Embracing Contradiction: Theoretical Inconsistency Will Not Impede the Road of Building Responsible AI Systems

Anonymous submission

Abstract

This paper argues that inconsistency among Responsible AI metrics is a feature, not a flaw. Treating divergent metrics as distinct objectives yields three benefits: normative pluralism that reflects diverse stakeholder values; epistemological completeness that preserves richer information about complex ethical concepts; and implicit regularization that prevents overfitting to any single metric and improves robustness. Forcing consistency by pruning metrics narrows values, reduces conceptual depth, and can harm performance. We call for defining acceptable inconsistency and the mechanisms that enable practical, robust alignment.

Introduction

The current practice of Responsible AI is built on metrics. Fairness audits invoke demographic parity (Feldman et al. 2015), equalized odds (Hardt, Price, and Srebro 2016a), or counterfactual consistency (Kusner et al. 2017); privacy claims are based on ε-differential privacy (Dwork et al. 2006); robustness tests span distribution shift (Quiñonero-Candela et al. 2009), adversarial risk (Goodfellow, Shlens, and Szegedy 2015), and calibration error (Guo et al. 2017). Each metric quantifies an abstract virtue into numbers to enable evaluation and optimization. Yet beneath this quantification enterprise lies a fundamental puzzle: **many of these metrics are mathematically incompatible**. Classical impossibility theorems show, for example, that no nontrivial predictor can simultaneously satisfy three most common fairness definitions except in degenerate cases of perfect prediction or equal base rates (Kleinberg, Mullainathan, and Raghavan 2017; Chouldechova 2017; Plecko et al. 2024). Similar tradeoffs plague accuracy versus privacy (Dwork et al. 2006; Zhao and Gordon 2022), political neutrality versus informativeness (Fisher et al. 2025), and interpretability versus expressive power (Rudin 2019).

Conventional wisdom treats these contradictions as *bugs to be patched*: choose a single "right" metric (Dwork et al. 2012; Hardt, Price, and Srebro 2016a) or derive consistency constraints (Agarwal et al. 2018; Zafar et al. 2017). We take the opposite stance. Drawing on work in moral philosophy, Human-Computer Interaction (HCI), and multi-objective optimization, we argue that **theoretical inconsistency is a feature, not a flaw**. The value of preserving theoretically inconsistent metrics emerges across three dimensions: Normatively, they encode distinct commitments from diverse social groups, essential for pluralistic alignment. Epistemologically, they enable a richer understanding of complex Responsible AI concepts. Practically, jointly optimizing these conflicting objectives acts as an implicit regularizer, steering learners away from brittle, single-metric shortcuts and towards solutions that generalize under realistic uncertainty (Neyshabur et al. 2017a; Yu et al. 2020b).

Overall, this paper makes three following contributions:

1. We formalize two kinds of metric inconsistency—*intra-concept inconsistency* (variants of the same ideal collide) and *inter-concept tradeoff* (distinct ideals compete)—and illustrate how each kind could project a complicated picture when applied to reality, with several canonical examples.
2. We synthesize evidence that inconsistent objectives improves ethical coverage, conceptual understanding, and out-of-sample performance, linking insights from optimization theory, Pareto-front geometry, Rashomon set exploration, and Goodhart's Law.
3. We propose a research agenda that shifts the field from eradicating inconsistency to *characterizing acceptable inconsistency*: defining tolerance bands, documenting normative provenance, and designing pluralistic evaluation dashboards.

In short, we invite the community to embrace contradiction as the necessary price and the promise of building Responsible AI systems that serve a pluralistic world. In the remainder of this paper, Section identifies two types of inconsistencies to provide a conceptual foundation for our central claim: theoretical contradictions between metrics—far from being flaws—serve practical and normative purposes. Section presents theoretical and empirical support for this position from the aforementioned dimensions. Finally, Section outlines actionable recommendations for theorists, tool builders, and regulators in engaging with theoretical contradiction.

Conceptual Framework: Two Forms of Metric Inconsistency

To make sense of the tension between Responsible AI metrics, we define two formal types of inconsistency that under-

lie these conflicts. The first, which we term *inconsistency in the concept*, occurs when multiple metrics derived from the same normative concept (e.g., fairness) conflict with each other (see Definition). The second, *inter-concept inconsistency*, arises when optimizing for one desirable metric (e.g., accuracy) degrades performance on another (e.g., privacy or fairness) due to structural tradeoffs (see Definition). Below, for each of these two inconsistencies, we will illustrate with canonical examples and show how, in each case, the inconsistency in theory formed a conversation with empirical results that more or less suggests a approximated consistency.

Definition : Intra-concept inconsistency

Let $\mathcal{H}$ be a hypothesis space (all possible models) and $\mathcal{A} = \{a_1, \ldots, a_n\}$ where $a_i : \mathcal{H} \to \{0, 1\}$ are Boolean metrics that all purport to measure the normative concept *same* A (e.g. fairness). 0 denotes unsatisfied and 1 denotes satisfied. We say $\mathcal{A}$ is inconsistent if

$$h \in \mathcal{H} \text{ such that } \forall i : \; a_i(h) = 1,$$

unless a trivial edge case holds (e.g: perfect prediction, identical base rates).

