Doing justice to algorithms. Integrating fairness metrics with a structural understanding of justice[a]

Enea Lombardi[*]

Abstract

Questo articolo esplora i limiti dell'equità algoritmica, in particolare il "teorema dell'impossibilità dell'equità", e discute come una comprensione strutturale della giustizia possa affrontare le relative questioni etiche. Dopo aver presentato i principali modelli di equità algoritmica, sostengo che essi trascurano questioni fondamentali di giustizia, dando priorità a metriche basate sui risultati e isolando il processo decisionale da contesti socio-storici più ampi. Inoltre, quando i tassi di base differiscono, diventa impossibile soddisfare contemporaneamente più di una metrica di equità. Per ovviare a queste carenze, propongo di integrare l'equità algoritmica con la nozione di ingiustizia strutturale di Iris M. Young, che tiene conto delle disuguaglianze radicate nell'interazione tra comportamenti, norme e istituzioni. Questo approccio colloca gli algoritmi nel loro contesto socio-storico più ampio, sottolineando i fattori sistemici che influenzano il processo decisionale e perpetuano risultati ingiusti. Sostengo inoltre che una prospettiva strutturale assegna agli algoritmi un duplice ruolo, in particolare nei casi controversi in cui sono in gioco questioni etiche. In primo luogo, una funzione diagnostica: mettendo in luce gli squilibri e i pregiudizi etici sottostanti, gli algoritmi possono evidenziare le aree critiche per le riforme sistemiche. In secondo luogo, possono fungere da strumenti di valutazione, consentendo la valutazione e la definizione delle priorità delle metriche di equità caso per caso.

Parole chiave: equità algoritmica; ingiustizia strutturale; metriche di equità; riforme sistemiche

This paper explores the limitations of algorithmic fairness, particularly the "impossibility theorem of fairness", and discusses how a structural understanding of justice can address the related ethical concerns. After presenting the main models of algorithmic fairness, I argue that they overlook key justice concerns by prioritizing outcome-based metrics and isolating decision-making from broader socio-historical contexts. Furthermore, when base rates differ, it becomes impossible to satisfy more than one fairness metric simultaneously. To address these shortcomings, I propose integrating algorithmic fairness with Iris M. Young's notion of structural injustice, which accounts for entrenched inequalities rooted in the interplay of behaviours, norms, and institutions. This approach situates algorithms within their broader socio-historical context, emphasizing systemic factors that influence

[a] Received on 31/01/2025 and published on 09/12/2025.
[*] Utrecht University, e-mail: e.lombardi@students.uu.nl.

decision-making and perpetuate unjust outcomes. I further contend that a structural perspective assigns algorithms a twofold role, particularly in contentious cases where ethical controversies are at play. First, a diagnostic function: by exposing underlying ethical imbalances and biases, algorithms can highlight critical areas for systemic reforms. Second, they can serve as evaluative tools, enabling the assessment and prioritization of fairness metrics on a case-by-case basis.

Keywords: algorithmic fairness; structural injustice; fairness metrics; systemic reforms

1. Algorithmic Metrics and the Impossibility of Fairness

Algorithms play an increasingly significant role in shaping social and institutional decision-making, particularly in critical areas such as healthcare resource allocation and pretrial risk assessments. As their application expands across various societal domains, concerns have grown regarding their potential to amplify unjust discrimination and biases. These risks are particularly pronounced when it comes to social groups with sensitive attributes, such as race and gender, which are often factors of systemic injustice. For instance, research has shown that facial recognition algorithms are systematically less accurate when applied to non-white women[1]. Additionally, recidivism prediction algorithms reportedly assign higher risk scores to non-white people, particularly in the U.S. criminal justice system[2]. To address these ethical imbalances, many authors have focused on algorithmic fairness[3].

The notion of fairness generally refers to the moral rightness of a given process, particularly in the context of decision-making. Despite differences among various approaches, the underlying intuition is tied to the idea of equality: a procedure is fair if it is equal, meaning it treats people equally in relevant respects[4]. In algorithmic design, fairness has predominantly been operationalized at the group level rather than focusing on individuals. While individual differences are recognized, current approaches seek to mitigate discrimination and biases by addressing fairness at the group scale, benefiting individuals as members of those groups[5]. Among the many existing metrics, the following overview focuses on the most significant and representative models.

