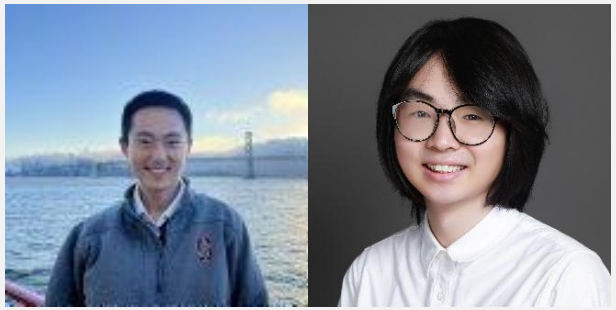
Carnegie Mellon University

NEURAL INFORMATION PROCESSING SYSTEMS

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



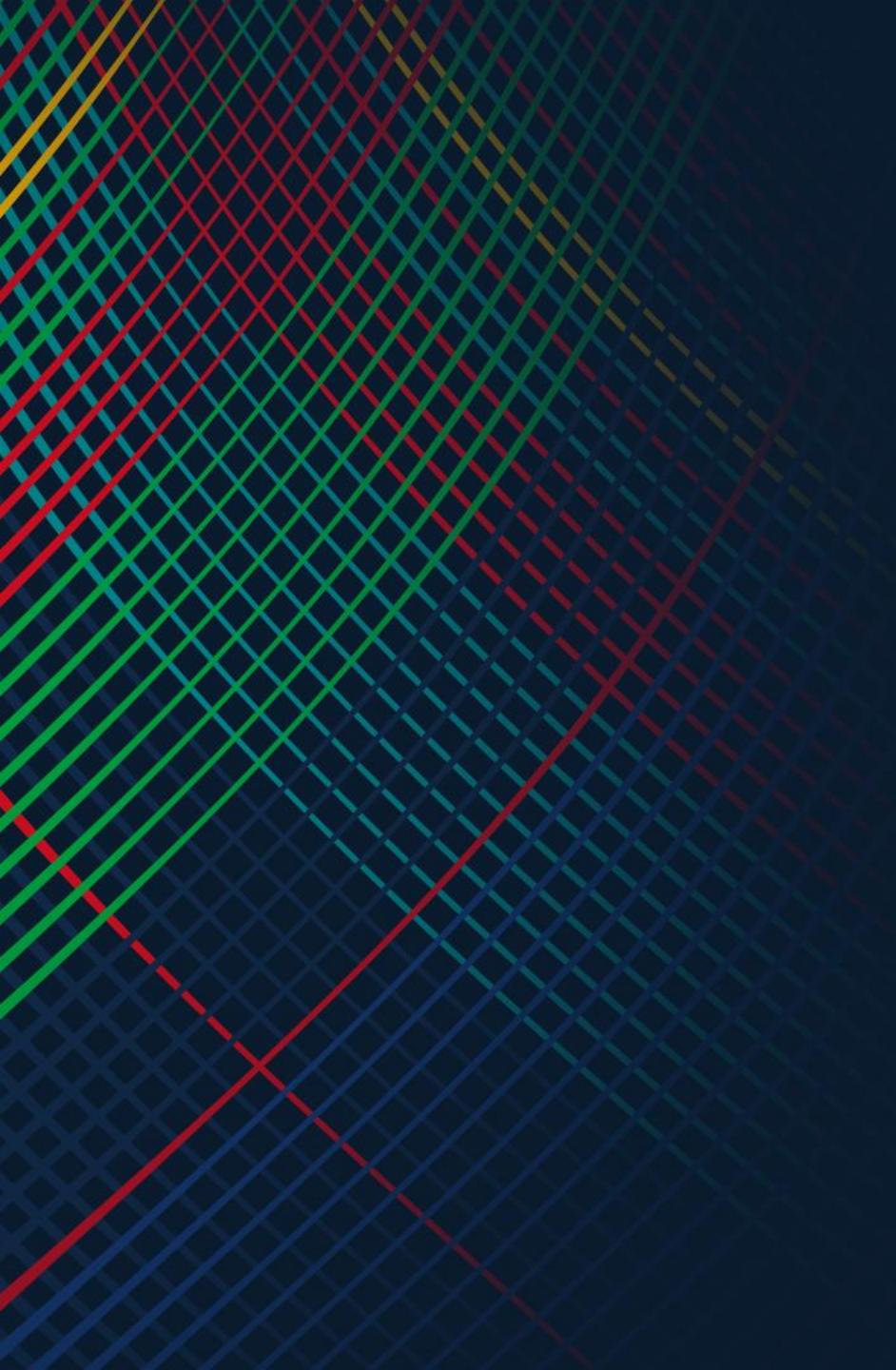
Gordon Dai and Yunze Xiao

NYU

LTI

Agenda

1. Introduction
2. Framework
3. The Value of Inconsistent Metrics
4. Recommendations and Future Directions
5. Response to Alternative View



Introduction

Current Responsible AI Practice

- Fairness Audits
  - Employ demographic parity, equalized odds, or counterfactual consistency to assess fairness.
- Privacy Claims
  - Rely on ε - differential privacy to ensure data privacy in AI systems.
- Robustness test
  - Evaluate models through distribution shift, adversarial risk, and calibration error checks.

