Necessity of Processing Sensitive Data for Bias Detection and Monitoring: A Techno-Legal Exploration

Ioanna Papageorgiou[*]
Leibniz University Hannover,
Germany
ioanna.papageorgiou@iri.uni-hannover.de

Carlos Mougan
University of Southampton, UK
c.mougan@soton.ac.uk

Abstract

This paper explores the intersection of the upcoming AI Regulation and fair ML research, specifically examining the legal principle of ”necessity” in the context of processing sensitive personal data for bias detection and monitoring in AI systems. Drawing upon Article 10 (5) of the AI Act, currently under negotiation, and the General Data Protection Regulation, we investigate the challenges posed by the nuanced concept of ”necessity” in enabling AI providers to process sensitive personal data for bias detection and bias monitoring. The lack of guidance regarding this binding textual requirement creates significant legal uncertainty for all parties involved and risks a purposeful and inconsistent legal application. To address this issue from a techno-legal perspective, we delve into the core of the necessity principle and map it to current approaches in fair machine learning. Our objective is to bridge operational gaps between the forthcoming AI Act and the evolving field of fair ML and support an integrative approach of non-discrimination and data protection desiderata in the conception of fair ML, thereby facilitating regulatory compliance.

1 Introduction

In response to societal concerns about the discriminatory potential of AI, researchers and practitioners have proposed new methods for detecting, mitigating and monitoring bias (Ntoutsi et al., 2020; Barocas et al., 2019; Mitchell et al., 2021), capturing the increasing attention of policymakers (Balayan and Gürses, 2021). These methods often imply the processing of (sensitive) personal data (Zliobaite and Custers, 2016), which is subject to extensive protection under the European Data Protection Regulation (GDPR, 2016)[2], creating a tension between the pursuit of fair ML and the EU Data Protection Law.

Against this background, the AIAct (2021), currently under negotiation, enables in Article 10 (5) the processing of sensitive personal data for the purpose of bias detection, correction and monitoring in

[*] First author.

[2] The GDPR protects all types of personal data, which include any information relating to an identified or identifiable individual (cf. the definition of personal data Article 4 (1) (GDPR, 2016)). Certain types of personal data are designated as “special category” data or “sensitive data” and granted a heightened level of protection (cf. definition of sensitive personal data Article 9 (1) (GDPR, 2016)). Sensitive data under EU data protection law partly overlap with the protected grounds under EU non-discrimination law. Gender and age though are in principle not considered sensitive but ”regular” personal data. Cf. Calvi (2023); Van Bekkum and Borgesius (2023).

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

relation to high-risk AI systems. However, to prevent excessive interference with the fundamental right to data protection, the AI Act ties this possibility to various textual requirements, notably to the condition of "necessity". Precisely, according to the Article 10 (5) AI Act, the processing of personal sensitive data is permitted:

"*. . . to the extent that it is **strictly necessary** for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems..*"

While the principle of necessity is well-established within EU law (Supervisor, 2017), its operationalization in various data processing contexts and particularly in the domain of fair ML lacks definition and clear guidance. Given that the application of "necessity" is inherently context-specific and involves a degree of discretion in decision-making, the absence of corresponding guidance gives rise to significant legal uncertainty for all parties involved and poses a significant risk of purposeful and inconsistent legal interpretation. As the necessity requirement is intrinsically tied to the legitimacy of data processing, a wrongful interpretation and application may not only compromise the fundamental right to data protection for data subjects but also expose AI providers[3] to serious consequences, ranging from reputational damage to severe financial penalties[4].

Our contribution aims to assist the lawful application of the necessity principle in the context of processing sensitive data for bias detection and monitoring in AI systems. Precisely, the paper provides the following contributions by order:

- We establish the technical and normative foundation for processing sensitive data in bias detection and monitoring, clarifying these two concepts and outlining the relevant legislative framework shaped by the AI Act.
- Drawing from the General Data Protection Regulation (GDPR) and Jurisprudence of the Court of Justice of the European Union, we examine the legal concept of necessity, shedding light on its core elements. As "necessity" is a recurring condition in nearly all legal bases for processing personal data in the GDPR[5], the relevance of our analysis extends beyond the Article 10 (5) of the AI Act and sensitive data to include a variety of potential uses of personal data for the purposes of bias detection and monitoring.
- We discuss the application of the necessity principle in the context of processing sensitive data for bias detection and bias monitoring through a techno-legal lens, identifying associated operational challenges.
- We assess existing fairness approaches in machine learning, such as individual and group fairness, in light of the core elements of the necessity principle. We thereby offer a conceptual mapping of data protection requirements to computational fairness metrics, broadening existing legal taxonomies (Binns et al., 2023; Wachter et al., 2021b; Xiang, 2021) and introducing new research challenges in the field of fair ML.