First, the disparate impact metric[6] evaluates fairness by requiring that the ratio of positive outcomes between groups exceeds a specific threshold, often 80%. For example,

[1] See J. Buolamwini, T. Gebru, *Gender shades: Intersectional accuracy disparities in commercial gender classification*, in «Conference on fairness, accountability, and transparency», 2018, pp. 77-91.

[2] See J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine bias*, in «ProPublica», 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] See J. Kleinberg, J. Ludwig, S. Mullainathan, A. Rambachan, *Algorithmic fairness*, in «AEA Papers and Proceedings», 108, 2018, 22-27; D. Pessach and E. Shmueli, *Algorithmic Fairness*, in L. Rokach, O. Maimon, and E. Shmueli (edited by), *Machine Learning for Data Science Handbook*, Springer, Dordrecht 2023, pp. 867-886; S. Barocas, M. Hardt, A. Narayanan, *Fairness and machine learning: Limitations and opportunities*, MIT Press, Boston 2023.

[4] See J. Broome, *Fairness*, in «Proceedings of the Aristotelian Society», 91, n. 1, 1991, pp. 87-102.

[5] See S. Verma, J. Rubin, *Fairness definitions explained*, in «2018 IEEE/ACM International Workshop on Software Fairness (FairWare)», 2018, pp. 1-7.

[6] See M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, *Certifying and removing disparate impact*, in «Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining», 2015, pp. 259-268.

if a hiring algorithm selects 20% of male applicants but only 12% of female applicants, the resulting disparate impact ratio of 0.6 falls below the 0.8 threshold, indicating potential gender bias. In contrast, demographic parity[7] focuses on equalizing the absolute rates of positive outcomes across groups, regardless of underlying differences and sensitive attributes. To exemplify, this model suggests that the hiring rate for women and men should be exactly the same, even if their average qualifications differ. While both metrics aim to promote equity, they may lead to unintended consequences when base rates differ significantly. The notion of base rate refers to the probability of an event occurring within a specific population before applying any predictive model. In the case of disparate impact, groups with higher base rates could be disadvantaged by suppressing positive outcomes, while demographic parity might disregard meaningful variations such as differences in qualifications or experience. To address these challenges, the strategy of equalized odds[8] proposes a more nuanced fairness criterion by requiring that both the true positive rate and the false positive rate are equal across groups defined by sensitive attributes. This ensures that the model's ability to correctly identify outcomes – and its likelihood of making mistakes – is not biased toward any particular group. For example, if a medical diagnostic algorithm correctly identifies 90% of actual positive cases (true positive rate) and falsely flags 10% of actual negative cases (false positive rate) in both white and Black patients, it satisfies equalized odds. Differently, predictive parity[9] ensures that the positive predictive value – the proportion of predicted positives that are actually correct – is consistent across groups. For instance, if a loan approval algorithm predicts repayment correctly for 80% of approved applicants in both male and female groups, it satisfies predictive parity. Error rate parity[10], on the other hand, focuses solely on balancing specific types of errors (e.g., false positives or false negatives), without requiring both to be equal simultaneously, as equalized odds does. For example, a recidivism risk algorithm that has a false positive rate of 20% and a false negative rate of 10% for both Black and white defendants satisfies error rate parity, even though it may not satisfy equalized odds if only one type of error is balanced.

All these metrics are closely tied to the broader notion of accuracy, which defines the proportion of correct predictions made by an algorithmic model. However, accuracy alone is insufficient to guarantee fairness. For a model with high accuracy may still perpetuate systemic inequalities if it exhibits disparate error rates or ignores sensitive asymmetries in base rates[11]. The difference in base rates is critical in this context, as it implies that some metrics are mutually exclusive, making the achievement of complete fairness impossible. This leads to the "impossibility theorem of fairness", which states that when base rates are different, «it is not possible to satisfy multiple notions of fairness simultaneously»[12]. This is particularly problematic because in real life base rates often differ across groups, and issues of injustice frequently arise precisely due to these differences.