The Puzzle of Inconsistency

Mathematically incompatible metrics

Many Responsible AI metrics are mathematically at odds with each other.

Classical impossibility theorems

These theorems highlight the mathematical impossibilities among key metrics.

Tradeoffs between accuracy and privacy

Achieving high accuracy often comes at the expense of privacy.

Tradeoffs in other areas

There are tradeoffs in political neutrality, informativeness, and interpretability.

Conventional vs. Opposite Stance

1. ***Conventional wisdom***

   Treats contradictions as bugs needing fixing by choosing a single metric.

2. ***Opposite stance***

   Argues theoretical inconsistency is a valuable feature, not a flaw.

3. ***Embracing inconsistency***

   Advocates for viewing inconsistencies as beneficial rather than problematic.

Contribution of our work

01.	**Formalizing inconsistencies**	Defines intra - concept inconsistency and inter - concept tradeoffs.
02.	**Synthesizing evidence**	Provides theoretical and empirical support for the value of inconsistent metrics.
03.	**Proposing research agenda**	Shifts focus to characterizing acceptable inconsistency and pluralistic evaluation.

Conceptual Framework

Intra-concept inconsistency

Definition 1: 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

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

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.

Inter-concept inconsistency

Definition 2: Inter-concept tradeoff

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

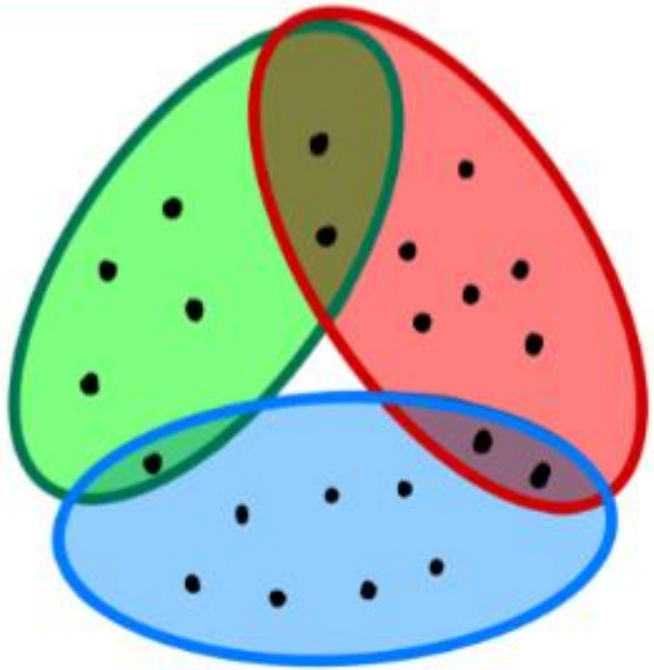
$$\sup_{h \in \mathcal{H}: B(h) \leq b} A(h) < \sup_{h \in \mathcal{H}} A(h) \quad \text{for some } b < \sup_{h} B(h).$$

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.

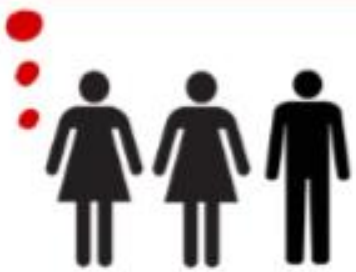
Fairness

- Kleinberg et al. (2016) findings
  - Demonstrated that three core fairness metrics cannot be simultaneously satisfied except in degenerate cases.
- Chouldechova's impossibility theorem (2017)
  - Established the mathematical relationship among PPV, FPR, FNR, and base rates, showing inherent fairness conflicts.
- Bell et al. perspective
  - Argued that approximate fairness is attainable when constraints on metrics are relaxed.

We think **democracy parity** should be the metric for fairness.



No, by our cultural convention, **equalized oddity** is what we mean by fairness!



Hum, why fairness can't be **perfect calibration**?