Interpretation. No single model can make *all* fairness metrics "satisfied" at once except in degenerate situations.

Fairness

Fairness illustrates a classic case of inconsistency between concepts, where multiple metrics derived from the single normative concept of algorithmic fairness cannot be simultaneously satisfied. Kleinberg et al. (Kleinberg, Mullainathan, and Raghavan 2017) demonstrated that three commonly used fairness metrics, equalizing calibration within groups, maintaining balance for the negative class, and maintaining balance for the positive class, could not be concurrently satisfied across multiple groups, with only two exceptions (Kleinberg, Mullainathan, and Raghavan 2017). These exceptions occurred: (1) when the algorithm achieved perfect prediction or (2) when there was no prevalence difference between the groups. Chouldechova (Chouldechova 2017) formulated a similar impossibility result, expressing it as a relationship between *Predictive Positive Value* (PPV), *False Positive Rate* (FPR), *False Negative Rate* (FNR), and *prevalence* (p), as shown in Equation 1:

$$\text{FPR} = \frac{p}{1-p} \cdot \frac{1 - \text{PPV}}{\text{PPV}} \cdot (1 - \text{FNR}) \tag{1}$$

More recently, Bell et al. (Bell et al. 2023) engaged in this theoretical impossibility: Instead of considering the perfectly fair case among multiple metrics, by slightly loosening the constraint from *zero* disparity to *minimum* disparity, one would find plenty of models that were approximately fair with respect to these theoretically inconsistent metrics (Bell et al. 2023). Their empirical studies on 18 real-world datasets revealed that theoretical impossibility results often overstate practical tradeoffs. This perspective gained support from Wick et al. (Wick, Panda, and Tristan 2019), who demonstrated that carefully engineered feature representations could mitigate fairness tradeoffs, and Liu et al. (Liu, Simchowitz, and Hardt 2019), who argued that impossibility results often stemmed from implicit assumptions about data generation processes.

Political Neutrality

Like fairness, political neutrality exemplifies intra-concept inconsistency, where multiple interpretations derived from this single normative concept cannot be simultaneously satisfied. Drawing from political philosophy, John Rawls argued that *procedural*, *aim* and *effect neutrality* could not be jointly satisfied (Rawls 1985, 1988). In particular, the third sense was "undoubtedly impossible" and "futile trying to counteract": educational institutions inevitably tilted the social climate, so we must abandon the neutrality of effect (Rawls 1988). Joseph Raz made a similar point on a parallel concept: comprehensive neutrality, being neutral with respect to the ideals people will adopt in the future, was also unattainable in practice, given that any serious political morality unavoidably shaped the comparative fortunes of conceptions of the good (Raz 1982).

Yet Raz also argued that neutrality "can be a matter of degree" (Raz 1986). Recent empirical work by Fisher et al. (Fisher et al. 2025) inherited and operationalized Raz's "degrees of neutrality" idea by proposing eight mathematically formalized techniques for approximating political neutrality across three levels: output level (refusal, avoidance, reasonable pluralism, output transparency), system level (uniform neutrality, reflective neutrality, system transparency) and ecosystem level (neutrality through diversity). Their evaluation of 9 LLMs in 7,314 political queries demonstrated that these approximations could be practically implemented with measurable tradeoffs between utility, safety, fairness, and user agency. For example, while refusal techniques achieved 100% safety scores, they scored poorly on utility, while reasonable pluralism maintained high fairness but risked information overload.

Definition : Inter-concept tradeoff

Let $\mathcal{H}$ be a hypothesis space and $A, B : \mathcal{H} \to R_{\geq 0}$ be *different* metrics (e.g. accuracy and demographic parity or loss of privacy). There is an (A, B) tradeoff if

$$\sup_{h \in \mathcal{H}: B(h) \leq b} A(h) \; < \; \sup_{h \in \mathcal{H}} A(h) \quad \text{for some } b < \sup_{h} B(h).$$

Interpretation. Constraining B (say, requiring loss of fairness $\leq b$ or privacy $\varepsilon \leq b$) reduces the maximum achievable A (say accuracy) below its unconstrained optimum.

Accuracy–Fairness

An illustrative inter-concept tradeoff involves accuracy and fairness. From an information theory perspective, Zhao & Gordon (Zhao and Gordon 2022) derived lower bounds to

show that satisfying *independence-based* parity notions such as demographic (statistical) parity would impose extra error when group base rates differed. Specifically, for binary class labels $Y \in \{0, 1\}$ and a protected attribute $A \in \{0, 1\}$, given $\mathrm{Err}_g(h)$ denotes the misclassifications rate of hypothesis h on group $A = g$, they proved that

$$\mathrm{Err}_0(h) + \mathrm{Err}_1(h) \geq \left|\Pr(Y = 1 \mid A{=}0) - \Pr(Y = 1 \mid A{=}1)\right|. \tag{2}$$

Thus, when the base-rate gap on the right-hand side was large, at least one group must incur a proportionally large error.