[7] See T. Calders, S. Verwer, *Three naive Bayes approaches for discrimination-free classification*, in «Data Mining and Knowledge Discovery», 21, 2010, 277-292.
[8] See M. Hardt, E. Price, N. Srebro, *Equality of opportunity in supervised learning*, in «Advances in Neural Information Processing Systems», 2016, pp. 3315-3323.
[9] See D. Hellman, *Measuring algorithmic fairness*, in «Virginia Law Review», 106, n. 4, 2020, pp. 811-866.
[10] See J. Kleinberg, J. Ludwig, S. Mullainathan, A. Rambachan, *Algorithmic fairness*, cit., pp. 22-27.
[11] See J. Herington, *Measuring Fairness in an Unfair World*, in «Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society», 2020, pp. 286-292.
[12] D. Pessach and E. Shmueli, *Algorithmic Fairness*, cit., p. 873.

Yet, as demonstrated by Chouldechova, «if the base rate […] differs across groups, any instrument that satisfies predictive parity at a given threshold […] must have imbalanced false positive or false negative errors rates at that threshold»[13]. This implies that predictive parity, equalized odds, and demographic parity end up being mutually exclusive for a well-calibrated classifier. For achieving calibration – i.e., ensuring an accurate alignment between predicted probabilities and actual outcomes – contrasts with attaining equality in false negative and false positive rates. This is because groups with different base rates inherently require different thresholds to align probabilities with actual outcomes, but changing thresholds affects the error rates. As a result, ensuring fairness in one aspect will necessarily lead to unfairness in another, as these metrics pull decision thresholds in opposite directions[14].

The impossibility of fairness results in a decisional impasse that offers no baseline for evaluating the moral status of algorithms, particularly in contentious cases – i.e., cases characterized by ethical controversies, such as entrenched inequalities, and contextual variables that are difficult to operationalize. Consequently, this renders some critical ethical concerns unsolvable by fairness metrics alone. These concerns encompass a range of issues, from the case of COMPAS, a pretrial risk assessment algorithm criticized by ProPublica for disadvantaging African American individuals[15], to matters of epistemic hermeneutical injustice in healthcare[16]. As I will show in the next section, the COMPAS case further illustrates that current metrics are inherently incapable of addressing systemic inequalities across groups. To tackle these challenges, I will argue that contentious cases should be addressed by complementing algorithmic fairness with a structural understanding of justice.

2. Algorithmic Fairness and the Structural Understanding of Justice

First proposed by Iris M. Young[17], the notion of structural injustice was developed in contrast to distributive paradigms, which conceive justice as the morally proper allocation of societal goods – framed as resources, opportunities, or welfare – to address unfair social inequalities. While distributive accounts are undoubtedly effective in certain contexts, such as for the allocation of economic resources following an environmental disaster, Young argues that it does not fully capture the scope of justice. To introduce this criticism, she advances a "structural objection", which can be divided into two main claims. First, (i) distributive paradigms overlook the institutional and socio-historical context that shapes the allocation of goods. This oversight results in a twofold harm: it reinforces the underlying context while failing to address the systemic origins of entrenched inequalities. For instance, a fair distribution of economic resources to unemployed people may be beneficial in the short term, but it fails to tackle the systemic factors that necessitated such a distribution in the first place. Consequently, Young argues that (ii) the exclusive focus on distribution neglects the structural dimension of justice. A social structure is an organized

[13] A. Chouldechova, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, in «Big Data», 5, 2017, pp. 153-163: 158.
[14] See R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, *Fairness in criminal justice risk assessments: The state of the art*, in «Sociological Methods & Research», 50, n. 1, 2021, pp. 3-44.
[15] See J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine bias*, cit.
[16] See G. Pozzi, *Automated opioid risk scores: A case for machine learning-induced epistemic injustice in healthcare*, in «Ethics and Information Technology», 25, n. 3, 2023, pp. 1-12.
[17] See I.M. Young, *Justice and the Politics of Difference*, Princeton University Press, Princeton 1990.

field of social positions stemming from «accumulated outcomes of actions of masses of individuals»[18] and institutions acting «according to normally accepted rules and practices»[19]. It represents the skeleton of the social fabric, situating people in positions that are relational, mutable, and characterized by varying degrees of power relations. For example, a specialized working woman may be subjected to the power of her superiors due to her gender, but she can also exert power over unspecialized subordinates. Given that this approach prioritizes systemic constraints, structural injustice occurs when «social processes put large groups of persons under systematic threat of deprivation of the means to develop and exercise their capacities»[20]. Accordingly, since injustice arises from the multifaceted position occupied within the social structure, it cannot be traced back to individual actions or isolated policies. Rather, it emerges from the intertwined relationships among individuals and institutions that are entrenched in everyday practices. This means that structural injustice cannot be rectified solely through ex-post distribution, as the structure functions as a pre-existing field of norms and social positions.

