

A Human Dignity Metric to Evaluate AI Models

Hüseyin S. Kuyumcuoğlu¹  ¹ Sabancı University, Faculty of Arts and Social Sciences, Istanbul, Türkiye

Abstract

The general moral concern regarding AI models refers to statistical unfairness (or discrimination) these systems impose. Statistical fairness demands that individuals from socially disadvantaged groups receive equal consideration from AI systems as other individuals. Moreover, it is widely accepted that we must also consider avoiding inaccurate results. However, statistical fairness and accuracy cannot be achieved simultaneously. This result leaves us with insufficient tools to pick the morally best AI model, even among a reduced set of models. Therefore, we need a further moral evaluation method (against using fairness as the only criterion) to assess AI models' moral performances to pick the morally superior one(s). This article investigates the extent to which human dignity is a suitable moral criterion to assess these models further. As a result, the benefits and limitations of using human dignity as an assessment criterion of AI models are laid out.

Keywords

Human dignity · AI model evaluation · fairness-accuracy tradeoff · statistical unfairness



“ Kuyumcuoğlu, Hüseyin S. "A Human Dignity Metric to Evaluate AI Models." In *Philosophy, Society and Artificial Intelligence* Vol. VI, edited by M. Ertan Kardeş, and M. Halit Çelikyön, 183-193. *Social Sciences and Artificial Intelligence Series*, Series Ed. by S. Yazıcıoğlu. Istanbul: Istanbul University Press, 2026. <https://doi.org/10.26650/BS/AH11.2026.002-6.14>

© This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License. 

© 2026. Kuyumcuoğlu HS.

✉ Hüseyin S. Kuyumcuoğlu huseyin.kuyumcuoglu@sabanciuniv.edu

Introduction

The general moral concern regarding AI models refers to statistical unfairness (or discrimination) these systems impose.¹ Statistical fairness demands that individuals from socially disadvantaged groups receive equal consideration from AI systems as other individuals. Moreover, it is widely accepted that we must also consider avoiding inaccurate results. In short, we want our AI systems to perform as accurately as possible. However, statistical fairness and accuracy cannot be achieved simultaneously. This conflict is widely accepted in computer science and is sometimes called fairness-accuracy tradeoff (Berk et al. 2018, Corbet-Davies et al. 2017, Kearns and Roth 2020).

Many who write on the issue suggest algorithmic improvements on both metrics (Kearns et al. 2018; Kearns et al. 2019; Kearns and Roth 2020; Liu and Vicente 2022; Valdivia et al. 2021; Wei and Niethammer 2021; Wong 2020). These studies provide tools for decision-makers to assess the balance between fairness and accuracy in machine learning models, offering insights into the limits of unfairness mitigation interventions and the potential for attaining fairer results with minimal impact on accuracy. Nevertheless, these authors concede that the fairness-accuracy tradeoff cannot be annihilated.

When authors compare AI models on fairness and accuracy, they find that some models dominate others. An AI model dominates another if the dominating model is superior to the other model, at least in one of the fairness and accuracy metrics. Hence, when one is to pick a model to solve a particular problem, it is evident that one should pick the model that dominates the alternative models. Nevertheless, some models are not dominated by each other. In these cases, one model can be superior to another in the fairness metric, whereas it might fare worse in the accuracy metric or vice versa. The models that cannot dominate each other form a curve on a two-dimensional graph with the axes showing fairness and accuracy values. This curve is called the Pareto front (or the Pareto frontier).

This result leaves us with insufficient tools to pick the morally best AI model, even among a reduced set of models that fall on the Pareto front. Therefore, we need a further moral evaluation method to assess AI models' moral performances to pick the morally superior one(s). This article investigates the extent to which human dignity is a suitable moral criterion to assess these models further.

Section 1 explains the concept of fairness in AI models and further discloses statistical fairness. It also explains the fairness-accuracy tradeoff. Section 2 introduces the concept of human dignity and its different interpretations. Finally, section 3 explains how the concept of human dignity can inform AI decision-makers to pick the morally superior model when faced with the fairness-accuracy tradeoff.

The fairness-accuracy tradeoff

Multiple authors point out various unfair impacts of machine learning algorithms on members of socially protected groups (Barocas and Selbst 2016, Mittelstadt et al. 2016). Other authors count cases of algorithmic unfairness that also affect individuals who are not necessarily members of these protected groups or who face algorithmic unfairness not necessarily due to their group membership (O'Neil 2016). In return, various methods are suggested to improve algorithms in terms of fairness (Friedler et al. 2019, Lepri et al. 2018).