2 Bias Detection and Bias Monitoring in AI systems

2.1 Computational Fairness Metrics

Aiming to measure bias in AI systems, various definitions of fairness in machine learning have been proposed (see Mehrabi et al. (2021); Finocchiaro et al. (2021); Barocas et al. (2019) for recent overviews). They can be categorized into two central notions: individual and group fairness. The concept of *individual fairness* revolves around treating similar individuals similarly. In essence, it aims to ensure that "any two individuals who are similar for a particular task should be classified similarly" Dwork et al. (2012). Similarity-based metrics such as Lipschitz condition (Dwork et al., 2012) or similarity graphs between individuals (Petersen et al., 2021) calculate this notion on different scenarios.

In contrast, *group fairness* seeks to establish some form of parity between groups of individuals based on protected attributes, such as gender and race. Various metrics have been pro-

[3] According to Article 3 (1) AI Act, AI providers are actors who develop AI systems or have an AI system developed with a view to placing it on the market or putting them into service under their own name or trademark, whether for payment or free of charge.

[4] cf. Article 83-84 GDPR (2016) and Article 71 of the AIAct (2021).

[5] See Article 6 (1) (b), (c), (d), (e), (f) GDPR.

posed in the literature to measure parity between protected groups. These include i.a. statistical parity (Corbett-Davies et al., 2017), equal opportunity (Hardt et al., 2016), equal treatment (Candela et al., 2023), and equal misclassification rates (Zafar et al., 2017).

To support our subsequent discussions, we outline below the notation we use to define a subset of metrics which will be central to our argumentation.

Let X and Y be random variables taking values in $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$, respectively. A *predictor* is a function $f\colon \mathcal{X} \to \mathcal{Y}$, also called a model. For conceptual simplicity, the domain of the target feature is $\operatorname{dom}(Y) = \{0, 1\}$ (binary classification). We assume a binary feature modelling protected social groups denoted by Z, called *protected feature*, that is not used to train the model, known as fairness through unawareness (Pedreschi et al., 2008).

Definition 2.1. *(Individual Fairness) (IF)* We say that a model achieves individual fairness w.r.t. two individual samples if $d(f(x), f(x')) \leq \mathcal{L} \cdot d(x, x') \forall x, x' \in X$ where $\mathcal{L} > 0$ is a constant and $d(\cdot, \cdot)$ is a distance function.

Definition 2.2. *(Demographic Parity (DP)).* A model f achieves demographic parity if $f(X) \perp Z$. We can derive an unfairness metric as $d(P(f(X)|Z = z), P(f(X))$, where $d(\cdot)$ is a distance between probability distributions.

Definition 2.3. *(Equal Opportunity (EO))* A model f achieves equal opportunity if $\forall z.\, P(f(X)|Y = 1, Z = z) = P(f(X) = 1|Y = 1)$.

2.2 Detection vs Monitoring: Technical Disentanglement

We disentangle the concepts of bias detection and monitoring to provide the technical foundation for the subsequent presentation of the normative framework and acknowledge the significance and nuances of addressing biases at different stages of the AI system's lifecycle.

Bias detection, also known as bias *bias evaluation*, involves measuring discrimination and fairness using held-out in-distribution and labeled data in a static point of time (Suresh and Guttag, 2021). It aims to assess the presence and extent of biases in the model during the development and training phase. Detecting bias in AI systems is a topic that has received much attention in recent years with surveys and taxonomies developed (Ntoutsi et al., 2020; Barocas et al., 2019). By analyzing the training data, practitioners and researchers can identify potential sources of bias, understand its impact on the model's predictions, and take steps to mitigate it before deployment.

On the other hand, *bias monitoring* focuses on evaluating the same metrics of discrimination and fairness once the model has been deployed and makes predictions on new data batches, making it an iterative continuous process (Diethe et al., 2018; Paleyes et al., 2023). The monitoring phase is crucial because biases can emerge or evolve in real-world applications due to various factors, such as changes in the data distribution or external factors influencing the model's performance (Huyen, 2022; Burkov, 2020). Bias monitoring techniques are employed to continuously assess the model's behaviour and performance in different contexts, identify any unintended biases that may have surfaced after deployment, and take corrective actions if necessary (Burkov, 2020; Huyen, 2022; Mougan and Nielsen, 2023; Bhatt et al., 2020).

Different research approaches have aimed to monitor fairness under distribution shift, when training and deployment distributions differ (Quiñonero-Candela et al., 2009), for example (Rezaei et al., 2021) explore fairness under covariate shift.

