Fairness and Data Protection Impact Assessments

Atoosa Kasirzadeh
University of Toronto
Australian National University
Australia
atoosa.kasirzadeh@anu.edu.au

Damian Clifford
Australian National University
Australia
damian.clifford@anu.edu.au

ABSTRACT

In this paper, we critically examine the effectiveness of the requirement to conduct a Data Protection Impact Assessment (DPIA) in Article 35 of the General Data Protection Regulation (GDPR) in light of fairness metrics. Through this analysis, we explore the role of the fairness principle as introduced in Article 5(1)(a) and its multifaceted interpretation in the obligation to conduct a DPIA. Our paper argues that although there is a significant theoretical role for the considerations of fairness in the DPIA process, an analysis of the various guidance documents issued by data protection authorities on the obligation to conduct a DPIA reveals that they rarely mention the fairness principle in practice. Our analysis questions this omission, and assesses the capacity of fairness metrics to be truly operationalized within DPIAs. We conclude by exploring the practical effectiveness of DPIA with particular reference to (1) technical challenges that have an impact on the usefulness of DPIAs irrespective of a controller's willingness to actively engage in the process, (2) the context dependent nature of the fairness principle, and (3) the key role played by data controllers in the determination of what is fair.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics** → **Governmental regulations**; • **Computer systems organization** → **Embedded systems**.

KEYWORDS

Ethics of Artificial Intelligence; Regulation of Artificial Intelligence; Fairness Principle; Algorithmic Fairness; General Data Protection Regulation; Data Protection Impact Assessments

ACM Reference Format:
Atoosa Kasirzadeh and Damian Clifford. 2021. Fairness and Data Protection Impact Assessments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3461702.3462528

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462528

1 INTRODUCTION

The employment and deployment of machine learning algorithms in social contexts is widespread. These algorithms which are trained on massive amounts of data learn, without being programmed with special rules and principles, to predict with respect to a particular task (e.g., classification) about unobserved data. Machine learning algorithms impact various aspects of our private and public life by being embedded into socio-technical environments in areas as diverse as facial recognition [4], allocation of scarce medical goods [36], credit scoring [27], and criminal justice decision making [2, 8].

With a growing public awareness of their increasing impact, mitigating the unintended consequences of these algorithms for high-stake decision making has become the focus of much discussion in academic, business, and policy circles. One of the most popular ethical solutions to the mitigation of these unintended consequences is the operationalization of various fairness metrics in machine learning ecosystems. Indeed, in the academic and business circles, there has been a growing literature under the umbrella term of "fair machine learning" which claims to positively accommodate, via fairness metrics, some of the negative and/or unintended consequences of unfair prediction-based decision making [18, 34].

In policy making and the legal literature, there has been an ongoing discussion regarding the need for legislative reforms, also reflecting the need to cater for the negative impacts of algorithmic decision making [39]. Authors such as Nemitz (2018) have highlighted the importance of the General Data Protection Regulation (GDPR) in the regulation of algorithmic decision making [35]. In addition, some data protection authorities such as the Information Commissioner's Office (ICO) in the United Kingdom have emphasized the importance of the data protection framework in the policy guidance on developments in Artificial Intelligence [23, 37]. Through the lens of the GDPR, the ICO in particular appears to place significant weight on the importance of the fairness principle, as specified in Article 5(1)(a), to appropriately regulate machine learning systems [37].

The fairness principle has been described as a core principle [16] and the cornerstones upon which the other principles contained in Article 5 of the GDPR are built [6]. Despite its key importance, however, this principle remains somewhat of a nebulous concept and its relationship with the more accountability principle-orientated obligations on those processing personal data remains hard to characterize in a precise manner.

The purpose of this paper is to assess the role of the fairness principle in the requirement to conduct a Data Protection Impact

Assessment (DPIA) contained in Article 35 of the GDPR. If the fairness principle is to play a key role in mitigating the negative consequences of machine learning systems, it is certainly important to understand (1) how it is operationalized in the obligation to conduct DPIA and (2) how to evaluate the success of this mechanism in tackling the negative and/or unintended consequences of automated decision making. Building on this doctrinal legal analysis, we then explore how the fairness principle in the GDPR might be (or indeed, might not be) operationalized through fairness metrics. In addition, this paper critically examines the capacity of DPIAs to effectively accommodate the negative impacts of machine learning systems in light of insights gathered from the academic and business literature on the formalization and operationalization of various metrics of algorithmic fairness through the operation of the data protection fairness principle in the requirement to conduct a DPIA. This paper, therefore, takes some initial steps to connect these distinct bodies of literature.