The empirical findings paint a more nuanced picture. Hardt et al.(Hardt, Price, and Srebro 2016b) demonstrated that, via post-processing optimization, *error-rate-matching* criteria—such as equalized odds or equality of opportunity—can be satisfied with negligible loss in overall accuracy. Rodolfa et al.(Rodolfa, Lamba, and Ghani 2021) corroborated this across several public-policy tasks and observed virtually no reduction in precision after post-hoc mitigation of recall disparity. Furthermore, Li et al.(Li, Wu, and Su 2022) showed that enforcing their causal-path fairness constraint can even *improve* accuracy.

The key insight is that the **accuracy–fairness trade-off is not universal but depends on the particular fairness metric**. Independence-based metrics tend to incur an accuracy cost, whereas error-rate-matching criteria, calibration, or certain causal formulations can often be achieved at little or no cost. Analogous debates have arisen over trade-offs between accuracy–interpretability (Rudin 2019; Bell et al. 2022) and accuracy–privacy (Ziller et al. 2021).

The Value of Inconsistent Metrics

Reviewing three of the aforementioned case studies, one might naturally ask: given the theoretical inconsistencies of Responsible AI metrics, does this suggest that the underlying concepts such as "fairness" and "neutrality" behind these metrics are *ill-defined*? Are these concepts such as "fairness" and "neutrality" remain meaningful and valuable in Responsible AI?

We note that this question can be generalized as: if a goal is contradictory (here making the machine learning model political neutral), should one pursue it (to optimize the model for it)? In other words, one only pursue different goals that are consistent with each other? **Our position is that we should embrace this inconsistency**.

In this section, we emphasize the value of preserving theoretically inconsistent metrics by ensuring that these inconsistencies serve three key purposes: 1) Normatively, they uphold value pluralism: each metric captures a distinct moral stance, ensuring that diverse stakeholder perspectives remain visible. 2) Epistemologically, these inconsistent metrics better preserves information of the underlying concept. 3) Practically, conflicting metrics act as regularizers, guiding models toward more robust and generalizable behavior under real-world complexity. Rather than impeding progress, inconsistency enables both ethical inclusivity and technical resilience.

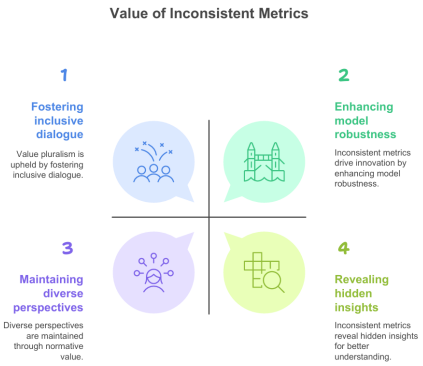

Figure 1: An Overview of The Values in Theoretical Inconsistency

Inconsistent metrics encodes the (unfortunately) inconsistent human values required in pluralistic alignment.

In AI Alignment, pluralistic alignment is an approach that acknowledges and embraces the diversity of human values, perspectives, and preferences rather than attempting to align AI systems with a single, universal set of values (Leike et al. 2018; Ji et al. 2024; Sorensen et al. 2024a). The core premise of pluralistic alignment is that there isn't one "correct" set of human values that AI systems should adopt. Instead, it recognizes that different cultures, communities, and individuals have varying and sometimes conflicting values, and that AI systems should be designed to accommodate this diversity, rather than propagating bias and systematic injustice (Sorensen et al. 2024b; Alamdari et al. 2024; Gabriel et al. 2024; Mehrabi et al. 2021; Crawford et al. 2019).

From the famous proverb, "There are a thousand Hamlets in a thousand people's eyes", it is natural for a concept to be understood differently among people, social groups, and culture. There is no exception for core concepts in Responsible AI such as "fairness", "privacy", and "political neutrality". Each metric of each of the concept exactly represents one way of understanding the concept by an individual or a social group. In turn, people from different social group may have divergent conception of a single concept. For example, psychology literature showed that people from individualist cultures tend to favor the rule of equity and the distributive principle of equity, while people from collectivist cultures emphasize equality and need rules, especially with members of the group (Morris and Leung 2000; Bond, Leung, and Schwartz 1992; Tyler et al. 1997). This might suggest that if a model is intended to be applied to a society where people from divergent social backgrounds blend, to ensure pluralistic alignment, multiple fairness metrics including demographic parity and individual fairness or equal opportunity are needed.

On the other hand, diversified human values are always inconsistent with each other. Isaiah Berlin argued that fundamental human values are inherently pluralistic and sometimes cannot be reconciled theoretically (Berlin 1998). As

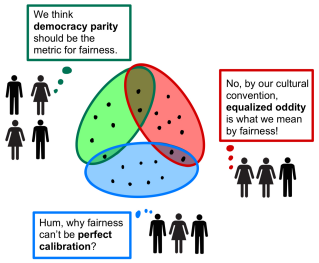

Figure 2: Represented as clusters in different colors, different groups have different ideas about what "fairness" means and formalized them into divergent metrics. By the Impossibility Theorem of Fairness, no models, as represented by black dots, can satisfy all of the three formalized metrics of fairness.