2.3 Bias Detection and Bias Monitoring in the AI Act: Trilogue Insights

In April 2021, the European Commission submitted a Proposal for a Regulation laying down harmonised rules on artificial intelligence (AIAct, 2021)[6], which seeks, i.a. to ensure a high level of fundamental rights protection in the context of AI development and use. The objective of minimising

[6] For readers not familiar with EU legislative procedures, the ordinary legislative procedure (Article 294 (2) TFEU (2007)) starts with the submission by the Commission of a proposal for a legislative act to the Parliament and the Council. Based on their respective positions on the proposal, the Parliament and the Council enter into negotiations (*Trilogues*) facilitated by the Commission. The AI Act is currently at this stage of interinstitutional negotiation. Only when the Parliament and the Council reach an agreement on the precise text is the AI Act finally adopted (Bauerschmidt, 2021).

the risk of algorithmic discrimination and safeguarding the fundamental right to non-discrimination is particularly emphasized in the explanatory memorandum and the recitals.[7] However, Article 10 is the only one in the operative text that explicitly and directly deals with "bias". The fifth paragraph of the Article is in turn the sole provision that explicitly refers to the notions of *bias detection* and *bias monitoring*, by allowing the processing of sensitive personal data:" *. . . to the extent that it is strictly necessary for ensuring bias monitoring, detection and correction in relation to the high-risk AI systems. . .* " Through the adoption of this provision, the AI Act provided on the grounds of substantial public interest an exception to the GDPR's principal prohibition to process sensitive data [8].

In December 2022, the European Council adopted its general approach to the Artificial Intelligence Act (European Council, 2022), without proposing any amendments to Article 10 (5) of the AI Act, preserving the inclusion of both bias detection and monitoring within its scope.

In June 2023, the European Parliament concluded the first reading of the AI Act and adopted its negotiating position (Parliament, 2023). Among various revisions, the Parliament has notably amended Article 10 (5) AI Act (amendment Nr. 290).[9] In addition to enhancing the conditions for its application, the Parliament's negotiating position omits the notion of *bias monitoring* from the provision's material scope, thereby permitting the processing of sensitive personal data only "*to the extent that it is strictly necessary for the purposes of ensuring negative bias detection and correction in relation to the high-risk AI systems.*" While this amendment has not yet captured significant attention in the public discourse surrounding the AI Act, the preceding technical section has demonstrated that it signifies more than a mere shift in nomenclature. On the contrary, whether or not the concept of bias monitoring is included within Article 10 (5) of the AI Act is deemed to have a substantial effect on the material and personal scope of the provision and on the course of fair ML per se. Examining the motivations behind this legislative decision and its implications in terms of data protection exceeds the scope of this paper. Given that both approaches are currently under negotiation, our analysis will address the necessity principle within the contexts of both bias detection and bias monitoring.

3 Processing Sensitive Personal Data under the Principle of Necessity

3.1 The Indefinite Legal Concept of "Necessity"

The AI Act enables in Article 10 (5) the processing of sensitive personal data for the purposes of bias monitoring and detection on the grounds of public interest. Since this type of data is extensively protected by the General Data Protection Regulation and in order to prevent a disproportionate interference with the right to data protection, this possibility is tied to specific legal requirements, notably the requirement of necessity. Precisely, AI providers are permitted to process sensitive personal data, such as ethnic origin, only if and to the extent that this is *strictly necessary* for the purposes of bias monitoring or detection.

Generally, *necessity* stands out as an overarching principle within the GDPR (Kühling and Buchner, 2020), taking the form of a data quality principle[10] and a recurrent condition in nearly all requirements governing the lawfulness of processing personal data[11]. The principle of necessity essentially dictates that data processing is permissible only to the extent required for achieving the intended purpose. Its origins can be traced to Article 52 of the Charter of Fundamental Rights of the European Union (2000), where "necessity" is enshrined as a fundamental principle governing any limitation on fundamental human rights, including the right to data protection (Supervisor, 2017; Simitis et al., 2019).

[7] cf. explanatory memorandum 1.2, 3.5; Recitals 15, 17, 28, 35 - 39, 44 AIAct (2021).

[8] Despite the fundamental prohibition on processing sensitive data enshrined in Article 9 (1) of the GDPR, the second paragraph of the same Article outlines strict exceptions under which the data processing is legitimized. Among these exceptions, sensitive data processing is permitted "*as long as this is necessary for reasons of substantial public interest, on the basis of Union or Member State law*" (cf. Article 9 (2) (g) GDPR). Article 10 (5) of the AI Act aims to serve as the corresponding Union law, incorporating the purpose of bias monitoring, detection, and correction in the "public interest" and providing on this ground a new exemption from the prohibition of processing sensitive personal data.

[9] For an illustrative comparative view of all three EU institutional positions see European Parliament (2023).