As Kasirzadeh suggests, «most mathematical metrics of algorithmic fairness are inherently rooted in a locally distributive conception of justice», insofar as «they are concerned with how the algorithm would allocate the relevant computational or material goods across different groups»[21]. This resemblance stems from the fact that fairness metrics typically assess how benefits or harms – such as loans, jobs, or medical treatments – are distributed among predefined groups based on sensitive attributes like race or gender. Much like distributive theories of justice, these metrics aim to ensure proportional or equal allocation, without necessarily questioning the structural conditions that shape individuals' positions or access to resources in the first place. Rather, these metrics are predominantly outcome-focused, evaluating fairness at the point of decision-making while abstracting away from the broader social, historical, and institutional contexts that shape both opportunities and data.

As a result, the obstacles faced by algorithmic metrics of fairness are akin to those encountered by distributive paradigms of justice. On closer inspection, the primary obstacles arise from the focus of the normative lenses: the metrics of algorithmic fairness focus solely on current outcomes, overlooking the underlying dynamics that produce them and relying «on a narrow frame of analysis restricted to specific decision points, in isolation from the context of those decisions»[22]. This is exemplified by the case of COMPAS. For the exclusive focus on outcomes has led to conflicts among fairness metrics, which are applied without consideration of the socio-historical context that systematically disadvantages African American individuals. In turn, the implementation of algorithmic metrics results in a situation where «even the best-case scenario – a perfectly accurate risk assessment – would perpetuate racial inequity»[23]. Therefore, since «fairness is operationalized in terms of isolated decision-making processes»[24] and algorithmic metrics

[18] I.M. Young, *Responsibility for Justice*, Oxford University Press, Oxford 2011, p. 62.
[19] Ivi, p. 100.
[20] Ivi, p. 56.
[21] A. Kasirzadeh, *Algorithmic fairness and structural injustice: Insights from feminist political philosophy*, in «AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society», 2022, pp. 349-356: 351.
[22] B. Green, *Escaping the impossibility of fairness: From formal to substantive algorithmic fairness*, in «Philosophy & Technology», 35, n. 90, 2022, pp. 1-32: 4.
[23] B. Green, *Escaping the impossibility of fairness: From formal to substantive algorithmic fairness*, cit., p. 3.
[24] Ivi, p. 4.

are outcome-based, they are inherently incapable of preventing the perpetuation of long-lasting cycles of inequality»[25]. In contrast, a structural understanding of justice combines the evaluation of outcomes with a normative analysis of the context, systemic factors, and social processes that shape these results.

By focusing on the systemic origins of entrenched inequalities, the structural approach introduces three key elements to the normative analysis of algorithms[26]. First, since injustice cannot be fully attributed to individual actors, but emerges from the complex interplay of norms and institutions, decision-making processes must be understood within their broader, systemic contexts. Algorithms, as integral components of these structures, are "sociotechnical entities"[27] that should not be assessed in isolation but must be evaluated as part of the larger fabric of social, economic, and political forces. Second, structural injustice is forward-looking. Rather than simply focusing on offering reparations for past wrongs, this approach prioritizes reforms to current and future conditions. It calls for proactive measures that seek to address the causes of entrenched inequalities by detecting and implementing structural conditions that can guide the decision-making process toward more equitable outcomes. Third, structural injustice is not a static notion. Since structures can accumulate and compound over time, the temporal dynamics of injustice – i.e., how it unfolds and deepens across time – are crucial to the normative analysis of algorithms. This highlights the importance of continuously auditing and reassessing datasets, which are never entirely neutral or value-free. For datasets are shaped by historical and contextual factors that can encode biases and reproduce inequalities, thereby placing algorithms trained on them at risk of perpetuating the very injustices they aim to mitigate[28].