Berlin argued, **the incompatibility of values does not make them less valid** - it is a natural condition of human life that we must navigate. Value pluralism offers an alternative to both moral relativism (all values are equally valid) and moral absolutism (only one set of values is true)(Fisher et al. 2025). In this view, the lack of theoretical consistency between Responsible AI metrics isn't a flaw but reflects the genuine plurality of human values. Thus, if one finds several metrics of a single Responsible AI concept inconsistent (i.e: facing intra-concept inconsistency as defined in), this suggests that the diversified understandings of that concept are inconsistent.

Therefore, if one manages to fix the inconsistency by simply considering a subset of the set of all metrics around the concept, it would harm the goal of pluralistic alignment, as it made some of the respective understanding of the concept behind the deleted metric underrepresented in the evaluation and optimization. For instance, in the case of fairness, if one focuses on solely optimizing for Predictive Positive Value and False Negative Rate, the idea of fairness as understood by people who proposed False Positive Rate is underrepresented. Therefore, optimizing machine learning models with respect to these inconsistent metrics preserves the plural human values, thus helping the agenda of pluralistic alignment.

Inconsistent metrics better preserves information of the underlying concept.

Besides the ethical dimension of inconsistent metrics, we further argue that epistemologically, preserving inconsistent metrics around a concept is to best preserve the information contained within that concept.

Wittgenstein's analysis of 'family resemblance' terms shows that many everyday and moral concepts such as 'game', 'fair', 'neutral' do not have one set of necessary and sufficient conditions. There is no one property (or fixed list of properties) that every "game" has and that only games have. Some games are competitive, some are cooperative; some require skill, others chance; some have rules, others only conventions. Instead, different games share different combinations of features. Chess and soccer both have rules, but chess and tag both involve winning–losing. Each pair overlaps on some trait, but there's no single trait common to all (Wittgenstein 1953).

A formal metric of a concept selects one cluster of it and thereby projects only part of the resemblance network. Any projection as such might be useful, but inevitably left to be a 'rational reconstruction' that sacrifices some of the original content (Carnap 1950). For instance, to capture human "intelligence", psychologists proposed several metrics, such as the genuine Intelligence Quotient (IQ) (Wechsler 1958), Emotional Quotient (EQ) that captures social and emotional understanding (Salovey and Mayer 1990; Goleman 1995), Gardner's Multiple Intelligences that capture linguistic, spatial, musical, and other domains (Gardner 1983), and Practical Intelligence that captures real-world problem solving (Sternberg 1985; Sternberg and Wagner 1986). Each of these tests captures a limited cluster of what we mean by "intelligent." Similarly, for concepts in responsible AI such as political neutrality, each of the breakdown dimension such as "refusal", "avoidance", or "reflective neutrality" as specified in (Fisher et al. 2025) captures a limited cluster of "political neutrality".

Hence, **the best attempt to capture a concept is to incorporate all the possible metrics (which are potentially inconsistent) into evaluation**, so that all the possible clusters are getting involved, and therefore the original content is best preserved. One would have a better understanding of their "intelligence" by considering their performance on all of the possible tests; one would have their best capturing of the degree of political neutrality of an AI model if its performance on all the possible dimensions of neutrality is considered. Conversely, if one manages to fix the inconsistency by simply considering a subset of the set of all metrics around the concept, it will inevitably lead to the loss of some information contained in the concept.

Metric Inconsistency Has Potential Practical Values

Furthermore, we emphasize the potential practical value of metric inconsistency. The conventional wisdom that *one metric suffices* is often fragile in optimization practice: once a single score becomes the sole optimization target, models exploit idiosyncrasies of the training distribution—a phenomenon formalized by Goodhart's Law (Goodhart 1975). Here, we show that maintaining and jointly optimizing several conflicting objectives counteracts such specification overfitting. **It acts as an *endogenous regularizer* that enhances out-of-sample accuracy and robustness**. Three complementary mechanisms are identified and supported by theory and evidence.

Mechanism I: Gradient Conflict as Semantic Regularization In multi-objective learning, when optimizing $\mathcal{L}(x;\theta) = \sum_i \lambda_i L_i(x;\theta)$, objectives are said to conflict when the angle between their gradients exceeds 90°(Yu et al. 2020b). This misalignment bounds the effective gradient norm via the parallelogram law and prevents any one loss from dominating updates(Boyd and Vandenberghe 2004). Crucially, these conflicting gradients introduce *semantic regularization*: unlike stochastic methods such as dropout,

this regularization arises from meaningful tension between goals and narrows the train–validation gap (Yu et al. 2020b; Liu et al. 2021a). Algorithms like PCGrad (Yu et al. 2020a) and CAGrad (Liu et al. 2021b) leverage this structure to promote task-agnostic feature learning. For example, in NYUv2 semantic-depth co-training, PCGrad improves mean IoU by 2.3 pp while also reducing depth RMSE by 0.04m (Yu et al. 2020a). Theoretically, when the L_i are L-Lipschitz and of VC-dimension d, the generalization bound under joint optimization degrades only by $\mathcal{O}(\sqrt{k})$—a modest price for increased robustness (Neyshabur et al. 2017b).

