging (mis)behaviors from the practice. To this end, a realistic (impact) evaluation of such systems, which is lacking nowadays, is needed before, during and after their operation (Leeuw, 2025). The outcomes of such learning can be used to enrich the conceptualization and design of algorithmic fairness.

One of the concepts close to fairness that has changed during the time is privacy (Bargh, 2019). As the early principled discussion of privacy, Aristotle made distinction between public and private spheres of life (Nissim & Wood, 2018). Afterwards, there have been many definitions for privacy introduced, particularly within legal regimes like the right to be let alone by Warren and Brandeis in 1890, limited access to the self by Godkin in 1880, and control over personal information by Westin (Westin, 1968). An interesting point within the evolution of privacy concept is the role or rise of certain technological developments. The definition of Warren and Brandeis, i.e., the right to be let alone, brought up as a reaction to Kodak's new snap cameras and widespread newspaper circulation in 1890's. Later on, we notice ICT developments which made it possible to transfer data between remote servers (like development of the DARPANET in the 1970's), coincides with Westin's definition in 1968.

5. Conclusion

Despite the success that data-driven decision making systems offer, there are still many obstacles in their large-scale and real-life deployments like mitigating the security issues of largely distributed data-reliant systems, managing the quality issues of loosely coupled data sources, and dealing with social biases and wrongful discrimination of individuals and groups (i.e., unfairness) as embedded in collected data or induced due to inattentive data processing. Not handling these soft challenges appropriately and adequately would harm individuals, groups, and society, adversely affecting basic human rights like privacy, liberty, autonomy, and dignity.

Inspired by non-technical definitions of social biases and discrimination, IS developers define and realize algorithmic fairness solutions to deal with these issues, especially when using AI/ML based algorithms. In practice, there are discrepancies between the ways that algorithmic fairness solutions are viewed in formal and social domains. Such discrepancies can contribute to an inefficient system design, which may lead to to unjustified biases and discrimination of social groups and individuals. Such a practice becomes highly risky, especially when applying the outcomes of decision-support systems in social contexts (i.e., where humans, social groups and society are affected directly or indirectly).

In this contribution we focused on the issue of algorithmic fairness, which, in itself, can be framed as a sociotechnical system with interacting social and formal components. The information construct is considered as a key element of sociotechnical systems. It is concerned with creating and exchanging information in a well-formed, meaningful, and order-giving way. As part of the information construct, we presented a conceptual (mental) model of the sociotechnical systems associated with algorithmic fairness. The model conceptualizes the development stages of these systems for both formal and social views. As such, it can serve as a mental model for both formal and social stakeholders. Having such a common view on the complex algorithmic fairness sociotechnical systems can facilitate information exchange among the multi-disciplinary stakeholders involved and thus help bridging the gap between the formal and social views. In addition, from the literature, we noted that providing information eases the opacity of the components of the algorithmic fairness sociotechnical system. As another contribution, we identified 7 types of the information types and flows, which can collectively constitute the information construct of the sociotechnical system of algorithmic fairness.

In our future research, we will elaborate on the existing inconsistencies among some statistical fairness metrics from the literature, which ask for making tradeoffs among contending values. Further, we will sketch and discuss several strategies to deal with some aspects of algorithmic fairness (like making trade-offs). Despite being developed within or inspired by the formal perspective, we foresee that these strategies can inform the design of integrated formal and social solutions that aim at adhering to fairness principles and ideals.

Contributor Statement

The authors confirm sole responsibility for study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. The first author's roles are Writing – original draft, Visualization, Investigation, Conceptualization. The second and third authors' roles are Writing – review & editing, Conceptualization, Investigation.

Use of AI

There is no use of AI in this work.

Conflict Of Interest (COI)

There is no conflict of interest.

References

Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2023). Contestable ai by design: Towards a framework. *Minds and Machines*, *33*(4), 613–639.

Altman, M., Wood, A., & Vayena, E. (2018). A harm-reduction framework for algorithmic fairness. *IEEE Security & Privacy*, *16*(3), 34–45.