[10] cf. Supervisor (2017) and Article 5 (1) (c) GDPR.

[11] Supra note 5.

Despite its pivotal role in both primary and secondary EU law, neither of these sources provides a precise definition of the principle's scope. As a result, it emerges as an *indefinite* legal concept (Gola and Heckmann, 2022), which must be filled in and fleshed out in each individual case.

An examination of relevant provisions of the General Data Protection Regulation, legal literature and case law of the Court of Justice of the European Union (CJEU) provides some interpretational guidance and unveils tangible elements of the necessity requirement. Precisely, according to CJEU jurisprudence (Huber v Bundesrepublik Deutschland, 2006), *necessity* is a concept which has its independent meaning in community law and must be interpreted in a manner that fully reflects the objective of the data protection regulation. ECJ Jurisprudence such as TK v Asociaţia de Proprietari bloc M5A-ScaraA (2018) and Meta Platforms and Others (Conditions générales d'utilisation d'un réseau social) (2023) highlight that necessity should be examined in conjunction with the *data minimisation principle*[12], a fundamental tenet of EU data protection law and a direct expression of necessity in the body of the GDPR (2016). This principle limits processing to data that is adequate, relevant, and not excessive in relation to the intended purpose and together with the elements examined below provides the normative reference for applying necessity in a given context.

3.2 Necessity's Core Elements

a. Suitability To establish the necessity of data processing, it must first and foremost be objectively suitable to enable or facilitate the intended purpose in the specific case at hand (Schantz and Wolff, 2017). This first element of necessity, described as "suitability"[13], represents a relatively value-neutral component, assessing whether the means that imply the processing of personal data are well-chosen at a conceptual or empirical level. Specifically, it requires a direct connection between the processing of (sensitive) personal data and the pursued objective, with mere convenience or cost-effectiveness falling short of satisfying the criterion [14]. In terms of data minimisation, this element maps to the data quality standard of "adequacy" and "relevance". Accordingly, the processing should be limited to data appropriate for fulfilling the objective at hand in terms of their function, content and scope (Simitis et al., 2019).

b. Requirement of Alternatives The principle of necessity extends beyond assessing whether processing sensitive data is suitable for the intended purpose. It also imposes that no alternative option is available which is both *(i)* **less intrusive** and *(ii)* **similarly effective** in achieving the desired objective. This aspect, known as ”requirement of alternatives (Schantz and Wolff, 2017)” - , is found in Recital 39 of the GDPR:

"*. . . personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means*".

In this second step of the necessity assessment, data controllers[15] are required to assess whether the intended purpose could be attained equally effectively with alternative means which are more "data-protection friendly". Being more data-protection friendly generally implies that the data protection rights of data subjects are restricted to a lesser extent (Gola and Heckmann, 2022). Key criteria for evaluating the severity of the restriction or the interference with data protection rights include the type of data - sensitive or non-sensitive - , the volume of data and the number of people affected, the risk of data misuse (Schantz and Wolff, 2017) and the duration of data processing (Gola and Heckmann, 2022) . In terms of data minimisation, this element maps to the data standard of ”non-excessiveness”, which requires that the amount of data processed and the duration for which data is stored is kept to a necessary minimum (Kuner et al., 2019).

The required assessment of alternatives is conditioned by a third element, this of *(iii)* **reasonableness** for the actor responsible for the data processing. To elaborate, according to the prevailing opinion, necessity does not imply that data processing must be absolutely indispensable

[12] Article 5 (1) (c) GDPR.

[13] This element can be found also as "effectiveness"(Huber v Bundesrepublik Deutschland, 2006). Terminological variations in EU legal scholarship or jurisprudence and potential nuances are beyond the scope of this paper.

[14] Cf. Art29WP (2012); European Data Protection Supervisor (2010).

[15] Data controllers are according to Article 4 GDPR the actors which, alone or jointly with others, determine the purposes and means of the processing of personal data.

due to technical or other reasons for achieving the purpose at hand (Schantz and Wolff, 2017; Information Commissioner's Office, 1 11) [16], nor does it suggest that the purpose becomes unattainable in its absence. It "solely" means that there is not an effective milder alternative available, which is reasonable in terms of personal, operational or financial feasibility (Gola and Heckmann, 2022). In particular, nothing *illegal* or *practically impossible* can be demanded. The question of which alternatives are to be considered reasonable or effective vis-à-vis their intrusiveness is not always straightforward and often imply value-laden judgements, since different methods may carry distinct sets of advantages and disadvantages.

4 Applying Necessity in Bias Detection and Monitoring: A Techno-Legal Discussion

AI providers have a margin of discretion on how to design their bias detection and monitoring processes. However, if they wish to collect, store, or in another way process sensitive personal data for those purposes, they must demonstrate that the procedures in place comply with legal requirements such as the requirement of necessity stipulated in Article 10 (5) AI Act[17].