Mechanism II: The Existence of Pareto Fronts and the Rashomon Set Since conflicting objectives rarely admit a unique minimizer, they may induce a *Pareto front* with a ε-optimal region $\mathcal{R}_\varepsilon = \left\{\theta \mid \exists\, \theta' :\ L_i(\theta) \leq L_i(\theta') + \varepsilon,\ \forall i\right\}$. This coincides with the notion of *Rashomon set*, defined as a set of near-optimal models with some $0 << \epsilon << 1$ accuracy loss (Fisher, Rudin, and Dominici 2019; Semenova, Rudin, and Parr 2022; Rudin et al. 2024). Here, a larger $\mathcal{R}\varepsilon$ yields two tangible benefits that can now be computed. First, it allows practitioners to *swap* models to satisfy downstream constraints such as fairness, interpretability, or energy budgets, without retraining or sacrificing accuracy. This flexibility is enabled by recent algorithmic advances, such as TreeFARMS and the GAM Rashomon Set algorithm, which make it feasible to enumerate near-optimal models across the Pareto front in minutes (Xin et al. 2022; Zhong et al. 2023). Second, the functional diversity within $\mathcal{R}\varepsilon$ enhances the ensemble and majority vote strategies, improving the robustness to adversarial perturbations and distributional shifts(Durrant and Kabán 2020). These practical benefits are supported by both classical methods like NSGA-II (Deb 2001) and modern enumeration techniques for decision trees (Bell et al. 2022), which show that Pareto-style search can reliably uncover diverse, near-optimal hypotheses—transforming ambiguity in model selection into a powerful tool for structured flexibility.

Mechanism III: Complementary Metrics Block Shortcut Features Requiring simultaneous performance on inconsistent metrics forces the model to abandon brittle shortcut features. For instance, in medical imaging, injecting a differential-privacy (DP) loss caps memorization, thereby *improving* external-hospital AUC by 5% despite a 2% decline on the internal test set (Ziller et al. 2021). In text classification, optimizing both sentiment accuracy and gender independence removes name-related artifacts, raising cross-domain F_1 (Hardt, Price, and Srebro 2016b). Empirically, errors on one metric often flag spurious correlations exploited by the other, creating a form of *cross-metric debugging* unavailable in single-objective training.

From the above three mechanisms, inconsistent metrics are not an impediment but a *safeguard*: Especially in high-stakes domains—health, finance, criminal justice—joint optimization delivers a principled trade between slightly lower headline scores and substantially higher reliability.

Recommendations and Future Directions

Advancing Practice-Driven Theories on Responsible AI. A question yet to discussed from Section is the gap between the inconsistencies in theory and the more complicated empirical results: Given that in practice, one can solve the inconsistency by loosing constraints, do the theoretical results really matter? How should we build theories on Responsible AI topics?

We highlight that the gap between theoretical inconsistencies and practical consistencies of Responsible AI metrics does not suggest that *any* theories on Responsible AI are not needed. Instead, it pushes researchers to **build theories that fit to everyday Responsible AI practices**. While current theoretical formulations such as the Impossibility Theorem of Fairness often focus on optimality, the models that work well in practice are frequently sub-optimal, approximate, and constrained. What Responsible AI currently lacks a theoretical framework that adequately explains everyday cases of Responsible AI Practices. We should develop responsible AI theories that fit the problem context, rather than demanding universal consistency.

One of the very recent theoretical developments alongside this agenda is the theories on the Rashomon set, which theorizes properties of models with near-optimal performance, a practically feasible setting. In a recent work by Dai et al.(Dai et al. 2025), researchers explored several theoretical properties of the Rashomon set. In particular, they derived that the asymptotic size of the Rashomon Set (and thus the possibility of finding the desired models) grows exponentially with $\sqrt{\epsilon}$. This entails that in practice, a company searching for fairer models within the Rashomon set should use the largest error tolerance acceptable to their business. We argue that in the field of Responsible AI, practice-driven theories such as this should be the direction of future theoretical work.

Defining Acceptable Inconsistency Thresholds. Instead of pursuing the unattainable goal of perfect alignment for all fairness metrics, we recommend specifying explicit tolerance ranges. These ranges serve as normative guardrails, defining acceptable divergence levels among metrics before ethical concerns arise (Ruf and Detyniecki 2021). This approach preserves pluralism, provided that no single objective can override or weaken the legitimate claims of others. Crucially, determining an acceptable inconsistency is a highly *contextual* process; it encompasses not only the magnitude of the divergence but also its nature. Suitable thresholds will be dependent on empirical risk, stakeholder priorities, legal norms, and furthermore the specifics of each deployment scenario (Holstein et al. 2019). Therefore, a key challenge is providing guidance to evaluate whether particular inconsistencies are beneficial, such as by enhancing generalization, or detrimental, such as by leading to failures in essential model operation. Through this approach, we can preserve the benefits of pluralism, accommodating diverse values alongside perspectives, without descending into arbitrary or unchecked tradeoffs.