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

Bargh, M. S. (2019). *Realizing secure and privacy-protecting information systems: Bridging the gaps*. Hogeschool Rotterdam.

Bargh, M. S. (2024). Data lineage for the justice system: Scope, potentials, and directions. http://hdl.handle.net/20.500.12832/3443

Beynon-Davies, P., & Wang, Y. (2019). Deconstructing information sharing. *Journal of the association for information systems*, *20*(4), 1.

Binns, R. (2018). What can political philosophy teach us about algorithmic fairness? *IEEE Security & Privacy*, *16*(3), 73–80.

Binns, R. (2020). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 514–524.

Carey, A. N., & Wu, X. (2023). The statistical fairness field guide: Perspectives from social and formal sciences. *AI and Ethics*, *3*(1), 1–23.

Chatterjee, S., Sarker, S., Lee, M. J., Xiao, X., & Elbanna, A. (2021). A possible conceptualization of the information systems (is) artifact: A general systems theory perspective 1. *Information Systems Journal*, *31*(4), 550–578.

Choenni, S., Netten, N., Bargh, M. S., & van den Braak, S. (2021). Exploiting big data for smart government: Facing the challenges. In J. C. Augusto (Ed.), *Handbook of smart cities* (pp. 1035–1057). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-69698-6_82.

Choenni, S., Netten, N., Shoae-Bargh, M., & Choenni, R. (2018). On the usability of big (social) data. *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 1167–1174. DOI: https://doi.org/10.1109/BDCloud.2018.00172.

Choraś, M., Pawlicki, M., Puchalski, D., & Kozik, R. (2020). Machine learning–the results are not the only thing that matters! what about security, explainability and fairness? *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20*, 615–628.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153–163. DOI: https://doi.org/10.1089/big.2016.0047.

Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *Washington Post*, *17*.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, *7*(4), 1–36.

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, *32*(4), 754–818.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Engin, Z., Gardner, E., Hyde, A., Verhulst, S., & Crowcroft, J. (2024). Unleashing collective intelligence for public decision-making: The data for policy community. *Data & Policy*, *6*, e23.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, *80*, 38.

Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review*, *106*(4), 811–866.

Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2022). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, *55*(5), 1–29.

Kleinberg, J. M., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *Information Technology Convergence and Services*. https://api.semanticscholar.org/CorpusID:12845273

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the compas recidivism algorithm [Retrieved on 14 October 2024 from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm].

Lee, A. S. (2004). Thinking about social theory and philosophy for information systems. *Social theory and philosophy for information systems*, *1*, 26.

Leeuw, F. L. (2025). The algorithmization of policy and society: The need for a realist evaluation approach. In *Artificial intelligence and evaluation* (pp. 242–265). Routledge.

Manning, P., & Ravi, S. (2013). Cross-disciplinary theory in construction of a world-historical archive. *Journal of World-Historical Information*, *1*(1), 15–39.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1–35.

Misuraca, G., & van Noordt, C. (2020). Overview of the use and impact of ai in public services in the eu. *Publications Office of the European Union: Luxembourg*.

Mogul, J. C. (2006). Emergent (mis) behavior vs. complex software systems. *ACM SIGOPS Operating Systems Review*, *40*(4), 293–304.

Netten, N., Bargh, M. S., & Choenni, S. (2018). Exploiting data analytics for social services: On searching for profiles of unlawful use of social benefits. *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*, 550–559.

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

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & management*, *52*(2), 183–199.

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the is discipline: Its historical legacy and its continued relevance. *MIS quarterly*, *43*(3), 695–720.

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, *9*(2). DOI: https://doi.org/10.1177/20539517221115189.

Toli, A. M., & Murtagh, N. (2020). The concept of sustainability in smart city definitions. *Frontiers in Built Environment*, *6*, 77. DOI: https://doi.org/10.3389/fbuil.2020.00077.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the international workshop on software fairness*, 1–7.

Westin, A. F. (1968). Privacy and freedom. *Washington and Lee Law Review*, *25*(1), 166.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th international conference on world wide web*, 1171–1180.