The paper is structured as follows. In Section 2, we briefly introduce the GDPR and its key concepts and we examine the importance of the data protection fairness principle and the principle's role in the obligation to conduct a DPIA. Building on the insights gathered, in Section 3, we explore the interpretations of the data protection fairness principle in light of the pervasive literature on fairness metrics. Moreover, we investigate the potential roles that these interpretations can have in conducting a DPIA in Article 35 of the GDPR. We diagnose the pitfalls of using fairness metrics in light of the multiplicity of the interpretations of fairness and examine how fairness could play a more defined role, and how policy agendas distinct from data protection might effectively characterize what fairness means in specific contexts. The paper concludes in Section 4 by positioning the role of DPIAs and by calling for a more open debate regarding the role of the fairness principle and the intersection of data protection with different policy agendas in the determination of what processing operations should be deemed de facto unfair.

2 THE FAIRNESS PRINCIPLE AND THE REQUIREMENT TO CONDUCT DPIA

The GDPR aims to mitigate the power and information asymmetries between controllers (and processors) and data subjects or the natural persons to whom the personal data relates [32]. The Regulation, as the key pillar of the European Union's data protection framework, formulates standards for the processing of personal data with personal data defined in Article 4(1) GDPR as,'[...] any information relating to an identified or identifiable natural person ('data subject') [...]'. The Regulation affords rights to data subjects (e.g. erasure, access, rectification), imposes obligations on data controllers and processors, and assigns a monitoring role for data protection authorities. Data controllers are the natural or legal persons 'which, alone or jointly with others, determine the purposes and means of the processing of personal data' (Article 4(7)), whereas the processor is the 'natural or legal entity that processes personal data on behalf of the controller' (Article 4(8)). The requirements that controllers and processors are subject to stem from the principles relating to the processing of personal data contained in Article 5 of the Regulation with these principles guiding the interpretation of the rights and obligations contained therein. The fairness principle, as stated in Articles 5(1)(a), alongside the lawfulness and transparency principles, is one of these key principles.

The fairness principle is also mentioned specifically in Article 8(2) of the Charter of Fundamental Rights of the European Union, with Article 8 stipulating the right to data protection. Despite its fundamental role in debates about the preservation of human rights and artificial intelligence, however, the fairness principle has been largely unexplored and remains undefined in the data protection framework and case law. This is despite the fact that the requirement to process personal data 'fairly' is a standard-bearer in data protection. This in turn presents challenges for controllers in the fulfillment of their obligations.

2.1 Fairness, the fairness principle, and the GDPR

As briefly mentioned in the introduction, the fairness principle is understood by many as the cornerstone upon which the other data protection principles are built. For instance, Bygrave observes that 'it embraces and generates the other core principles of data protection laws' [6] and when positioned as such, fairness is connected to the protection against any negative consequences even in the absence of an intent to deceive on behalf of the controller when personal data are processed. When processing personal data, controllers are therefore obliged to consider the interests and reasonable expectations of the data subject. In a similar vein, Malgieri [33] notes that 'fairness refers to a substantial balancing of interests among data controllers and data subjects', and the principle is, therefore, effect-based in that 'what is relevant is not the formal respect of procedures (in terms of transparency, lawfulness or accountability), but the substantial mitigation of unfair imbalances that create situations of "vulnerability"'. This demonstrates the important connection between the fairness, lawfulness and transparency principles but also the accountability principle provided in Article 5(2) of the Regulation.

Previous literature has explored the overlaps between the fairness, lawfulness and transparency principles in an attempt to delineate the precise role for fairness. Clifford and Ausloos [9], for instance, divide the operation of the fairness principle into a process-oriented manifestation and an outcome-driven fair balancing. They argue that both run concurrently and inter-dependently throughout the application of the Regulation with respect to the ex ante and ex post rights and obligations contained in the GDPR. To clarify, the ex ante application of the fairness principle refers to the rights and obligations which apply prior to the processing of personal data such as the application of the conditions for lawful processing in Article 6(1) or the requirement to conduct a DPIA in Article 35; the ex post safeguards relate to the rights and obligations which apply during personal data processing and are most clearly manifested in the application of data subject rights.

The more process- or procedure-orientated manifestation of the fairness principle results in the burdening of controllers with an obligation to be mindful of data subject's interests and capacities with reference to the ex ante and ex post operation of the information provision requirements. Therefore, the fairness principle is

strongly connected to the transparency principle. The fair-balancing manifestation, on the other hand, refers to the weighing of the rights and interests of data subjects in determining the fairness of a processing operation in a more outcome-orientated manner, again with both ex ante and ex post manifestations. Interestingly, Malgieri [33] observes that this dualist understanding of the role/manifestation of the fairness principle seems to have also been established in the modernized Council of Europe Convention 108. Irrespective of such a division, it is clear that the fairness principle is primarily concerned with mitigating the negative impacts of the power and information asymmetries between the controller (and processor) and data subject.