When we talk about statistical unfairness, we refer to algorithms' disparate treatment of or disparate impact on individuals who belong to socially protected groups. These socially protected groups are usually formed based on attributes mentioned in the international human rights law: "race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status" (UDHR, Article 2).

¹In this article, the terms AI model or AI system refer widely to algorithmic models mainly based on machine learning to accomplish tasks like job hiring, mortgage lending, college admissions, or medical diagnoses.

It is widely accepted that even though these protected attributes are not used in training machine learning algorithms, the algorithms still make predictions that are eventually in line with the possible predictions they would have made were they trained on data that includes the protected attributes. Moreover, one's verdict on whether an algorithm is fair depends on the fairness metric one commits. For example, after ProPublica criticized the notorious recidivism algorithm named COMPAS of Northpointe, the company did not take long to defend itself, holding that its algorithm was fair in its understanding of the concept.² Their dispute stems from one party maintaining false positive rates and the other holding positive recidivism prediction rates to be the relevant fairness metric. According to Northpointe, the positive prediction rates of recidivism were sufficiently similar between Caucasian and African American defendants, showcasing the fairness of their algorithm. On the other hand, ProPublica claimed that it was twice as likely for an African American defendant to be falsely accused of reoffending compared to a Caucasian defendant, giving ProPublica enough reason to accuse the algorithm of leading to unfair results (Corbett-Davies et al. 2017).

Regardless of which statistical fairness metric one picks, one faces a tradeoff between fairness and accuracy for the algorithm in question. Corbett-Davies et al. (2017) prove the tradeoff mathematically before showing its effect on the actual data from Broward County, Florida, initially compiled by ProPublica.³ These authors apply three fairness criteria to measure their impact on accuracy expectations: statistical parity, predictive equality, and conditional statistical parity. The authors then compare the accuracy expectations from the fairness-constrained algorithms with those from the unconstrained algorithm. They observe that the expected crime rate increases, albeit in different amounts, when the algorithm uses a fairness metric compared to the expected crime rates when it runs unconstrained with these metrics.

These analyses show that AI decision-makers can first pick the models that fare better on fairness and accuracy, thus dominating other models. Second, these decision-makers can benefit from suggestions in the literature to improve their models' performance on both fairness and accuracy. Even when AI decision-makers apply these strategies, they cannot eliminate the fairness-accuracy tradeoff. The following sections will explain the concept of human dignity and investigate the extent to which it can be used to evaluate the competing AI models further.

Human dignity as a moral criterion

The literature on human dignity distinguishes various uses of the concept, where the main distinction is between status (or aspirational) dignity and intrinsic dignity (Killmister 2020, Stahl et al. 2023, Rueda et al. 2025). Status dignity is further divided into three categories. The first use refers to dignity that signifies an individual's family belonging as in aristocratic relations of the relevant historical periods. The second use of status dignity refers to the comportment of behavior, showing that an individual behaves in line with the expectations of a specific milieu. Such a use of the concept assigns dignity to someone who follows the proper etiquette. The third use of status dignity refers to how virtuous an individual is. Hence, according to this sub-category, someone would be dignified to the extent they act virtuously.

On the other hand, intrinsic dignity has two separate foundations: the Catholic Christian and the Kantian. The Catholic version assigns dignity to each human being because they are created as equals in the image and likeness of God (Rueda et al. 2025: 2). And finally, according to the Kantian understanding of human dignity, each human has dignity due to their innate possession of a rational nature and the capacity for autonomy. This final account of dignity and its interpretations is what this article takes on to develop and utilize to evaluate AI models in light of the fairness-accuracy tradeoff.

According to Kant, each human being, regardless of whether or not they act morally, possesses dignity stemming from their rational nature. Humans are rational and free, and thus, they can govern their behaviors.

²See Corbett-Davies et al. (2016) for a summary of the discussion. See Angwin et al. (2016) for the initial claim that the COMPAS algorithm was unfair towards African American defendants.

³See Larson et al. (2016).

Governing their behaviors makes them capable of moral action or acting with moral concern. If humans are capable of moral action, they are responsible for what they do and refrain from doing. They are not merely acting as a result of natural impulses or other people's commands. They give themselves their own rules, meaning they are autonomous. This autonomy is why humans have dignity (Kant 2011: 97, 99. [435, 436]).⁴

Kant conceptualizes dignity as a value that cannot be ranked comparatively with any other value or any amount of the same value. Hence, Kant dismisses any possible comparisons between different persons regarding their dignity. Moreover, Kant would not permit the sacrifice of a person's dignity in one place to save more people's dignity elsewhere just because more dignity is better than less. This result does not mean Kant's idea of dignity can be used to oppose any sacrifice of human dignity. Instead, it means one cannot morally justify the sacrifice of human dignity based on its comparison with something more valuable, be it a more dignified person or a greater number of people with dignity (Hill 1980: 92).