Political Neutrality

- Rawls' argument
  - Claimed that procedural, aim-based, and effect-based neutrality cannot all be satisfied simultaneously.
- Raz's viewpoint
  - Argued that full or comprehensive neutrality is impossible and should instead be understood as "a matter of degree."
- Fisher et al.'s work
  - Proposed eight practical techniques to approximate political neutrality across different system levels.
  - Assessed nine LLMs and identified trade-offs among utility, safety, fairness, and user autonomy.
  - Emphasized that neutrality is not binary but exists on a continuum with varying achievable degrees.

Why Inconsistent Metrics?

Normative Pluralism

Pluralistic alignment

Acknowledges diverse human values and accommodates them in AI systems.

Diverse value representations

Each metric captures a distinct moral stance from different social groups.

Human value inconsistencies

Highlights that fundamental human values are often pluralistic and conflicting.

Pluralistic alignment benefits

Ensures AI systems consider diverse perspectives, reducing bias and injustice.

Epistemological Completeness

Wittgenstein's analysis

Showed many concepts lack a single set of necessary and sufficient conditions.

Conceptual projection limitations

Formal metrics capture only part of a concept's meaning, sacrificing original content.

Information preservation

Preserving inconsistent metrics helps retain all aspects of a concept's information.

Concept complexity

Concepts like fairness and neutrality are complex, with multiple valid interpretations.

Practical Benefits

Gradient conflict

Misaligned gradients in multi - objective learning prevent overfitting to a single metric.

Rashomon set

Set of near - optimal models offers flexibility and enhances ensemble robustness.

Improved robustness

Joint optimization of conflicting metrics leads to models with better generalization and robustness.

Pareto fronts

Existence of Pareto fronts allows swapping models without sacrificing accuracy for fairness.

Complementary metrics

Simultaneous performance on inconsistent metrics helps models avoid shortcut features.

Recommendations

Practice - Driven Theories

- Gap between theory and practice
  - 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.
- Rashmon set
  - Rashomon set theory studies all near-optimal models; Dai et al. show its asymptotic size grows exponentially with $\sqrt{\varepsilon}$, meaning a small increase in tolerated error yields many more candidate models.
  - This yields a clear rule of thumb for model search: pick the largest error tolerance your business can accept to expand the pool of viable, potentially fairer models, turning a formal result into a concrete selection strategy.

Defining Acceptable Inconsistency

Specifying tolerance ranges	Establishes clear boundaries for acceptable divergence among metrics.
Contextual determination	Considers empirical risk, stakeholder priorities, and deployment specifics.
Evaluating inconsistency impact	Assesses whether inconsistencies enhance generalization or cause failures.
Tradeoff between plurality and consistency	Balances the need for diverse values with avoiding arbitrary tradeoffs.
Benefits of clear standards	Provides clarity for practitioners and ensures ethical accountability.

Documenting Normative Assumptions

Model Cards	Detail intended use cases, evaluation conditions, and ethical considerations.
Data Statements	Document data set creation rationale, demographic coverage, and limitations.
Metric Provenance Sheet	Explain what each metric measures, its limitations, and stakeholder values.
Importance of transparency	Promotes stakeholder trust by making model evaluation criteria clear.
Enhancing interpretability	Helps users understand model performance across different metrics.

SPHERE: An Evaluation Card for Human-AI Systems

Qianou Ma[1]*, Dora Zhao[2]*, Xinran Zhao[1], Chenglei Si[2], Chenyang Yang[1], Ryan Louie[2], Ehud Reiter[3], Diyi Yang[2]†, Tongshuang Wu[1]†

[1]Carnegie Mellon University, Pittsburgh, USA,
[2]Stanford University, Stanford, USA,
[3]University of Aberdeen, Aberdeen, UK

Abstract

In the era of Large Language Models (LLMs), establishing effective evaluation methods and standards for diverse human-AI interaction systems is increasingly challenging. To encourage more transparent documentation and facilitate discussion on human-AI system evaluation design options, we present an evaluation card SPHERE, which encompasses five key dimensions: 1) What is being evaluated?; 2) How is the evaluation conducted?; 3) Who is participating in the evaluation?; 4) When is evaluation conducted?; 5) How is evaluation validated? We conduct a review of 39 human-AI systems using SPHERE, outlining current evaluation practices and areas for improvement. We provide three recommendations for improving the validity and rigor of evaluation practices.