Documenting Normative Assumptions Explicitly. As we discussed in Section , each metric represents a diversified interpretation behind it. We recommend that all Respon-

sible AI evaluations should capture the normative assumptions embedded in each metric. This proposal builds on the success of Model Cards (Mitchell et al. 2019), which aimed to foster transparency in model reporting by detailing intended use cases, evaluation conditions, and ethical considerations. Similarly, Data Statements (Bender and Friedman 2018) provided schema for documenting data set creation rationale, demographic coverage, and limitations, helping practitioners understand what system behavior can (and cannot) be trusted to generalize. We suggest adapting these principles into a "Metric Provenance Sheet", a structured documentation explaining what each metric measures, its limitations and which stakeholder values it reflects or omits. We expect that this explicit normative tracking will improve interpretability when models excel in one metric but falter in another, promoting stakeholder trust through transparency and accountability (Chmielinski et al. 2024).

Testing Human–Metric Interaction Empirically. To validate the practical relevance of theoretical inconsistency, we call for empirical studies involving stakeholders such as users, domain experts, and regulators in the negotiation and selection of metrics. Previous work in HCI and Responsible AI has shown that participatory methods effectively capture diverse notions of fairness and stakeholder-specific priorities. For example, Cheng et al.(Cheng et al. 2021) involved child-welfare practitioners in defining fairness criteria, revealing substantial variation in ethical interpretations.Future studies should examine how people interact with pluralistic evaluation tools, make tradeoffs between conflicting objectives, and interpret explanations of inconsistency. Such insights will inform the design of more intuitive interfaces, inclusive optimization strategies, and accountable AI policies.

Response to Alternative View

This plan sounds great, but practically speaking, conflicting metrics will confuse end-users and regulators!

Response: One may argue that in practice, regulatory and deployment environments require single and clear standards and that conflicting metrics complicate this picture. A single metric that represents all Responsible AI concepts and diverse perspectives is indeed an appealing ideal, but this ideal is often impractical. To truly represent plural perspectives, multiple metrics are often needed, and using them also helps incorporate complete information. Although this approach is more complex than a single score, it accurately reflects the multifaceted nature of Responsible AI. Multi-metric evaluation can lead to inconsistencies; however, regulators and deployment teams can address these using the approach suggested by Bell et al. (Bell et al. 2023). This method avoids strictly satisfying one specific metric. Instead, it incorporates several different metrics, allowing each to be 'approximately satisfied' within a small, defined tolerance. This enables regulators to define clear standards based on acceptable profiles across key metrics, rather than relying on a single, potentially oversimplified one. Empirical results in the literature support this approach (Bell et al. 2023; Dai et al. 2025; Laufer, Raghavan, and Barocas 2025), as research shows, this tolerance allows for a great variety of models that satisfy established constraints. Therefore, such a multi-metric evaluation framework is not only practically feasible but also essential for robust and responsible AI governance in real-world deployment.

Be careful here! It is the diversity of metrics, not the inconsistency of metrics, that helps pluralistic alignments.

Response: We acknowledge that it is not a logical necessity that inconsistency *entails* plurality. And with no doubts, we wish all the metrics to be consistent with each other while not harming plurality: all concepts do not suffer from any internal consistencies so that we can possibly achieve zero loss for all the metrics, and there will be no tradeoffs among any pair of concepts central in Responsible AI. However, it is believed that inconsistent perspectives and incommensurable values are ubiquitous among real social interactions, as we established earlier in Section . In practice, a multitude of metrics will almost inevitably exhibit inconsistencies. Therefore, attempts to enforce consistency by reducing the number of metrics inherently sacrifice valuable diversity and plural representation. Hence, the viable approach to preserve plurality is to preserve inconsistent metric.

I see your point on the value of inconsistent metrics, but instead of considering all these metrics, shouldn't one pick a metric that works best for the application tasks, which does not involve inconsistency?

Response: We indeed admit that one should choose the metric that works best for their related application tasks. However, as we illustrated in Section there are always cases in which the targeted population shares conflicting perspectives, which requires the "best fit" evaluation method to be itself incorporating plural perspectives. This means that multiple theoretically inconsistent metrics are still needed.

Besides, even for tasks that do not assume to represent plural perspective, inter-concept tradeoff as defined in remains present. For example, Differential Privacy is legally mandated (U.S. Department of Health and Human Services 2024) and ethically required to protect patient identity, yet it is exactly the privacy metric that has a tradeoff with task precision (ROC-AUC) (Ziller et al. 2021).

Furthermore, the effectiveness of each metric in evaluating the respective domain is not static. Some metric might be good for the application at the beginning, but, by Goodhart's law, when a good metric became a target to explicitly optimize for, it ceased to be a good one. Yet, if multiple (inconsistent) metrics are present in evaluation and optimization, as we established earlier in Section , the model will have less chance of suffering from the negative impacts of Goodhart's law.

Conclusion

Metric inconsistencies are essential, not defects. They preserve diverse values, capture ethical complexity, and regularize models for better generalization. Empirical results show that impossibilities become workable tradeoffs when we allow near-optimal, approximate satisfaction. The field should define acceptable inconsistency thresholds and build tools that navigate value tensions rather than erase them.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In *Proc. 35th International Conference on Machine Learning (ICML)*, volume 80, 60–69. PMLR.

Alamdari, S.; Klassen, T. Q.; Toro Icarte, R.; and McIlraith, S. A. 2024. Being Considerate as a Pathway Towards Pluralistic Alignment for Agentic AI. *arXiv preprint arXiv:2411.10613*.