Nevertheless, the Kantian account of dignity is not devoid of differing interpretations. Cummiskey mentions three distinct interpretations of the concept: the substantive, the direct, and the contractualist interpretations (Cummiskey 2008).

According to the "substantive interpretation" of the concept, humans possess an inviolable value named dignity based on their rational nature. Since human dignity is inviolable, it is always morally wrong to violate someone's fundamental rights or endanger someone's security to satisfy one's other desires. Such prohibition grounds what Cummiskey calls first-order inviolability (Cummiskey 2008: 403). The problem of the substantive interpretation of dignity is with second-order inviolability. According to second-order inviolability, violating someone's fundamental rights or endangering their security is prohibited even when such acts could save the fundamental rights or secure the safety of even more people. This conclusion is unacceptable because it urges us to intentionally ignore the rights violations of many (Cummiskey 2008: 404).

The "direct interpretation" of Kantian dignity focuses on promoting rational nature rather than non-violating it. This interpretation is what Cummiskey endorses. He gives two principles, which always ensure that human dignity is treated in the right way. According to the first principle, one must always promote the flourishing of rational agency in the forms of fundamental rights, security, and capabilities for autonomous choice. The second principle instructs one to support conditions for individuals to effectively pursue their own ends, for instance, regarding social, economic, and political opportunities. However, the first principle constrains the second, ensuring that rational nature is never sacrificed for mere preference satisfaction (Cummiskey 2008: 392).

Cummiskey argues that traditional deontological interpretations of dignity equate this concept with inviolability or non-interference, whereas his direct interpretation respects dignity by actively promoting rational agency. According to this account, ensuring access to education or healthcare respects dignity more than merely refraining from coercion. The direct interpretation distinguishes between the intrinsic value of rational nature (dignity) and the derivative value of happiness or preference satisfaction. This distinction mirrors Kant's distinction between dignity and price but frames it in terms of promoting autonomy rather than imposing constraints.

Finally, Cummiskey suggests what he names the contractualist interpretation of Kantian dignity. Among various versions of contractualist ethics, T. M. Scanlon's version comes to the fore. According to Scanlon, an action or policy is wrong if and only if it is based on a principle which can be reasonably rejected by someone who holds a particular standpoint (Scanlon 1998: chapter 5). One reason to reject a principle that grounds conduct could be a violation or even threat of violation to human rationality, including the capacity to give oneself rules and determine one's own goals.

Cummiskey's criticism of the Scanlonian contractualist interpretation of dignity is two-fold. First, he suggests that Scanlon's reasonable rejection criterion is redundant because it must point at some moral value other than itself. But if this is the case, the other moral value pointed at is why human dignity is mistreated. Moreover,

⁴Numbers in brackets refer to page numbers in the corresponding volumes of the Prussian Academy edition of Kant's works.

Scanlon's contractualist theory is not helpful in finding this other value. Second, if this other value is human rational nature, then Scanlon's contractualism falls in the same trap as the substantive interpretation, mentioned above, holding that human dignity is inviolable (Cummiskey 2008: 396).

In what follows, I will investigate the extent to which the concept of human dignity can be handy in picking morally superior AI models among the ones that fall on the fairness-accuracy Pareto curve. I will defend that if this concept is to be helpful at all for the mentioned purpose, we must pick either the direct or the contractualist interpretation rather than the substantive one.

A human dignity metric to assess AI models

Understanding the fairness metrics

Corbett-Davies et al. (2017) apply various fairness criteria to actual data from Broward County, Florida, to observe the results of these criteria on pretrial release decisions. In particular, they apply three different fairness criteria (statistical parity, conditional statistical parity, and predictive equality) and compare their results to those stemming from what they name the unconstrained algorithm, the decision procedure that does not use any fairness criterion.

The unconstrained algorithm suggests the decision makers detain a defendant if their predicted risk, i.e., the probability of committing a violent crime if released, exceeds the cost of detention. This rule applies a single threshold to all individuals, regardless of race. Statistical parity requires equal detention rates across racial groups. The decision rule under this fairness constraint applies group-specific thresholds, detaining individuals in each group whose risk exceeds their group's threshold. Conditional statistical parity extends statistical parity by conditioning on "legitimate" factors like prior convictions. The conditional statistical parity rule uses thresholds that depend on both group membership and legitimate attributes. Finally, predictive equality requires equal false positive rates across groups. This rule also uses group-specific thresholds but adjusts them to ensure fairness in error rates.