The mere fact that sensitive data might be computationally useful or even essential for the application of a specific bias detection or monitoring method is *per se* not sufficient to satisfy the necessity requirement. To comply with necessity in a legal sense, AI providers must also evaluate the *(a)* **suitability** of the data processing in the context of the employed method and, importantly, *(b)* the existence of available **alternatives** concerning the factors of *(i) intrusiveness*, *(ii) effectiveness*, and *(iii) reasonableness*.

Crucially, by requiring *strict* necessity in Article 10 (5), the AI Act sets a higher necessity threshold for the case of bias monitoring and detection compared to the one foreseen by the GDPR for cases of public interest[18]. This may have implications on the approach and the rigour of the necessity assessment in the context of fair ML. It may either require an absolute imperative necessity when processing sensitive data for bias detection and monitoring - in a sense of a ”conditio sine qua non” - or a less demanding standard when considering alternative options. Satisfying necessity in the latter case would imply prioritizing less intrusive methods, even if they are not ”equally” but moderately effective for the purpose of bias detection or monitoring.

While acknowledging that applying necessity is a highly context-dependent task that relies on the specificities of the case at hand rather than abstract situations, we will now conceptually explore some of the aspects and challenges of its operationalization in the context of state-of-the-art bias detection and monitoring approaches.

4.1 Structural Application Challenges

The necessity of processing sensitive data is not assessed in a vacuum but against the purposes defined as legitimate by the EU legislator in Article 10 (5) AI Act:

"*the purposes of* ***ensuring*** *bias monitoring, detection [. . .] in relation to high-risk AI systems.*"

This purpose definition already sets a high threshold for meeting the necessity requirement. Despite the abundance of fairness metrics and methods proposed in the ML realm, there is limited practical evidence for bias detection and monitoring techniques in real-world scenarios. As a result, their applicability and adaptability in different contexts remains uncertain (Lee and Singh, 2021; Balayan and Gürses, 2021), challenging the feasibility of providing an ex-ante "guarantee" for detecting or monitoring bias in a specific industrial setting.

Moreover, evaluating necessity of data processing in the context of fair ML requires knowledge of the availability, efficacy and trade-offs of bias detection and monitoring methods. This task is further challenged by the growing availability of fair ML research and fairness toolkits, demanding

[16] For an opposite view Gola and Heckmann (2022).

[17] The statement is based on the premise that the AI Act comes into force in one of the currently negotiated forms. The version of the European Parliament (Parliament, 2023) states explicitly that AI Providers shall draw up documentation explaining why the processing of special categories of personal data was necessary to detect biases.

[18] See Article 9 (g) GDPR.

a continuous and close examination of the field's state of the art from the AI providers. The challenge becomes even more pronounced when dealing with interpretative legal notions, such as the principle of necessity, which are inherently vague and context-dependent, lacking a mathematical formulation.

4.2 Group Fairness Metrics

Suitability. Group fairness metrics aim to measure statistical disparities across groups and directly rely on access to the protected attributes of the individuals, which represent their group membership. Those protected attributes correspond to the protected characteristics under EU non-discrimination law and most of the sensitive data EU under data protection law[19]. Therefore, in the case of group fairness metrics, there is a *de facto* direct link between the processing of sensitive data and bias detection and monitoring.

Requirement of Alternatives We distinguish between two types of metrics: *(a)* metrics that are independent of ground truth data, such as demographic parity (definition 2.2) and *(b)* metrics that rely on ground truth data to be computed, such as equal opportunity (definition 2.3).[20]

To illustrate, consider a lending application: even when sensitive data (e.g. ethnic origin) of the prospective borrowers becomes available[21], the calculation of Equal Opportunity (as proposed by Hardt et al. (2016) and defined in Section 2.2) requires access to ground truth data, which includes information on default events or repayments among successful applicants. In contrast, fairness metrics that rely on the model predictions (Verma and Rubin, 2018), such as demographic parity (as proposed by Barocas and Selbst (2016); Zafar et al. (2017) and defined in Section 2.2) do not require ground truth data and can directly be measured. We now discuss the requirements of intrusiveness, effectiveness and reasonableness among group fairness metrics.

Intrusiveness: We argue that metrics that rely on ground truth data are more *intrusive* than those that rely exclusively on the model prediction. In the given example, computing equal opportunity would occur after the closure of the loan cycle, which, in turn, would imply a prolonged duration of data storage and, thus, a more intrusive alternative compared to demographic parity under the criterion of ”data processing duration”. [22] Furthermore, since ground truth labels, e.g. the fact whether an individual paid off or defaulted a loan, constitute personal data protected by the GDPR [23], computing metrics that rely on ground truth data requires a bigger amount of personal data for bias monitoring and thereby increase their general intrusiveness under the criterion of “data volume”.