These elements suggest the need for a structural approach that incorporates systemic and contextual considerations into the analysis of algorithmic decision-making. According to Green[29], this model involves three main steps. The first is identifying inequalities by detecting entrenched disparities and examining how social, political, and institutional arrangements reinforce them. This involves analyzing power structures and understanding how deep-rooted practices contribute to systemic inequities. The second stage focuses on enacting reforms, determining what changes could mitigate the identified inequalities, and emphasizing the need to restructure decision-making processes to reduce their role in exacerbating social hierarchies. This phase is not just about implementing technical fixes, but about transforming the systems and practices that give rise to inequality. The third step involves assessing the role of algorithms by analyzing whether and how they can support these reforms. While algorithms can be integral to certain aspects of systemic change, they should be viewed not as standalone solutions, but as contextual tools operating within broader reform agendas. Therefore, this approach suggests that a structural understanding of justice can address the impossibility of fairness in three ways, particularly in contentious cases where ethical controversies are at play.

---

[25] A. Kasirzadeh, *Algorithmic fairness and structural injustice: Insights from feminist political philosophy*, cit., p. 352.

[26] Cfr. J. Himmelreich, D. Lim, *AI and structural injustice*, in J.B. Bullock, Y. Chen, J. Himmelreich, V.M. Hudson, A. Korinek, M.M. Young, B. Zhang (edited by), *The Oxford Handbook of AI Governance*, Oxford University Press, Oxford 2022, pp. 210-231.

[27] See A.D. Selbst, D. Boyd, S.A. Friedler, S. Venkatasubramanian, J. Vertesi, *Fairness and abstraction in sociotechnical systems*, in «Proceedings of the Conference on Fairness, Accountability, and Transparency», 2019, pp. 59-68.

[28] See C. Stinson, *Algorithms are not neutral*, in «AI Ethics», 2, 2022, pp. 763-770.

[29] See B. Green, *Escaping the impossibility of fairness*, cit.

First, algorithms can function more as diagnostic tools that reveal systemic imbalances rather than offering direct solutions. Consider the case of COMPAS, which can be viewed as an instance of structural injustice[30]. Rather than being used to solve pretrial risk assessment, it can be assigned an *ex-negativo* role. This implies that its primary task is not to provide prescriptive guidance on what should be done – such as extending incarceration – but rather to highlight the ethical controversies, biases, and technical ambiguities involved, which can help refine the focus of reform agendas. Accordingly, COMPAS might be employed to expose systemic discrimination against African American people and illuminate potential areas for targeted reforms. To illustrate, this could involve addressing criminogenic conditions in disadvantaged communities, such as through the redesign of urban environments[31] or reforms to the educational system[32], to amend the social structure and reduce systemic recidivism from the outset.

Second, as tackling entrenched injustice is a long-term and multifaceted process, this structural approach can also serve as an evaluative tool to assess and prioritize fairness metrics on a case-by-case basis. Rather than dismissing algorithmic metrics altogether, the structural approach allows for a context-sensitive deployment of such tools by treating fairness metrics not as universally applicable standards, but as evaluative instruments whose relevance must be determined in light of the specific injustices at play. This case-by-case orientation acknowledges that no single metric can capture the full scope of fairness in every situation, but that certain metrics – when interpreted through a structural lens – can help reveal, monitor, and eventually mitigate the particular forms of inequality embedded in distinct socio-institutional contexts. To illustrate, consider again the case of COMPAS. While it failed to satisfy the metric of error rate parity, since African American non recidivists were more likely to be classified as high risk[33], it complied with predictive parity, as outcomes were predicted at the same rate across all groups[34]. In this framework, the structural approach would suggest assessing the two metrics from a systemic perspective. This involves evaluating the algorithm within the broader socio-historical context and determining whether it mitigates or exacerbates existing biases and discrimination. Accordingly, error rate parity performs better from a structural standpoint because, while not exhaustive, its outcomes align with the systemic discrimination experienced by African American people and are responsive to the demands of the social context. This suggests that, when faced with the impossibility of fairness, the structural approach can overcome the decisional impasse by introducing an additional normative layer that broadens the locus of assessment and selects one metric over another. This can be viewed either as a transitional step toward achieving perfectly just outcomes, or as a permanent stage in which fairness is continuously evaluated and refined by justice.