Indeed, the division of the manifestation of the fairness principle could arguably be categorized within an overarching notion of the fairness principle as concerned with fair balancing in that the procedural fairness manifestations are also indicative of the need to take the rights and interests of the data subjects into account. This seems to align with the idea that fairness, at its core, refers to the need to prevent adverse effects and balance conflicting rights and interests. Here reference can be made for example to various guidance documents demonstrating the link between the fairness principle and non-discrimination [15, 23].

More broadly, the ICO has noted that fairness involves three elements: (1) a consideration of the effects on individuals, (2) the expectations of the data subject, and (3) the transparency of the data processing. Similarly, the French Data Protection Authority (CNIL) has stated that the fairness principle should be interpreted as a means of preventing unfair outcomes or impacts with the effects incorporating not only the perspective of the data subject but also a more collective one. It is important to note, however, that discriminatory effects are just one form that an unfair or unbalanced outcome may take. This appears to reflect the approach taken by the European Data Protection Board (EDPB) in its guidance on Data Protection by Design and by Default in which specific design elements are proposed as a means of considering the implementation of the fairness principle [15]. These design elements are listed in table 1.

This list of design elements aligns well with viewing fairness as a principle intended to counteract power and information asymmetries and thus as a protection against negative consequences stemming from personal data processing even in the absence of an intent to deceive on behalf of the controller. Indeed, some items on the list appear to be abstract examples of design elements that are unfair (Non-discrimination, Non-exploitation, Consumer choice, Power balance, No risk transfer, No deception and Truthful), whereas others seem to relate to specific counter-measures that may be used to prevent unfair outcomes (Interaction, Human intervention and Fair algorithms). The rest appear to reflect important underlying rights and values (Autonomy, Expectation, Respect rights, and Ethical). Thus, it seems uncontroversial to suggest that fairness is inherently linked with balancing competing rights and interests in an ecosystem dictated by power and information asymmetry. As a result, determining what is fair is couched in terms of balancing and analyzing the necessity and proportionality of the processing which feeds into the amorphous and context dependent nature of what might be deemed a fair outcome.

Design element	EDPB explanation
Autonomy	Data subjects should be granted the highest degree of autonomy possible to determine the use made of their personal data, as well as over the scope and conditions of that use or processing.
Interaction	Data subjects must be able to communicate and exercise their rights in respect of the personal data processed by the controller.
Expectation	Processing should correspond with data subjects' reasonable expectations.
Non-discrimination	The controller shall not unfairly discriminate against data subjects.
Non-exploitation	The controller should not exploit the needs or vulnerabilities of data subjects.
Consumer choice	The controller should not lock-in their users in an unfair manner. Whenever a service processing personal data is proprietary, it may create a lock-in to the service, which may not be fair, if it impairs the data subjects' possibility to exercise their right of data portability in accordance with Article 20.
Power balance	Power balance should be a key objective of the controller-data subject relationship. Power imbalances should be avoided. When this is not possible, they should be recognized and accounted for with suitable countermeasures.
No risk transfer	Controllers should not transfer the risks of the enterprise to the data subjects.
No deception	Data processing information and options should be provided in an objective and neutral way, avoiding any deceptive or manipulative language or design.
Respect rights	The controller must respect the fundamental rights of data subjects and implement appropriate measures and safeguards and not impinge on those rights unless expressly justified by law.
Ethical	The controller should see the processing's wider impact on individuals' rights and dignity.
Truthful	The controller must make available information about how they process personal data, they should act as they declare they will and not mislead the data subjects.
Human intervention	The controller must incorporate qualified human intervention that is capable of uncovering biases that machines may create in accordance with the right to not be subject to automated individual decision making in Article 22.
Fair algorithms	Regularly assess whether algorithms are functioning in line with the purposes and adjust the algorithms to mitigate uncovered biases and ensure fairness in the processing. Data subjects should be informed about the functioning of the processing of personal data based on algorithms that analyze or make predictions about them, such as work performance, economic situation, health, personal preferences,reliability or behavior, location or movements.