Effectiveness: Besides being less intrusive, metrics that rely solely on model predictions, such as demographic parity, are deemed effective in capturing bias that may result in discrimination prohibited by European Union law. This type of metric belongs to the category of “bias transforming” metrics, as defined by Wachter et al. (2021a), aligning with some of the tests used by the European Court of Justice and Member State courts to measure indirect discrimination and the aims of EU non-discrimination law (Wachter et al., 2021a,b).

Reasonableness: Under some situations, prioritizing demographic parity for bias detection and bias monitoring may not be a reasonable alternative. For example, forcing to achieve demographic parity in the examined lending scenario may imply that we accept qualified applicants in the demographic $Z = 0$, but unqualified individuals in $Z = 1$, as long as the predicted probability matches (Hardt et al., 2016). However, granting loans to individuals with a potential lack of creditworthiness might be considered financially irresponsible or even illegal, thus failing the reasonability criterion. For example, according to German law[24] the lender must assess the creditworthiness of the borrower before concluding a consumer credit agreement and may only agree if the assessment shows that it is likely or there are no significant doubts that the borrower will comply with their credit obligations.

[19] Supra note 2.

[20] See Verma and Rubin (2018) for a systematic review of the fairness metrics of both categories.

[21] For consumer credit, the Article 18 (2) of the EU Directive of 18 October 2023 on credit agreements for consumers (The European Parliament and the Council of the European Union, 2023) prohibits the use of sensitive data for assessing the consumer's creditworthiness. While the interaction of this provision with Article 10 (5) AI Act remains uncertain, the given example serves merely as an illustration.

[22] For the key criteria for evaluating the severity of interference see Section 3.2.

[23] See Article 4 (1) GDPR.

[24] See Article 505a, German Civil Code BGB (2023) and §18a German Banking Act KWG (2023).

4.3 Individual Fairness Metrics[25]

Definitions of individual fairness, as exemplified by 2.1, which do not necessitate the use of protected attributes, render prima facie the use of sensitive data **unsuitable** for their computation. Similarly at the level of **alternatives** and from a purely technical standpoint, they appear to be *less intrusive* than group fairness metrics since they eliminate the requirement for processing both sensitive and ground truth data.

However, when framed in this manner, individual fairness notions seem to lack relevance and *effectiveness* in capturing bias that infringes upon non-discrimination rights, as established within the EU legal framework for non-discrimination. Since a "protected characteristic" is a constitutional element in defining discrimination within EU non-discrimination law [26], individual fairness metrics of this kind become irrelevant within the scope of Article 10 (5) of the AI Act which maintains consistency with existing secondary Union legislation on non-discrimination[27].

To address related issues pertaining to individual fairness, researchers have suggested individual fairness metrics which incorporate the protected attribute in the definition of the similarity metric (Castelnovo et al., 2022; Fleisher, 2021; Yurochkin et al., 2020). By doing so, individual fairness metrics are no longer *less intrusive* than group fairness metrics with respect to accessing sensitive data.

Under both scenarios, it is questionable whether notions of individual fairness can be deemed equally *effective* or operationally *reasonable* alternatives for bias monitoring or detection. First of all, defining similarity between individuals and determining the appropriate mathematical similarity metric becomes a complex task. This complexity extends beyond mathematical formalization and involves considerations of social values and normative assumptions. Moreover, individual fairness approaches may fall short of identifying algorithmic biases that manifest only at the group level, which are at the core of indirect discrimination under EU non-discrimination law. Critiques of individual fairness notions along these lines, such as those by (Fleisher, 2021; Xiang, 2021), question the *effectiveness* and *reasonableness* of processing sensitive data for the application of individual fairness in the context of bias detection and monitoring.

5 Conclusions

In an effort to facilitate compliance with the forthcoming AI Act and the General Data Protection Regulation as well as to bridge operational gaps between these regulations and the field of fair ML, this paper explored the application of the necessity principle in the context of bias detection and bias monitoring in AI systems.

We first outlined the technical nuances of bias detection and bias monitoring and the corresponding legislative framework as is being shaped by the AI Act. This provided the operational and normative context of the analysis.

Subsequently, we drew on pertinent legal provisions and Jurisprudence of the Court of Justice of the EU and shed light to the core elements of the necessity principle. We suggested that legal necessity substantially differs from technical necessity in both nature and scope and we established the normative reference for applying necessity in the context of bias detection and monitoring.