Importantly, this perspective acknowledges that datasets will never be neutral, as they are shaped by historical power asymmetries, institutional practices, and societal

[30] See A. Kasirzadeh, *Algorithmic fairness and structural injustice*, cit.

[31] See P.M. Cozens, *Sustainable urban development and crime prevention through environmental design for the British city: Towards an effective urban environmentalism for the 21st century*, in «Cities», 19, n. 2, 2002, pp. 129-137.

[32] See L. Lochner, E. Moretti, *The effect of education on crime: evidence from prison inmates, arrests, and self-reports*, in «American Economic Review», 94, n. 1, 2004, pp. 155-189.

[33] See J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine bias*, cit.

[34] See A.W. Flores, K. Bechtel, C.T. Lowenkamp, *False positives, false negatives, and false analyses: a rejoinder to "Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks"*, in «Federal Probation», 80, 2016, pp. 38-46.

inequalities. Yet, by embedding fairness metrics within a structural evaluative lens, such biases can be rendered visible, interrogated, and, to some extent, mitigated – turning algorithmic evaluation into a site of critical reflection and corrective intervention. To illustrate, the structural approach is particularly well-equipped to account for the epistemic risks posed by confirmation bias and self-fulfilling prophecies. Rather than accepting algorithmic outputs as neutral or dispositive, this framework embeds them within a recursive process of critical scrutiny. Tools such as COMPAS are not interpreted as delivering authoritative verdicts, but as indicators of deeper systemic dynamics that warrant interrogation. By assigning algorithms an *ex-negativo*, diagnostic function, the structural approach resists the closure of interpretive loops that confirmation bias typically exploits. Instead of allowing predictive scores to validate entrenched assumptions – such as the presumed higher risk of certain demographic groups – the algorithm's role is to surface and destabilize such associations, thereby revealing the socio-political structures that generate them. In this sense, algorithmic metrics serve not to confirm pre-existing beliefs, but to provoke epistemic friction and redirect inquiry toward underlying institutional and historical causes. Thus, rather than erasing confirmation bias, the structural approach renders it visible and accountable, integrating its acknowledgment into the broader pursuit of justice.

Third, this approach promotes a forward-looking, shared responsibility[35] among all stakeholders, including the developers and the implementers of algorithms. This implies that responsibility does not rest solely with manufacturers but extends to users – such as judges in the case of COMPAS – who cannot disregard ethical contentions due to the involvement of multiple actors. Rather, they must adopt a responsible and transparent approach to the use of algorithms, recognizing that they are active participants who could contribute to exacerbating systemic inequalities[36]. This entails that since the impossibility of fairness implicates all actors, conflicting fairness metrics require stakeholders to reassess datasets and contexts in line with their roles within the broader structure. In the case of COMPAS, this could involve judges establishing a committee to evaluate ethical concerns, identify risks, clarify ambiguities and biases – particularly in datasets – and prevent potential misuses. Ultimately, this underscores a further point of distinction of the structural approach: responsibility is a critical component of justice, as it represents the first step toward forward-looking reforms.

One might object that this approach underestimates the role of algorithmic metrics, rendering the notion of fairness redundant and normatively useless. It might seem that normative work is solely carried out by the structural approach, which ends up overshadowing the importance of current metrics. I reply in three ways. First, this approach is designed to complement, not replace, fairness metrics. This is for two reasons. On the one hand, a structural approach is much more complex to operationalize compared to the notion of fairness, as the variables involved are contextual, non-linear, and interrelated[37]. On the other hand, it acknowledges that there are cases where fairness metrics are sufficient – for instance, when base rates are significantly similar, and the context is not characterized by systemic inequalities. Second, although this approach emphasizes the need for human

[35] See I.M. Young, *Responsibility for Justice*, cit.

[36] See R.E. Goodin, C. Barry, *Responsibility for structural injustice: A third thought*, in «Politics, Philosophy & Economics», 20, n. 4, 2021, pp. 339–356.