Authors suggest that fairness constrained algorithms use different thresholds for different race groups to ensure fairness. Their decision leads to the algorithm's perfect alignment with the relevant fairness metric at the cost of an estimated increase in violent crime. The estimated number of violent crimes increases because authors assume risk scores reflect the accurate probability of committing a crime. So, if there are 20 defendants with a risk score of 0.7 (on a scoring scale of 0 to 1), one would expect precisely 14 more crimes in case all these defendants are released. The authors set the algorithm such that for each fairness metric, the algorithm detains 30% of all defendants. Hence, even the unconstrained algorithm, i.e., the decision rule that uses a single risk threshold to minimize expected crime, will come up with a non-zero expected crime number. If the unconstrained algorithm yields 7 expected crimes, then the fairness metric with 14 more expected crimes would give a 50% increase in expected crimes compared to the unconstrained algorithm.

In the calculation that Corbett-Davies et al. (2017) pursue, the fairness metrics ensure that the relevant metric ensures fairness across racial groups. However, this time, different fairness metrics differ in the percentage of detained low-risk defendants. They show that the conditional statistical parity metric gives the lowest percent of low-risk detainees, with again the lowest percent of the increase in violent crime. If we accept a lower percentage of low-risk detainees as a criterion for a more fair model, then in the authors' example, the AI model using conditional statistical parity fairness criterion *dominates* other models because it fares better in terms of both fairness and accuracy. Hence, one would not need a human dignity metric in this case because the best model has already been picked: the one with the conditional statistical parity fairness criterion.

To remind, authors use different risk thresholds for different racial groups to decide detention. For example, suppose African American defendants are accumulated in higher risk scores. In this case, one might have to detain African Americans whose risk scores are 0.6 and higher to ensure conditional statistical parity fairness. Likewise,



if Caucasian defendants are accumulated in lower risk scores, their detention threshold could be 0.3. Such a difference in risk threshold would then ensure the algorithm's alignment with the relevant fairness metric.

However, using different risk thresholds for different groups would face multiple problems. First, even though the above example uses race as the defendants' relevant feature to ensure fairness, one could imagine gender as another relevant feature. Now, the decision-makers would have to use different thresholds for the four combinations resulting from the two sets of binary feature couples (African American vs. Caucasian and male vs. female).⁵ They would need different thresholds for all these new sub-groups to prevent fairness gerrymandering (Kearns et al. 2018). One can imagine how difficult this task would become as more relevant features, like age, disability, etc., are also considered. Second, including race as a feature may lead to legal and policy issues because someone who faces an individually unfair result can demand *strict scrutiny*, at least according to US laws (Corbett-Davies et al. 2017: 804).

Therefore, one might have to use a single threshold value for both races to determine which defendants to detain. For example, regardless of how defendants of different races accumulate around different risk values (as African Americans accumulate around higher risk values whereas Caucasians around lower values), one might determine a single risk value and detain any individual whose risk score is higher. Obviously, it is plausible to expect to detain more African American defendants than Caucasians since more African American defendants than Caucasians have higher risk scores.

In such cases, we can calculate a "fairness gap score" by finding the difference between rates of detainees for each racial group. Here, we can claim that an AI model is fairer as it has a lower fairness gap score. We can also calculate the total expected crime by multiplying the relevant risk scores (a number between 0 and 1) by the number of detainees for each group and then adding them up. The total expected crime number will be an indicator of accuracy of the AI model in our case. A higher total expected crime number indicates a lower accuracy because it will show that our AI model fails to determine those defendants who are not expected to commit a crime if released.

Just as fairness and accuracy cannot be achieved simultaneously (Berk et al. 2018, Corbet-Davies et al. 2017, Kearns and Roth 2020), fairness gap score and total expected crime cannot be minimized at the same time either. In the example we have been following, this result is not surprising. The exemplary numbers in Table 1 illustrate the suggested hypothesis. The numbers there do not reflect the actual data from Broward County, Florida. They are given only for illustrative purposes.

Table 1

This table uses exemplary numbers to illustrate how the detainees could be distributed among races for different risk score thresholds. Each racial group is supposed to represent 400 defendants. The numbers here do not reflect the actual data from Broward County, Florida.