Finally, we discussed the challenges of operationalising the necessity principle in the context of bias detection and monitoring and we examined computational fairness metrics in light of its core elements. By assessing fairness metrics along the axes of *intrusiveness*, *effectiveness* and *reasonableness*, we demonstrated that the choice between different fairness metrics carries distinct data protection implications. We introduced thus new research questions in the field of legal fair ML that extend beyond an adversarial conceptualization of "fairness" versus "privacy", thereby supporting

[25] Given the partial overlap of the necessity elements in the context of individual fairness, the *Suitability* and the *Requirement of Alternatives* are in this section addressed collectively to support readability and coherence.

[26] Cf. also Weerts et al. (2023).

[27] See Explanatory memorandum 1.2., Recital 44 AI Act. The European Parliament establishes a clear link between Article 10 (5) AI Act and the material scope of the non-discrimination directives by explicitly addressing only "negative" bias, i.e. bias that creates direct or indirect discriminatory effect against a natural person.

an integrative approach of non-discrimination and data protection desiderata in the conception of Fairness in machine learning.

Limitations: Assessing the necessity of processing sensitive personal data in AI bias detection and monitoring presents a multifaceted challenge that requires continuous consideration of evolving fair ML research and different legal frameworks. The principle of necessity is an inherently vague legal term with its application being highly context-dependent. This makes the formulation of any definitive guidance or rigid formula that can be generalised for all cases of bias detection and monitoring an impossible task. Furthermore, the paper draws on the AI Act, a legislative text that is still under negotiation. Developments through the legislative process might affect aspects of the analysis. Finally, it is essential to acknowledge that bias detection and bias monitoring methods are still in the realm of ongoing research. This poses challenges in exploring their data protection implications within real-world settings, where applicability and effectiveness are mostly untested.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions (grant agreement number 860630) for the project: "NoBIAS - Artificial Intelligence without Bias". Furthermore, this work reflects only the authors' view, and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

References

AIAct (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Art29WP, A. . W. P. (2012). Article 29 data protection working party opinion 3/2012 on developments in biometric technologies.

Balayan, A. and Gürses, S. (2021). Beyond debiasing: Regulating ai and its inequalities. Technical report, Delft, Netherlands.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California law review*, pages 671–732.

Bauerschmidt, J. (2021). The basic principles of the european union's ordinary legislative procedure. In *ERA Forum*, volume 22, pages 211–229. Springer.

BGB (2023). German civil code (bürgerliches gesetzbuch - bgb, federal law gazette. Accessed on 2023-10-01.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., and Eckersley, P. (2020). Explainable machine learning in deployment. In *FAT**, pages 648–657. ACM.

Binns, R., Adams-Prassl, J., and Kelly-Lyth, A. (2023). Legal taxonomies of machine bias: Revisiting direct discrimination. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1850–1858.

Burkov, A. (2020). *Machine learning engineering*, volume 1. True Positive Incorporated Montreal, QC, Canada.

Calvi, A. (2023). Exploring the synergies between non-discrimination and data protection: What role for eu data protection law to address intersectional discrimination? *European Journal of Law and Technology*, 14(2).

Candela, J. Q., Wu, Y., Hsu, B., Jain, S., Ramos, J., Adams, J., Hallman, R., and Basu, K. (2023). Disentangling and operationalizing AI fairness at linkedin. In *FAccT*, pages 1213–1228. ACM.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209.

Charter of Fundamental Rights of the European Union, O. J. o. t. E. C. (2000). Eu charter.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *KDD*, pages 797–806. ACM.

Diethe, T., Borchert, T., Thereska, E., Balle, B., and Lawrence, N. (2018). Continual learning in practice. In *Continual Learning Workshop at NeurIPS 2018*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012). Fairness through awareness. In Goldwasser, S., editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM.

European Council (2022). Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts - general approach. Official European Council Document.

European Data Protection Supervisor (2010). THE EDPS VIDEO-SURVEILLANCE GUIDELINES.

European Parliament (2023). Ai mandates: Document on proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

Finocchiaro, J., Maio, R., Monachou, F., Patro, G. K., Raghavan, M., Stoica, A.-A., and Tsirtsis, S. (2021). Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 489–503.

Fleisher, W. (2021). What's fair about individual fairness? In *AIES*, pages 480–490. ACM.

GDPR (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council. Official Journal of the European Union.

Gola, P. and Heckmann, D. (2022). *Datenschutz-Grundverordnung, Bundesdatenschutzgesetz: DS-GVO / BDSG*. C.H. Beck, 3 edition.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323.

Huber v Bundesrepublik Deutschland (2006). Judgment of the court (grand chamber)of 16 december 2008. ECLI:EU:C:2008:724.

Huyen, C. (2022). *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. O'Reilly.