Bell, A.; Bynum, L.; Drushchak, N.; Herasymova, T.; Rosenblatt, L.; and Stoyanovich, J. 2023. The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice. arXiv:2302.06347.

Bell, A.; Solano-Kamaiko, I.; Nov, O.; and Stoyanovich, J. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 248–266. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Bender, E. M.; and Friedman, B. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 587–604. Association for Computational Linguistics.

Berlin, I. 1998. Isaiah Berlin on Pluralism. https://www.cs.utexas.edu/~vl/notes/berlin.html. Archived at the University of Texas at Austin. Copyright: The Isaiah Berlin Literary Trust and Henry Hardy.

Bond, M. H.; Leung, K.; and Schwartz, S. H. 1992. Explaining choices in procedural and distributive justice across cultures. *International Journal of Psychology*, 27(2): 211–225.

Boyd, S.; and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press.

Carnap, R. 1950. Empiricism, Semantics, and Ontology. *Revue Internationale de Philosophie*, 20–40.

Cheng, H.-F.; Stapleton, L.; Wang, R.; Bullock, P.; Chouldechova, A.; Wu, Z. S. S.; and Zhu, H. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Chmielinski, K.; Newman, S.; Kranzinger, C. N.; Hind, M.; Vaughan, J. W.; Mitchell, M.; Stoyanovich, J.; McMillan-Major, A.; McReynolds, E.; Esfahany, K.; Gray, M. L.; Chang, A.; and Hudson, M. 2024. The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers. Harvard Kennedy School Shorenstein Center discussion paper.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163.

Crawford, K.; Whittaker, M.; Chwialkowski, K.; Barocas, S.; Boyd, D.; Gangadharan, S. P.; Wallach, H.; D'Ignazio, C.; Virdi, R. B.; and Green, B. 2019. AI bias and the problems of ethical locality. *Science and Engineering Ethics*, 25(4): 993–1010.

Dai, G.; Ravishankar, P.; Yuan, R.; Neill, D. B.; and Black, E. 2025. Be Intentional About Fairness!: Fairness, Size, and Multiplicity in the Rashomon Set. arXiv:2501.15634.

Deb, K. 2001. *Multi-objective optimization using evolutionary algorithms*. Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons. ISBN 978-0471873396.

Durrant, R.; and Kabán, A. 2020. A Diversity-aware Model for Majority Vote Ensemble Accuracy. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 1310–1320. PMLR. Discusses modeling ensemble accuracy via Condorcet–based assumptions.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. S. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ACM.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 265–284. Springer.

Feldman, M.; Friedler, S. A.; Moeller, C.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Fisher, A.; Rudin, C.; and Dominici, F. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81.

Fisher, J.; Appel, R. E.; Park, C. Y.; Potter, Y.; Jiang, L.; Sorensen, T.; Feng, S.; Tsvetkov, Y.; Roberts, M. E.; Pan, J.; Song, D.; and Choi, Y. 2025. Political Neutrality in AI is Impossible—But Here is How to Approximate it. *arXiv preprint arXiv:2503.05728*.

Gabriel, I.; Ghazavi, A.; Kumar, A.; Mooij, J.; Schroeder, J.; Smith, L. S.; and Weidinger, L. 2024. A matter of principle? AI alignment as the fair treatment of claims. *Philosophical Studies*, 1–31.

Gardner, H. 1983. *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books.

Goleman, D. 1995. *Emotional Intelligence: Why It Matters More Than IQ*. New York: Bantam Books.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Goodhart. 1975. Problems of Monetary Management: The U.K. Experience. In *Papers in Monetary Economics*, volume 1 of *Papers in Monetary Economics*, 1–20. Sydney: Reserve Bank of Australia.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, 1321–1330.

Hardt, M.; Price, E.; and Srebro, N. 2016a. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29, 3323–3331.

Hardt, M.; Price, E.; and Srebro, N. 2016b. Equality of Opportunity in Supervised Learning. arXiv:1610.02413.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudík, M.; and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; Zeng, F.; Ng, K. Y.; Dai, J.; Pan, X.; O'Gara, A.; Lei, Y.; Xu, H.; Tse, B.; Fu, J.; McAleer, S.; Yang, Y.; Wang, Y.; Zhu, S.-C.; Guo, Y.; and Gao, W. 2024. AI Alignment: A Comprehensive Survey. arXiv:2310.19852.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, volume 67, 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30, 4066–4076.

Laufer, B.; Raghavan, M.; and Barocas, S. 2025. What Constitutes a Less Discriminatory Algorithm? In *Proceedings of the Symposium on Computer Science and Law on ZZZ*, CSLAW '25, 136–151. ACM.

Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. arXiv:1811.07871.

Li, X.; Wu, P.; and Su, J. 2022. Accurate Fairness: Improving Individual Fairness without Trading Accuracy. arXiv:2205.08704.

Liu, B.; Liu, X.; Jin, X.; Stone, P.; and Liu, Q. 2021a. Conflict-Averse Gradient Descent for Multi-Task Learning. In *Advances in Neural Information Processing Systems*. Proposes CAGrad to mitigate conflicting gradients by structured regularization.