Table 1: EDPB and the development of the fairness design elements

2.2 The Fairness principle and DPIAs

In line with the above discussion, the requirement to conduct a DPIA and the preliminary assessment to determine whether a DPIA is required can be understood as examples of ex ante regulatory mechanism. Moreover, this requirement can be examined as 'early warning systems' that aim to identify the impact of potential risks, and also to fairly balance and mitigate the potential risks with a clear connection to the accountability principle [30]. Indeed, according to Recital 84 '[t]he outcome of the assessment should be taken into account when determining the appropriate measures to be taken in order to demonstrate that the processing of personal data complies with this Regulation.' The data protection fairness and accountability principles go hand in hand with the controller responsible for the fair balancing of rights and interests when processing personal data. Fairness then manifests itself in the implementation of the rights and requirements provided by the framework to ensure a fair personal data processing ecosystem.

Article 35(1) obliges controllers to perform a DPIA when a data processing operation (or set of similar operations), 'is likely to result in a high risk to the rights and freedoms of natural persons', and in particular if this operation makes use of new technologies. More specifically, Article 35(3) of the Regulation non-exclusively lists three specific cases when a DPIA is required, and Article 35(4) mandates the data protection authorities to publish a list of processing operations that are subject to the requirement to conduct a DPIA. Importantly, the key to determining what 'fairly balanced' personal data processing amounts to is to simply apply the checks and balances in the GDPR (i.e. including the obligation to conduct a DPIA). However, this does not eliminate the need to interpret what is 'fair' (or unfair) in the operation of the requirements through the lens of the fairness principle in Article 5(1)(a) of the Regulation. In other words, the Regulation and the requirement to conduct a DPIA are examples of the fair balance struck by the legislator between the competing rights and interests, but this does not eliminate the need to determine what the fairness principle as provided in Article 5(1)(a), as part of the balance struck by the legislator, means in a specific context. There is, therefore, a need for a better understanding of what fairness means through the lens of the requirement to conduct a DPIA and therefore, an exploration as to how this rather nebulous concept could be operationalized more effectively.

In her analysis of fundamental rights impact assessments in the context of automated decision making, Janssen [25] includes the 'balancing of risk and interests' as one of her key benchmarks and frames this as something coming within the scope of the data protection fairness principle. However, despite the merits of understanding fairness as holding a key role in the requirement to conduct a DPIA, it remains unclear how fairness as a core principle of the Regulation actually applies to the requirement to conduct a DPIA. This is indicative of the fact that a review of the guidance literature exploring the role of the DPIA process reveals a very limited discussion of the fairness principle. For example, there is no reference to the fairness principle in the Article 29 Working Party guidance on the requirement to conduct a DPIA [1], the recent European Data Protection Survey's Report on the DPIAs conducted by the European Union's institutions [17], the CNIL and the Irish Data Protection Commissioner's general guidance on the obligation to conduct a DPIA [13], the CNIL's privacy impact assessment methodology [11], or the ICO's DPIA template [22].

In contrast, some documents seem to make somewhat perfunctory references to the fairness principle. For instance, in its guidance on DPIAs, the ICO mentions that a DPIA may help demonstrate compliance with the 'fairness and transparency requirements' [24] and its guidance on artificial intelligence that a DPIA should include 'an explanation of any relevant variation or margins of error in the performance of the system may affect the fairness of the personal data processing' [23]. Given the lack of a more in-depth guidance, it is therefore necessary to analyze the role of the fairness principle in the requirement to conduct a DPIA in more detail.

More specifically, there are at least three places in which one can view a role for the fairness principle provided for in Article 5(1)(a) of the Regulation in Article 35. The first is in the determination of what is meant by 'high risk' in Article 35(1) where a failure to conduct a DPIA properly (or indeed at all) would seemingly breach the fairness principle. The second is in the interpretation of the situations identified as being specific cases of high risk in Article 35(3). Here reference can be made to Article 35(3)(a) which essentially makes a cross reference to the right not to be subject to an automated decision, including profiling contained in Article 22 of the Regulation. Indeed, it seems appropriate that the assessment of the impact of automated decision making should include a reference to the fairness of the processing operation in question given that such a decision may result in an unfair or biased outcome. As Hacker [19] notes, it would be odd to conclude that a model that racially discriminates processes personal data fairly. Therefore, this opens up the data protection toolbox to mitigate these challenges and therefore seemingly obliges controllers to consider the fairness of the processing in the DPIA. The third is, in a connected sense, the content of a DPIA including the items listed in Article 35(7) and for example, the assessment of the proportionality and necessity of the processing operations required under Article 35(7)(b).