Threshold	Caucasians detained	African Americans detained	Fairness gap	Total expected crime
0.3	250	400	37.5% (150)	30
0.5	170	290	30% (120)	95
0.6	110	190	20% (80)	177
0.7	110	150	10% (40)	207

Table 1 illustrates that at lower risk score thresholds, more Caucasians, as well as more African Americans, are detained compared to higher risk scores. For example, at a risk score threshold of 0.3, 650 defendants out of 800 (400 from each racial group) are detained. As the threshold for detaining defendants increases, detainees for both racial groups decrease. Hence, more defendants are released, increasing the expected crime number. The fairness

⁵Using binary feature couples for race and gender might already incur particular problems due to ignoring non-binary categories. However, even with the simplified binary system, decision-makers face the difficulty of determining multiple thresholds.

gap is higher for the lower threshold numbers. For example, at threshold 0.3, all African American defendants are detained, showing that all those defendants have a risk score above 0.3. On the other hand, only 250 out of 400 Caucasian defendants are detained for this threshold, showing that 150 Caucasian defendants have a risk score below 0.3.

Each line in Table 1 would illustrate a different AI model recommending detaining a different set of defendants. In this example, AI models that use a higher risk score threshold fare better in fairness but worse in accuracy. If one placed these models on a cartesian plane, they would fall on the Pareto front mentioned at the beginning. So, none of these models dominate any other. As a result, one does not have a moral criterion to pick among these four models as long as fairness is the only criterion we use to assess the AI algorithms. The following subsection explains how human dignity can be used to pick among the models that fall on the Pareto front.

Using the human dignity criterion

This section will apply a single fairness metric to the imaginary example in the previous subsection, where a single threshold is used for both races. Even though, in this example, the only difference between AI models is the risk score threshold they utilize, the results of this section can be expanded to models using other fairness criteria, including ones that are not covered in this article.

Remember that Cummiskey lists three interpretations of Kantian understanding of human dignity: the substantive, the direct, and the contractualist interpretations (Cummiskey 2008). This subsection will apply these three different human dignity conceptions to the previous section's example, where four AI models (with four different risk score thresholds) fall on the Pareto front.

The substantive interpretation prohibits violating someone's fundamental rights or endangering their security to satisfy one's other desires. Moreover, violating someone's fundamental rights or endangering their security is prohibited even to save the fundamental rights or secure the safety of more people. When an AI model recommends that an African American defendant is detained due to their risk score, while a Caucasian defendant with all the same relevant features—like prior convictions—is released, one has a moral ground to hold that the former risks unfair treatment before the law. Since it is plausible to hold that due process of law is a fundamental human right, it follows that an AI model that recommends unfair detention of a defendant violates a fundamental right, hence a person's human dignity.

However, except for some trivial cases, any AI model risks violating someone's human dignity in the way just explained. On the contrary, if an AI model were to release all defendants to avoid violating any person's human dignity, this time, it would risk the security of a resident living near a defendant who would commit a crime. Since endangering someone's security also violates their human dignity, this decision is unacceptable, like the previous one.

As a result, unless society renounces the use of AI models in determining recidivism risks, the substantive interpretation of Kant's human dignity is useless in assessing these AI models. One also has to remember that using these AI models as decision support systems for detention has a dual advantage. As algorithms that run on thousands of data points in fractions of seconds, they are pretty efficient compared to their human counterparts, and they are expected to avoid human biases.⁶

The second interpretation of the Kantian concept of human dignity is the direct interpretation. This version of Kantian dignity focuses on promoting rational nature instead of non-violating it. The direct interpretation of human dignity requires one to always promote the flourishing of rational agency in the forms of fundamental rights, security, and capabilities for autonomous choice and ensures that rational nature is never sacrificed for mere preference satisfaction.

⁶One study found that Miami and Philadelphia bail judges were racially biased against Black defendants and that this bias was implicit rather than conscious, explicit bias (Mitchell 2018). We can also think of the notorious "hungry judge effect" beyond the widely known bias types like racism or sexism (Danziger et al. 2011).

The direct interpretation deems "promotion" rather than non-violence as the right attitude toward the rational nature. In practical cases, this difference would require reducing the number of rights violations rather than avoiding any possible violations. Also, one must increase the security of people rather than avoid all breaches of security. Such a difference in general practical attitudes would also affect AI model assessment.

Table 1 shows that AI models with a higher risk score threshold recommend detaining fewer individuals. As long as the false negative rate is not 100%, meaning that the AI model is not inaccurately releasing all defendants, detaining fewer defendants means fewer rights violations for the inaccurately detained. Hence, even with different false negative rates among racial groups, as the AI model's threshold value increases, it better promotes the defendants' human dignity according to the direct interpretation of this concept.