Information Commissioner's Office (Accessed 2023-11-11). A guide to lawful basis.

Kuner, C., Bygrave, L. A., Docksey, C., and Drechsler, L. (2019). *The EU General Data Protection Regulation (GDPR): A Commentary*. Oxford University Press.

KWG (2023). Banking act (kreditwesengesetz - kwg), bundesministerium für justiz. Accessed on 2023-10-01.

Kühling, J. and Buchner, B. (2020). *Datenschutz-Grundverordnung, Bundesdatenschutzgesetz: DS-GVO / BDSG Kommentar*. C.H. Beck, 3 edition.

Lee, M. S. A. and Singh, J. (2021). The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–13, New York, NY, USA.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35.

Meta Platforms and Others (Conditions générales d'utilisation d'un réseau social) (2023). Judgment of the court (grand chamber) of 4 july 2023 (request for a preliminary ruling from the oberlandesgericht düsseldorf – germany)– meta platforms inc., formerly facebook inc., meta platforms ireland limited, formerly facebook ireland ltd. ECLI:EU:C:2023:537, para 109.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

Mougan, C. and Nielsen, D. S. (2023). Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. In *AAAI Conference on Artificial Intelligence*.

Ntoutsi, E. et al. (2020). Bias in data-driven artificial intelligence systems - an introductory survey. *WIREs Data Mining Knowl. Discov.*, 10(3).

Paleyes, A., Urma, R., and Lawrence, N. D. (2023). Challenges in deploying machine learning: A survey of case studies. *ACM Comput. Surv.*, 55(6):114:1–114:29.

Parliament, E. (2023). Amendments adopted by the european parliament on 14 june 2023 on the proposal for a regulation of the european parliament and of the council on laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. European Parliament Document.

Pedreschi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *KDD*, pages 560–568. ACM.

Petersen, F., Mukherjee, D., Sun, Y., and Yurochkin, M. (2021). Post-processing for individual fairness. In *NeurIPS*, pages 25944–25955.

Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., and Schwaighofer, A. (2009). *Dataset shift in machine learning*. Mit Press.

Rezaei, A., Liu, A., Memarrast, O., and Ziebart, B. D. (2021). Robust fairness under covariate shift. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 9419–9427. AAAI Press.

Schantz, P. and Wolff, H. A. (2017). *Das neue Datenschutzrecht: Datenschutz-Grundverordnung und Bundesdatenschutzgesetz in der Praxis*. Schantz/Wolff, Das neue Datenschutzrecht,(Fundstelle). Nr. 429 et seq.

Simitis, S., Hornung, G., and Spiecker gen. Döhmann, I. (2019). *Nomos Kommentar Datenschutzrecht DSGVO mit BDSG*. Spiros/Hornung, Gerrit/Spiecker, Indra (genannt Döhmann)(Hrsg.), Baden-Baden. Article 6, Nr. 67.

Supervisor, E. D. P. (2017). Necessity toolkit on assessing the necessity of measures that limit the fundamental right to the protection of personal data. Retrieved on 2023.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

TFEU (2007). *Treaty on the Functioning of the European Union*. Official Journal of the European Union.

The European Parliament and the Council of the European Union (2023). Directive (eu) 2023/2225 of the european parliament and of the council of 18 october 2023 on credit agreements for consumers and repealing directive 2008/48/ec. *Official Journal of the European Union*, L series:30.10.2023.

TK v Asociaţia de Proprietari bloc M5A-ScaraA (2018). Judgment of the court (third chamber) of 11 december 2019. ECLI:EU:C:2019:1064, para 47.

Van Bekkum, M. and Borgesius, F. Z. (2023). Using sensitive data to prevent discrimination by artificial intelligence: Does the gdpr need a new exception? *Computer Law & Security Review*, 48:105770.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *FairWare@ICSE*, pages 1–7. ACM.

Wachter, S., Mittelstadt, B., and Russell, C. (2021a). Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.*, 123(3):735–790.

Wachter, S., Mittelstadt, B., and Russell, C. (2021b). Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567.

Weerts, H., Xenidis, R., Tarissan, F., Olsen, H. P., and Pechenizkiy, M. (2023). Algorithmic unfairness through the lens of eu non-discrimination law: Or why the law is not a decision tree. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 805–816.

Xiang, A. (2021). Reconciling legal and technical approaches to algorithmic bias. *Tenn. L. Rev.*, 88:649.

Yurochkin, M., Bower, A., and Sun, Y. (2020). Training individually fair ML models with sensitive subspace robustness. In *ICLR*. OpenReview.net.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Barrett, R., Cummings, R., Agichtein, E., and Gabrilovich, E., editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1171–1180. ACM.

Zliobaite, I. and Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artif. Intell. Law*, 24(2):183–201.