Hence, although it remains somewhat implicit, it is certainly possible to plot the role of the fairness principle in the operation of the obligation to conduct a DPIA and the related provisions. What remains unclear, however, is how fairness should be effectively operationalized. The omission of a detailed analysis of the fairness principle from the various guidance documents issued by the data protection authorities listed above on the application of the requirement to conduct a DPIA is perhaps indicative of the process/procedural orientated ways that such accountability based requirements have been traditionally viewed. It seems, therefore, that generally speaking, as the starting point for the requirement to conduct a DPIA is strongly linked to the accountability principle, there is silence on the role of the fairness principle. This is despite the fact that fairness in Article 5(1)(a) must operate implicitly because it represents the need to fairly weigh the respective rights and interests at stake even in the operation of procedural rights and responsibilities. This omission is perhaps linked to the fact that a DPIA examines planned/future processing of personal data as opposed to ongoing processing operations (i.e. leaving the obligation to review to one side). Despite the lack of a specific and coherent analysis of the fairness principle in this DPIA documentation, there is clearly a role for the fairness principle in the DPIA

process. It would be counter-intuitive and seemingly bizarre to disregard this 'core' principle in what is effectively an accountability and legitimacy check for future processing operations.

3 INTERPRETING THE FAIRNESS PRINCIPLE IN LIGHT OF FAIRNESS METRICS

The discussion so far suggests that there is a need to think more deeply about the relationship between fairness as a 'core' principle of the GDPR and the methodological approach to a DPIA. Such an analysis would allow for a better understanding of the relationship between the principles provided for in Article 5 but also the malleable nature of the fairness principle. However, due to the broad role that fairness plays in the discussion, it is difficult to extract a concrete and actionable framework to assess fairness and inform a controller on how to undertake a balancing exercise given the complex context dependent nature of the principle in its operation. All the practical uncertainties of fairness in Article 5(1)(a) (i.e. in terms of the substantive outcome of a balancing exercise) effectively may mean that for a business, the only aspect that can really be entirely controlled is the conformity with the accountability-based process orientated requirements designed by the legislator to strike a fair balance between the competing rights and interests at stake. This statement will ring true where the underlying business model of the company (or a significant part of it) may in its entirety draw into question its compliance with the fairness principle. This is significant given the fact that guidance documentation on the application of the GDPR to developments in artificial intelligence (such as that issued by the ICO) seem to rely heavily on the fairness principle [23, 37].

The question thus becomes whether the literature on fair machine learning and fairness metrics could aid in formalizing the substantive content of what amounts to 'fair' processing in relation to the obligation to conduct a DPIA to solidify the more substantive outcome-driven (i.e. fair balancing) role for the fairness principle in practice. This points to the need to bridge the gap between the high-level abstract formulation of the fairness principle in the GDPR and the literature on the variety of fairness metrics for making artificial intelligence systems fair. Indeed, given the growth of the literature on fairness metrics as developed by computer scientists [3, 7, 12, 14, 18, 20, 28, 31, 34, 38], fairness metrics are a prominent option for resolving this interpretational issue. We believe that this interpretation is conceptually necessary because the fairness principle is inherently linked to balancing rights and interests in a socio-technical ecosystem dictated by asymmetrical information power. To analyze whether this balancing has happened, some kind of concrete quantified metrics of fairness as a bridge requirement from the regulation frameworks to the artificial intelligence systems will provide important insight.

According to the literature on algorithmic fairness and fairness metrics, there are several approaches to formalize and quantify the conceptions of fairness. These approaches can be categorized, broadly, depending on whether we want to examine the notion of fairness in a statistical or an individualistic sense. Each of these senses can be interpreted in various ways. To provide a flavor of some of the possible notions of fairness metrics, we briefly review a small list of them. This list is by no means exhaustive. For a comprehensive survey, see [18, 34]. Moreover, the appropriateness of the use of some of these metrics will depend on the kind of learning algorithm. For the purpose of this paper, however, we do not engage with this dependence. Our short review merely aims to provide some high level and basic insights in highlighting the challenges of this interpretation task.

The simplest and perhaps the most straightforward conception of fairness is fairness through blindness to some sensitive attributes (e.g., race) that could be the basis of unfair treatment of the subject. In the context of machine learning systems, this means that in order to make the results of predictive algorithms fair, we need to make sure that these algorithms have no information about the sensitive attributes of the subject.[1] Unfortunately, this conception does not appear helpful in many applications in practice because very often (the conjunction of) some features such as demographic information, the postcode (in segregated cities), or the annual income operate as a good proxy for informing the algorithm about the sensitive attributes. This means that although no sensitive attribute is directly given as an input information to the algorithmic system, the collection of some non-sensitive attributes can reliably approximate some sensitive features to which the algorithm is supposed to be blind.