On the other hand, AI models that use a higher threshold endanger the security of more residents, as reflected in Table 1, with the increase in total expected crime. So, AI models with higher thresholds hinder the residents' human dignity. This result contrasts with the previous one, showing that AI models with higher risk score thresholds promote the defendants' human dignity better. If the direct interpretation of human dignity is to be beneficial at all to solve this puzzle, it has to compare the rate of human dignity's promotion on the one hand with the rate of its hindrance on the other.

Table 1 shows that the AI model with a threshold of 0.3 recommends the detention of 650 out of 800 defendants, whereas the model with a threshold of 0.7 recommends the detention of 260 out of 800 defendants. Once we neglect the false negative values for simplicity, the above numbers yield a 60% decrease in the rate of detained defendants. On the other hand, the expected total crime increases by 86% if the model with the higher threshold is picked. Since both a decrease in false detentions and an increase in expected total crime are compared on the same metric of human dignity, we can conclude that these rates show that the model with the lower threshold better promotes human dignity. As a result, the direct interpretation of human dignity picks the AI model with a threshold of 0.3 as the morally superior model.

The calculation above has a couple of assumptions. First, we could assume that a decision-making body has decided that no less than 30% and no more than 90% of all defendants must be detained. These numbers are consistent with the rate of detainees for models that use 0.3 and 0.7 as thresholds respectively. So, to use the direct interpretation of human dignity to pick the best AI model, the decision makers must already reduce the set of models with a particular criterion before utilizing human dignity later. And second, we neglected false negatives for each racial group. The reasoning above without the data on false negatives is still valid; however, knowing these numbers could give a more sturdy result.

The final interpretation of Kant's concept of human dignity is the contractualist one. According to T.M. Scanlon's contractualism, an action or policy is wrong if and only if it is based on a principle that can be reasonably rejected by someone who holds a particular standpoint. One reason to reject an act or a policy is that it violates human rationality, hence human dignity. Cummiskey criticizes the contractualist theory in general and the contractualist interpretation of human dignity in particular, holding that the reasonable rejection criterion is either redundant or, if it is not redundant, then it yields no different results than the substantive account.

In his criticism of Scanlonian contractualism, Cummiskey overlooks the two distinct interpretations of the contractualist theory: the *ex-ante* and the *ex-post* versions. An *ex-ante* moral theory evaluates the rightness of acts or policies before one observes any results of that act or policy. On the other hand, an *ex-post* moral theory evaluates the rightness of acts or policies only after one observes the results. Many Scanlonian contractualists defend the superiority of the *ex-ante* version over the *ex-post* version for moral assessment (Frick 2015, Kumar 2015, Kuyumcuoglu 2021, Kuyumcuoglu 2022, Kuyumcuoglu 2024, Morgan-Knapp 2015).

The contractualist method requires a comparison of complaints of representatives of given standpoints who expect to suffer as a result of different principles' general conduct in a society (Scanlon 1998: Chapter 5). To apply the *ex-ante* contractualist interpretation of human dignity to the imaginary example above, we again compare the effects of two AI models, one with a threshold of 0.3 and the other with 0.7. This time, instead of comparing the

promotion rates of human dignity as we applied in the direct interpretation above, we compare the complaints of two representatives. The first representative is a defendant who expects to suffer a higher likelihood of false detention if the decision makers pick the AI model with a threshold of 0.3. The second representative is a resident who expects to suffer a higher likelihood of being the victim of a crime if the decision-makers pick the AI model with a threshold of 0.7.

Both representatives in this example complain on the same moral ground; they risk the violation of their human dignity. Hence, we can compare the strengths of their complaints and reject the AI model that instigates a greater complaint. The defendant complaining about the AI model with threshold 0.3, can claim that they would incur a 60% higher detention risk if the decision-makers picked this model instead of the model with threshold 0.7. In contrast, the resident complaining about the AI model with a threshold of 0.7 can claim that they would incur an 86% higher risk of being a crime victim if the decision makers picked this model instead of the model with a threshold of 0.3. Since the comparison is on the same moral metric, i.e., human dignity, the resident has a greater complaint than the defendant and reasonably rejects the AI model with a threshold value of 0.7. Therefore, the contractualist interpretation of human dignity, like the direct interpretation, picks the AI model with a threshold of 0.3 as the morally superior one.