Liu, L. T.; Simchowitz, M.; and Hardt, M. 2019. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, 4051–4060.

Liu, S.; Lin, Z.; Ma, Y.; Eriksson, B.; and Hsieh, C.-J. 2021b. Conflict–Aware Gradient Descent for Multi–Task Learning. In *Advances in Neural Information Processing Systems*, volume 34.

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

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 220–229. ACM.

Morris, M. W.; and Leung, K. 2000. Justice for all? Progress in research on cultural variation in the psychology of distributive and procedural justice. *Applied Psychology: An International Review*, 49(1): 100–132.

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017a. Exploring Generalization in Deep Learning. arXiv:1706.08947.

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017b. Exploring Generalization in Deep Learning. *NIPS*.

Plecko, D.; et al. 2024. Fairness-Accuracy Trade-Offs: A Causal Perspective. *arXiv preprint arXiv:2405.15443*.

Quiñonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2009. *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press.

Rawls, J. 1985. Justice as Fairness: Political not Metaphysical. *Philosophy & Public Affairs*, 14(3): 223–251.

Rawls, J. 1988. The Priority of Right and Ideas of the Good. *Philosophy & Public Affairs*, 17(4): 251–276.

Raz, J. 1982. Liberalism, Autonomy, and the Politics of Neutral Concern. *Midwest Studies in Philosophy*, 7: 59–94.

Raz, J. 1986. *The Morality of Freedom*. Oxford: Oxford University Press.

Rodolfa, K. T.; Lamba, H.; and Ghani, R. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10): 896–904.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5): 206–215.

Rudin, C.; Zhong, C.; Semenova, L.; Seltzer, M.; Parr, R.; Liu, J.; Katta, S.; Donnelly, J.; Chen, H.; and Boner, Z. 2024. Amazing Things Come From Having Many Good Models. arXiv:2407.04846.

Ruf, B.; and Detyniecki, M. 2021. Towards the Right Kind of Fairness in AI. arXiv:2102.08453.

Salovey, P.; and Mayer, J. D. 1990. Emotional intelligence. *Imagination, Cognition and Personality*, 9(3): 185–211.

Semenova, L.; Rudin, C.; and Parr, R. 2022. On the Existence of Simpler Machine Learning Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1827–1858. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; Althoff, T.; and Choi, Y. 2024a. A Roadmap to Pluralistic Alignment. arXiv:2402.05070.

Sorensen, T.; Sarkar, J.; Goldberg, S.; Anwar, U.; Sikder, R.; Rankin, C. H.; Qiu, X.; Phang, J.; Dhamala, J.; Gururangan, S.; Denton, E.; and Kiela, D. 2024b. A Roadmap to Pluralistic Alignment. *arXiv preprint arXiv:2402.05070*.

Sternberg, R. J. 1985. *Beyond IQ: A Triarchic Theory of Human Intelligence*. Cambridge: Cambridge University Press.

Sternberg, R. J.; and Wagner, R. K. 1986. Practical intelligence. *Nature*, 319(6055): 15–16.

Tyler, T. R.; Boeckmann, R. J.; Smith, H. J.; and Huo, Y. J. 1997. Distributive and procedural justice in seven nations. *Journal of Personality and Social Psychology*, 72(1): 157–169.

U.S. Department of Health and Human Services. 2024. HIPAA Privacy Rule: Uses and Disclosures of Protected Health Information (45 C.F.R. § 164.502). Code of Federal Regulations, Title 45, Part 164, Section 502. Available online: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.502.

Wechsler, D. 1958. *The Measurement and Appraisal of Adult Intelligence*. Baltimore: Williams & Wilkins, 4th edition.

Wick, M.; Panda, S.; and Tristan, J.-B. 2019. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems*, 8783–8792.

Wittgenstein, L. 1953. *Philosophical Investigations*. Blackwell.

Xin, R.; Zhong, C.; Chen, Z.; Takagi, T.; Seltzer, M.; and Rudin, C. 2022. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems*, 35: 14071–14084.

Yu, R.; Yu, T.; Gu, S. S.; Levine, S.; and Finn, C. 2020a. Gradient Surgery for Multi–Task Learning. In *Advances in Neural Information Processing Systems*, volume 33.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020b. Gradient Surgery for Multi-Task Learning. In *Advances in Neural Information Processing Systems*, volume 33, 5824–5836. Introduces PCGrad to alleviate conflicting gradients.

Zafar, M. B.; Valera, I.; Rodríguez, M. G.; and Gummadi, K. P. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proc. 26th International World Wide Web Conference (WWW)*, 1171–1180.

Zhao, H.; and Gordon, G. J. 2022. Inherent Tradeoffs in Learning Fair Representations. arXiv:1906.08386.

Zhong, C.; Chen, Z.; Liu, J.; Seltzer, M.; and Rudin, C. 2023. Exploring and interacting with the set of good sparse generalized additive models. *Advances in neural information processing systems*, 36: 56673–56699.

Ziller, A.; Usynin, D.; Braren, R.; Makowski, M.; Rueckert, D.; and Kaissis, G. 2021. Medical Imaging Deep Learning with Differential Privacy. *Scientific Reports*, 11(13524).