Another popular metric for the assessment of algorithmic fairness is statistical parity [14]. This characterization of fairness requires the predicted outcome of an algorithm to be statistically independent from the sensitive attributes. For instance, in the case of college admission, the predicted acceptance rates for both protected and unprotected groups should be the same (e.g. the acceptance rates of the applicants from different demographic groups must be equal). However, this notion of fairness might render misleading results, for instance when the underlying base rate for the protected and unprotected are different (e.g., fairness of arrest rate for violent crimes). Still another statistical fairness metric, equality of opportunity, measures whether those people who should qualify for an opportunity are equally likely to do so regardless of the group they are a member of [20]. One limitation of this metric is as follows: if one of the goals of a fairness metric is to close the gap between the two subgroups, the metric will not help to achieve that goal.

In addition to statistical fairness metrics, the conception of individual fairness aims to formalize and quantify the notion of fairness relative to the similar treatment of similar individuals [14]. However, measuring the similarity between two individuals in a metric space is an extremely tricky task. One way to provide a more concrete analysis of the notion of individual fairness is to use counterfactuals according to which a predictor's behavior must be compared across counterfactually similar individuals. For instance, Kusner et al. [31] define a fair predictor to be the one that gives the same prediction had the individual were different with respect to

[1] To understand how this idea might apply in practice, it is helpful to consider an algorithm that is used by a criminal justice system for making bail and parole decisions. This algorithm assigns a level of risk to each defendant. The risk assessment takes as input a set of individual's features such as their age and their previous offense history, and outputs the risk of the individual re-offending. Making the algorithm fair through blindness essentially means that any feature that prima facie is taken to be treated unfair should be removed, for instance, when a machine learning algorithm is being trained to evaluate the risk scores.

some attributes, for example had the individual been of another race or gender. This demands an implicit assumption that everything else (except for the tweaked attributes) will be presumed the same for that individual. Unfortunately, this fairness metric also comes with some difficulties in analyzing and evaluating the counterfactual statements. See Kasirzadeh and Smart [26] for some principled arguments against the prevalent use of counterfactual fairness in social contexts.

The discussion so far only specifies some of the primary variants of fairness metrics of the more than twenty definitions discussed in the algorithmic fairness literature (each with their own benefits and conceptual flaws). However, we believe that this brief discussion is sufficient in allowing us to draw some general conclusions about the use of fairness metrics in interpreting the fairness principle. For instance, one of the most significant results in the literature on algorithmic fairness is the impossibility results [7, 28], which show that the simultaneous satisfaction of some of the desirable fairness metrics (except in some trivial cases) is mathematically impossible. Indeed, the abundance of the metrics for capturing algorithmic fairness gives us an important lesson, namely, that if we base the interpretation of the data protection fairness principle in the literature on algorithmic fairness metrics, data controllers will have a high degree of liberty in claiming a "fair" personal data processing ecosystem. This would result in a pluralistic interpretation of the principle on the basis of a codified interpretation of what fairness means and would therefore, seem destined to fail to really move beyond the abstract notion of fairness in the GDPR.

Indeed, the move towards interpreting fairness metrics arguably belies the breadth of the fairness principle as provided for in the Regulation given that such metrics, for instance, seem focused on the removal of bias. Although non-discrimination can certainly be understood as a way in which the fairness principle plays a role in mitigating unfair outcomes, as discussed above, discriminatory effects are merely one example of an unfair outcome under the GDPR as opposed to the apparent equating of fairness and equality in the fairness metrics literature. As an example, in its guidance on Data Protection by Design and by Default, EDPB [15] states that

> 'Fairness is an overarching principle which requires that personal data should not be processed in a way that is unjustifiably detrimental, unlawfully discriminatory, unexpected or misleading to the data subject. Measures and safeguards implementing the principle of fairness also support the rights and freedoms of data subjects, specifically the right to information (transparency), the right to intervene (access, erasure, data portability, rectify) and the right to limit the processing (right not to be subject to automated individual decision making and non-discrimination of data subjects in such processes).'

This aligns with viewing fairness as a principle designed to counteract information or power asymmetries and hence, as a protection against negative consequences stemming from personal data processing even in the absence of an intent to deceive on behalf of the controller, as presented above. But what does the broadness of this role for the principle mean in terms of its capacity to effectively cater for developments in artificial intelligence systems and algorithmic decision making? And what does this mean with respect to the requirement to conduct a DPIA?

Practically speaking, given that defining what is fair in terms of a substantive outcome is context dependent, it is suggested that performing the DPIA process in a thorough fashion in itself will often constitute a large part of doing what is 'fair'. This point is indicative of the fact that, as mentioned earlier, a company's business model may run counter to certain interpretations as to how the fairness principle plays a role in the interpretation of key rights and requirements and thus the categorisation of certain processing operations as unfair and unlawful.