The assumptions in the direct interpretation also hold for the contractualist interpretation: 1) the decision makers must reduce the set of models with a particular criterion (for example, they can decide that they will detain between 30% and 90% of all defendants) before utilizing the human dignity criterion. And 2) false negatives are neglected for each racial group. Just as in the analysis of the direct interpretation, these assumptions do not deteriorate the moral strength of the analysis for the contractualist interpretation, although more information could give a stronger analysis.

Authors who utilize *ex-ante* contractualism in applied ethics, take their analysis one step further to claim that the *ex-ante* contractualist conclusion is morally acceptable only if the decision-makers achieve to (1) reduce expected harms, and (2) increase expected benefits for the expected victims of the relevant policy (Kuyumcuoglu 2021, Kuyumcuoglu 2022, Kuyumcuoglu 2024, Morgan-Knapp 2015). In our case, this approach implies that the AI model with a threshold of 0.3 is morally permissible only if the decision-makers execute the mentioned strategies for the defendants who expect false detention.

Further research could offer efficient tangible strategies to this end; however, one can already imagine some measures to take. To reduce expected harms, the AI models can first be trained on a larger data set to avoid false negatives, especially for African Americans who would plausibly expect a higher false detention rate. Second, AI systems could be improved to increase both their transparency and explainability. To be more precise, a more transparent AI model would permit AI audits on its training data and the results, and a more explainable AI model would clarify to the stakeholders how it reached the given risk scores. Increasing expected benefits for these defendants might involve requiring them to participate in community supervision programs instead of being held in jail (Corbett-Davies 2017: 804).

Conclusion

AI models unconstrained by any moral criterion can produce accurate results at the cost of generating unfairness. To solve this problem, multiple authors suggest various fairness criteria (like statistical parity, conditional statistical parity, predictive parity, and many more) to be incorporated into these models. However, fairness and accuracy cannot be achieved simultaneously. Authors have shown this result both mathematically (Corbett-Davies et al. 2017) and empirically (Berk et al. 2018, Corbett-Davies et al. 2017, Kearns and Roth 2020) and named the phenomenon as the fairness-accuracy tradeoff.

When placed on a two-dimensional graph with fairness and accuracy scores as the axes, some AI models form the Pareto front curve. The models that fall on this curve dominate other models that are above the curve. Dominating, in this context, means these models fare better than others, at least in one of the fairness or accuracy



metrics. However, the fairness criterion cannot provide us with a decision rule to pick the morally superior model among the ones on the Pareto front.

This article investigates how the concept of human dignity can help us as a secondary moral criterion, after fairness, to pick the superior AI model among the ones on the Pareto front. This article picks the Kantian concept of human dignity as the best definition of dignity for the given purpose. Later, the article discusses three interpretations of Kantian human dignity, showing how they can be incorporated as a secondary moral assessment criterion after fairness.

The discussion depicts that the substantive interpretation of human dignity is useless for the given problem. Therefore, one must renounce this interpretation. Both the direct and the contractualist interpretations can provide us with consistent moral frameworks to pick the morally superior AI model among the ones on the Pareto front. Moreover, the article shows that the contractualist interpretation is competent in enframing a more detailed method than the direct interpretation. This method still requires further improvements in the direction suggested in the final section.

Future research is required to start assessing AI models with the direct or the contractualist human dignity interpretations. First, one has to see how successful these metrics are with real-world data. The imaginary example investigated in this article provides plausible numbers that could reflect the real-world case in recidivism prediction algorithms. This example shows that the suggested human dignity metrics are philosophically consistent. Nevertheless, research on real-world data is essential to see how valuable results they give.



Author Details Hüseyin S. Kuyumcuoğlu (Dr.)

¹ Sabancı University, Faculty of Arts and Social Sciences, Istanbul, Türkiye

0000-0002-9267-4891

huseyin.kuyumcuoglu@sabanciuniv.edu

References

- Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23, 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104, no. 3: 671–732.
- Berk, Richard A., Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods and Research*, OnlineFirst. <https://doi.org/10.1177/0049124118782533>.
- Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel. 2016. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear." *Washington Post*, October 17, 2016. <https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/>.
- Corbett-Davies, Sam, 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 (KDD '17)*, 797–806. <https://doi.org/10.1145/3097983.3098095>.
- Cummiskey, David. 2008. "Dignity, Contractualism and Consequentialism." *Utilitas* 20, no. 4: 383–408.
- Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences* 108, no. 17: 6889–92. <https://doi.org/10.1073/pnas.1018033108>.
- Frick, Johann. 2015. "Contractualism and Social Risk." *Philosophy and Public Affairs* 43, no. 3: 175–223. <https://doi.org/10.1111/papa.12058>.
- Friedler, Sorelle A., 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 (FAT* '19)*, 329–38. <https://doi.org/10.1145/3287560.3287589>.