[37] See B. Green, *Escaping the impossibility of fairness: From formal to substantive algorithmic fairness*, cit.

supervision of the decision-making process, it does not stand in contrast to algorithmic metrics per se. Rather, it advocates for the design, implementation, and cross-use of different algorithms to incorporate systemic and contextual factors, while emphasizing the need for ongoing human oversight and assessment of both the datasets and the results. Third, this approach – as Young[38] emphasizes – moves from injustice rather than justice. In this context, this means that while it aims to design increasingly just algorithms, its normative task arises from the shortcomings of algorithmic metrics and the impossibility of achieving complete fairness. This implies that the operationalization of fairness remains essential, as the structural approach builds upon the results of these metrics – ultimately recognizing their normative validity and utility, as demonstrated by the prioritization of one metric over another. Therefore, this approach acknowledges both the limitations and potential of algorithmic metrics while emphasizing the necessity of ongoing human oversight, ethical reflection, and collective responsibility in designing and implementing fair decision-making processes and just outcomes.

One might further question the use of algorithms altogether. Given their entrenched biases and their *ex-negativo* role, it could be argued that the use of algorithmic metrics seems unjustified from the outset. I reply that, even when the ultimate goal is to transform upstream decision-making, quantitative diagnostics remain indispensable. Without such metrics, systemic injustices embedded in data sets might remain hidden, lacking the provisional benchmarks needed to make them visible, comparable, and manageable. Although fairness metrics often identify injustice only after it has occurred, this retrospective insight is vital, as it uncovers entrenched biases that would otherwise remain undetectable from a quantitative perspective. While this is surely a limitation inherent to algorithms, it simultaneously underscores a key strength of the structural approach: it does not passively wait for injustice to unfold, but rather integrates algorithmic metrics within a broader agenda of anticipatory and preventive interventions. By embedding these tools in a framework oriented toward systemic reform – through policies, institutional redesign, and cultural transformation – the structural approach ensures that the detection of bias is only one component in a wider strategy aimed at dismantling its root causes before they materialize in concrete harms. Properly contextualized, then, algorithmic tools can function as provisional instruments that "mathematize" injustice, offering a quantitative baseline from which to evaluate discrimination and enable structured comparisons across cases and contexts. In turn, these metrics serve an educative role, illuminating the persistence and extent of biases, thereby anchoring the political and institutional will necessary for upstream reform. In this way, algorithmic evaluation does not replace the imperative for structural transformation but reinforces it, providing the empirical foundation essential to dismantling injustice at its roots.

3. Conclusion

This paper explored how Iris M. Young's structural understanding of justice can address the limitations of algorithmic fairness metrics, in particular the "impossibility theorem of fairness". I began by demonstrating that operationalizing fairness in algorithms often

[38] See I.M. Young, *Justice and the Politics of Difference*, cit.

overlooks critical ethical concerns, as it is primarily outcome-based and isolates decision-making from its broader socio-historical context. As a result, systemic injustices risk being neglected, embedded within datasets as neutral information, and eventually reinforced by algorithmic decision-making. I further referred to the "impossibility theorem of fairness" to illustrate that when base rates differ, current metrics are mutually exclusive, rendering the achievement of full fairness unattainable. To tackle these challenges, I proposed adopting a structural understanding of justice to contextualize algorithms and facilitate systemic assessment. I argued that, from a structural perspective, algorithms can serve as (i) diagnostic tools that reveal systemic inequalities and ethical imbalances, thereby identifying critical areas for forward-looking reforms, and (ii) evaluative tools that, in contentious cases, assess and prioritize fairness metrics on a case-by-case basis. Furthermore, I emphasized that this approach fosters a shared account of responsibility among diverse stakeholders, including developers, users, and decision-makers. In conclusion, a structural understanding of justice can overcome the impossibility of fairness by evaluating the normative implications of algorithms and the value-laden nature of datasets, contextualizing decision-making processes, distributing responsibility among all stakeholders, and assuming a systemic perspective that can assess and rank fairness metrics tailored to each case. While this paper focused on contentious cases, a broader caution would suggest integrating the structural approach in all cases, as it consistently offers prescriptive guidance to complement, oversee, and enhance algorithmic fairness across various domains.