1 Introduction

The proliferation of LLMs has changed the way humans interact with AI systems. Compared to previously existing AI models, LLMs can better comprehend and generate human-like text, enabling

Given the wide diversity of human-AI systems, what factors should researchers consider when designing evaluations? How do we ensure that these evaluations are transparent and replicable? To address these questions, we need systematic methods for documenting how these evaluations are conducted. As a step in this direction, we propose the SPHERE evaluation card, which provides a comprehensive template for designing and documenting evaluation protocols used to assess human-AI systems.[1] Although we focus on systems powered by LLMs in this work, the evaluation dimensions we discuss are agnostic to the type of model and can be applied to AI systems more broadly.

SPHERE Evaluation Card for Human-AI Systems

(Subject) **What** is being evaluated?

- **Component**: Model, System
- **Design Goal**: Effectiveness, Efficiency, Satisfaction

(Process) **How** is the evaluation being conducted?

- **Scope**: Intrinsic, Extrinsic
- **Method**: Quantitative, Qualitative

(Handler) **Who** is participating in the evaluation?

Testing Human - Metric Interaction

Empirical studies	Involve users, experts, and regulators in metric selection and negotiation.
Participatory methods	Effectively capture diverse fairness notions and stakeholder priorities.
Interactive interfaces	Enable participants to navigate algorithmic tradeoffs and make informed decisions.
Insights from studies	Inform the design of intuitive interfaces and inclusive optimization strategies.
Importance of stakeholder input	Ensures AI systems reflect the needs and values of diverse stakeholders.

Hire Your Anthropologist!
Rethinking Culture Benchmarks Through an Anthropological Lens

Mai AlKhamissi[1*] **Yunze Xiao**[1*] **Badr AlKhamissi**[2] **Mona Diab**[1]

[1]Carnegie Mellon University [2]EPFL

{malkhami,yunzex,mdiab}@andrew.cmu.edu badr.alkhamissi@epfl.ch

Abstract

Cultural evaluation of large language models has become increasingly important, yet current benchmarks often reduce culture to static facts or homogeneous values. This view conflicts with anthropological accounts that emphasize culture as dynamic, historically situated, and enacted in practice. In this position paper, we qualitatively examine 20 handpicked cultural benchmarks to show the breadth of how culture is framed within NLP. First, we introduce a four-part framework that categorizes how benchmarks frame culture, such as *knowledge*, *preference*, *performance*, or *bias*. Using this lens and identify six recurring methodological issues: including treating countries as cultures, overlooking within-culture diversity, and relying on oversimplified survey formats. Drawing on established anthropological methods, we propose concrete improvements: incorporating real-world narratives and scenarios, involving cultural communities in design and validation, and evaluating models in context rather than isolation. Our aim is to guide the creation of cultural benchmarks that move beyond simple recall tasks and more faithfully reflect how models respond to complex cultural situations.

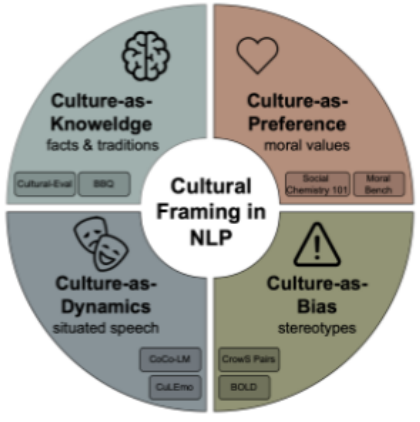


Figure 1: **Cultural Framing in NLP.** Our taxonomy of how culture is framed in NLP evaluation. Each quadrant represents a distinct theoretical lens on culture: defining what it entails, illustrating how it is expressed, and providing two representative benchmarks for each framing.

appropriately to regionally specific norms, moral frameworks, idioms and socio-political identities.

However, in this growing body of work, the concept of *culture* is often treated as a background variable rather than as a central analytical concern. Benchmark designers rarely engage with anthropology, a discipline that has developed rich tools to un-

Response to alternative view

Confusing Users and Regulators

- A single metric may oversimplify complex Responsible AI concepts.
- Multiple metrics are needed to represent diverse perspectives fully.
- Empirical results show tolerance - based evaluation is practical and effective.

Diversity vs. Inconsistency

- Plurality does not necessitate inconsistency but **often** leads to it.
- Attempts to force consistency can sacrifice valuable diversity and plural representation.
- Inconsistent metrics often reflect the real - world complexities of diverse stakeholder values.
- Preserving diverse perspectives is crucial for fair and inclusive AI systems.

Choosing the Best Metric?

- the importance of choosing metrics suited to specific application tasks.
- However, there are cases where diverse stakeholder perspectives require multiple metrics.
- Optimizing a single metric can lead to it losing its effectiveness over time.
    - Preserving inconsistent metrics helps mitigate the negative effects of Goodhart's Law.

Thank you for listening!