- Hill, Thomas E. 1980. "Humanity as an End in Itself." *Ethics* 91, no. 1: 84–99.
- Kant, Immanuel. 2011. *Immanuel Kant: Groundwork of the Metaphysics of Morals: A German–English Edition*. Edited and translated by Mary Gregor and Jens Timmermann. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511973741>.
- Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness." In *International Conference on Machine Learning*, 2564–72. PMLR.
- Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. "An Empirical Study of Rich Subgroup Fairness for Machine Learning." In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 100–109. <https://doi.org/10.1145/3287560.3287592>.
- Kearns, Michael, and Aaron Roth. 2020. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford: Oxford University Press.
- Killmister, Suzy. 2020. *Contours of Dignity*. Oxford: Oxford University Press.
- Kumar, Rahul. 2015. "Risking and Wronging." *Philosophy and Public Affairs* 43, no. 1: 27–51. <https://doi.org/10.1111/papa.12042>.
- Kuyumcuoğlu, Hüseyin S. 2021. "Sweatshops, Harm, and Interference: A Contractualist Approach." *Journal of Business Ethics* 169: 1–11. <https://doi.org/10.1007/s10551-019-04302-9>.
- Kuyumcuoğlu, Hüseyin S. 2022. "A Contractualist Defense of Sweatshop Regulation." *Business Ethics Journal Review* 10, no. 2: 8–13. <https://doi.org/10.12747/jtj02>.
- Kuyumcuoğlu, Hüseyin S. 2024. "Sweatshops, Disrespect, and Interference: How to Interfere in Sweatshops without Disrespecting the Workers." *Business and Professional Ethics Journal* 43, no. 2: 217–41. <https://doi.org/10.5840/bpej2024614156>.
- Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, May 2016.
- Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. "Fair, Transparent, and Accountable Algorithmic Decision-Making Processes." *Philosophy and Technology* 31, no. 4: 611–27. <https://doi.org/10.1007/s13347-017-0279-x>.
- Liu, Suyun, and Luis Nunes Vicente. 2022. "Accuracy and Fairness Tradeoffs in Machine Learning: A Stochastic Multi-Objective Approach." *Computational Management Science* 19: 513–37. <https://doi.org/10.1007/s10287-022-00425-z>.
- Mitchell, Gregory. 2018. "An Implicit Bias Primer." SSRN Scholarly Paper No. 3151740. Social Science Research Network. <https://papers.ssrn.com/abstract=3151740>.
- Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data and Society* 1: 1–21. <https://doi.org/10.1177/2053951716679679>.
- Morgan-Knapp, Christopher. 2015. "Nonconsequentialist Precaution." *Ethical Theory and Moral Practice* 18, no. 4: 785–97. <https://doi.org/10.1007/s10677-014-9552-6>.
- O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.
- Rueda, Jon, Txetxu Ausín, Mark Coeckelbergh, J. Ignacio del Valle, Francisco Lara, Belén Liedo, José Luis Albareda, Heidi Mertes, Robert Ranisch, Vera Lúcia Raposo, Bernd Carsten Stahl, Murilo Vilaça, and Ínigo de Miguel. 2025. "Why Dignity Is a Troubling Concept for AI Ethics." *Patterns* 6, no. 3. <https://doi.org/10.1016/j.patter.2025.101207>.
- Scanlon, Thomas M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
- Stahl, Bernd Carsten, Doris Schroeder, and Rowena Rodrigues. 2023. "Dignity." In *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges*, 79–93. Cham: Springer. https://doi.org/10.1007/978-3-031-17040-9_7.
- Universal Declaration of Human Rights (UDHR). <https://www.ohchr.org/en/human-rights/universal-declaration/translations/english>. Accessed March 4, 2025.
- Valdivia, Ana, Javier Sánchez-Monedero, and Jorge Casillas. 2021. "How Fair Can We Go in Machine Learning? Assessing the Boundaries of Accuracy and Fairness." *International Journal of Intelligent Systems* 36, no. 4: 1619–43. <https://doi.org/10.1002/int.22354>.
- Wei, Susan, and Marc Niethammer. 2022. "The Fairness-Accuracy Pareto Front." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 15: 287–302. <https://doi.org/10.1002/sam.11560>.
- Wong, Pak-Hang. 2020. "Democratizing Algorithmic Fairness." *Philosophy and Technology* 33, no. 2: 225–44. <https://doi.org/10.1007/s13347-019-00355-w>.