Of course, there are certainly clear instances where a processing operation will be unfair and reference can be made to those that result in direct discrimination as an example. However in practice, the lines to be drawn are far more blurred as the determination of what is fair is open to interpretation, at least until there is a Court ruling. Without Court of Justice rulings on the sticky issues running to the core of what fairness means in concrete contexts, there will always be uncertainty in the DPIA process in terms of the appropriate balance to be struck. There is a need to explore more deeply the relationship between fairness as a 'core' principle of the GDPR and the methodological approaches to the process of conducting a DPIA. Such an analysis would allow for a better understanding of the relationship between the principles provided for in Article 5 but also the various meanings attributed to fairness.

Here reference can also be made to Butterworth [5] who argues that the ICO's focus on fairness in relation to artificial intelligence and machine learning seems to stretch the GDPR and its fair processing requirement to address the challenges such as collective harms. Building on this point, Butterworth [5] finds that there may be a need for 'legislation defining socially acceptable limits and controls on the application of artificial intelligence, and providing effective rights of redress for individuals and groups that may suffer harm.'

Indeed, it has repeatedly been suggested that, for example, consumer law can act as a toolbox for the mobilization of the protection of consumers in order to facilitate more holistic protection [10, 21]. Such a development would allow legal protections to move beyond 'the exclusive realm of informational privacy and self-determination' [29]. This approach may also cater for the difficulties associated with trying to operationalize collective harms. That is, instead of focusing on a reconceptualization of how a group may fit within a fundamental rights framework, legislation may be adopted on the basis of collective concerns in the pursuit of human dignity, individual autonomy and personality in order to mitigate the negative effects of such developments. With this in mind, the recent publication of the draft of a Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislation acts, COM(2021) 206 final 2021/0106 (COD) is a clear indication that European Union policy makers appear to move in this direction as demonstrated by its proposed banning of certain Artificial Intelligence technologies and applications.

Such an approach seems to at least in part recognizes the limitations associated with the emphasis on controller accountability and the GDPR's decentered regulatory approach that is illustrative

of (1) the focus on risk and responsiveness and (2) the enhanced focus on accountability and the auditing of performance [9]. Indeed, these concerns have a clear impact on the operation of the fairness principle due to the fact that there is an inherent reliance on commercial entities to take fairness considerations into account. It should be noted, however, that our discussion does not negate the usefulness of the DPIA as a process but rather recognizes its limitations and that of the fairness principle to truly cater for developments in machine learning in a comprehensive manner.

Finally, our discussion does not render fairness metrics unsuitable for adoption within the requirement to conduct a DPIA. Controllers should be encouraged, and indeed are required, to consider the consequences of their personal data processing operations. Therefore, the DPIA process should be considered as an important element in the accountability principle linked trail established in the Regulation. Instead, it is suggested that the fairness principle in the context of DPIAs is unlikely in itself even if operationalized through fairness metrics (keeping in mind their limitations) to fully cater for the concerns associated with the development of automated decision making systems.

4 CONCLUSION

The data protection fairness principle plays an important role in the requirement to conduct a DPIA and the operation of this process. Fairness however, is rarely mentioned in the literature exploring the requirement to conduct a DPIA. Hence, there is a clear need for further research exploring the reasons for this omission more thoroughly and also in analyzing how this could be incorporated in the guidance issued by data protection authorities. As fairness is a core principle, it would be counter-intuitive to suggest that it plays no role in the determination of the potential impact of future processing operations. Given the rather nebulous nature of the fairness principle, this paper has explored the potential for fairness metrics to operationalize the principle in order to more adequately respond to the potential for unfair outcomes. Although there is certainly a role for fairness metrics in rendering the requirement to process personal data fairly more tangible, we have argued that such an approach also has significant limitations. Indeed, reference here can be made to (1) the technical challenges that have an impact on the usefulness of DPIAs irrespective of a controller's willingness to actively engage in the process, (2) the context dependent nature of the fairness principle and the narrow but also varying interpretations of fairness according to different fairness metrics, and (3) the key role played by data controllers in the determination of what is fair. Hence, although fairness is key to the operation of DPIAs it is unlikely to cater for all our concerns related to the processing of personal data in particular in the context of the employment and deployment of Artificial Intelligence. As such, the arguments in this paper justify the need for a more open debate regarding the role of the fairness principle, DPIAs, and the intersection of data protection with different policy agendas in the determination of what processing operations should be deemed de facto unfair. Our paper has therefore laid the foundation for a more detailed analysis of this topic in light of the forthcoming moves by, for instance, European Union policy makers to regulate Artificial Intelligence.

5 ACKNOWLEDGMENTS

This project was supported by the Humanizing Machine Intelligence Grand Challenge at the Australian National University.

REFERENCES

[1] A29WP. 2017. Article 29 Working Party, Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679 (No WP248 rev.01, 4 October 2017) 1. (2017).
[2] Julia Angwin, Larson Jeff, Mattu Surya, and Kirchner Lauren. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals and It's Biased Against Blacks. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.
[4] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.
[5] Michael Butterworth. 2018. The ICO and artificial intelligence: The role of fairness in the GDPR framework. *Computer Law & Security Review* 34, 2 (2018), 257–268.
[6] LA Bygrave. 2002. Data protection law: approaching its rationale, logic and limits,(Vol. 10). *Information Law Series. The Hague: Kluwer Law International* (2002).
[7] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[8] Angèle Christin, Alex Rosenblat, and Danah Boyd. 2015. Courts and predictive algorithms. *Data & Civil Right: Criminal Justice and Civil Rights Primer* (2015).
[9] Damian Clifford and Jef Ausloos. 2018. Data protection and the role of fairness. *Yearbook of European Law* 37 (2018), 130–187.
[10] Damian Clifford, Inge Graef, and Peggy Valcke. 2019. Pre-formulated Declarations of Data Subject Consent—Citizen-Consumer Empowerment and the Alignment of Data, Consumer and Competition Law Protections. *German Law Journal* 20, 5 (2019), 679–721.
[11] CNIL. 2018. Commission Nationale Informatique & Libertés, Privacy Impact Assessment Application to IOT Devices. (2018).
[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.
[13] DPC. 2019. Data Protection Commission, Guidance Note: Guide to Data Protection Impact Assessments (DPIAs). (2019).
[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[15] EDPB. 2020. European Data Protection Board, Guidelines 4/2019 on Article 25 Data Protection by Design and by Default (Adopted on 20 October 2020) 1. (2020).
[16] EDPS. 2014. *European Data Protection Supervisor, Privacy and Competitiveness in the Age of Big Data: The Interplay Between Data Protection, Competition Law and Consumer Protection in the Digital Economy*. European Data Protection Supervisor.
[17] EDPS. 2020. European Data Protection Supervisor, Survey on Data Protection Impact Assessments under Article 39 of the Regulation (case 2020-0066, 6 July 2020 1). (2020).
[18] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.
[19] Philipp Hacker. 2018. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. (2018).
[20] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
[21] Natali Helberger, Frederik Zuiderveen Borgesius, and Agustin Reyna. 2017. The perfect match? A closer look at the relationship between EU consumer law and data protection law. *Common Market Law Review* 54, 5 (2017).
[22] ICO. 2018. Information Commissioner's Office, Sample DPIA Template (No20180209v0.3). (2018).
[23] ICO. 2020. Information Commissioner's Office, Guidance on AI and data protection (20203006 0.0.39). *ICO* (2020).
[24] ICO. 2021. Information Commissioner's Office, Guidance on Data Protection Impact Assessments (DPIAs) (No 2021101 1.1.64). (2021).
[25] Heleen Janssen. 2020. An approach for a fundamental rights impact assessment to automated decision-making. *International Data Privacy Law* 10 (2020).

[26] Atoosa Kasirzadeh and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 228–236.
[27] Amir E Khandani, Adlar J Kim, and Andrew W Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34, 11 (2010), 2767–2787.
[28] Jon Kleinberg. 2018. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. 40–40.
[29] Bert-Jaap Koops. 2013. On decision transparency, or how to enhance data protection after the computational turn. *Privacy, due process and the computational turn: the philosophy of law meets the philosophy of technology* (2013), 189–213.
[30] Elini Kosta. 2020. Artilce 35. Data Protection Impact Assessment. In *The EU General Data ProtectionRegulation (GDPR): A Commentary (Christopher Kuner et al.)*.
[31] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.
[32] Orla Lynskey. 2014. Deconstructing data protection: the 'added-value'of a right to data protection in the EU legal order. *International & Comparative Law Quarterly* 63, 3 (2014), 569–597.
[33] Gianclaudio Malgieri. 2020. The concept of fairness in the GDPR: a linguistic and contextual interpretation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 154–166.
[34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).
[35] Paul Nemitz. 2018. Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180089.
[36] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[37] Information Commissioner's Office. 2017. Big Data, Artificial Intelligence, Machine Learning and Data Protection (No 20170904). *Version: 2.2* (2017).
[38] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.
[39] Karen Yeung. 2018. A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. *Council of Europe) I-AUT2018) 05 29* (2018).