Forschungsinstitut Cyber Defence
Universität der Bundeswehr München

Universität der Bundeswehr München

On the Tensions and Trade-Offs of Algorithmic Fairness

From Metric Incompatibility to Cross-Pillar Trade-Offs

Eirini Ntoutsi

AIML

European Conference on Algorithmic Fairness, 3 September 2026, Ghent

ECAF 2026

Algorithmic fairness: a rich and evolving field

Substantial progress but fundamental tensions remain

- Since the seminal paper by Pedreschi et al. (KDD 2008), algorithmic fairness research has expanded rapidly:
  - Multiple fairness definitions
  - Detection and mitigation methods across the ML pipeline
  - Fairness-aware methods across tasks and modalities
  - Growing practical and regulatory relevance
- Yet, **fundamental tensions remain insufficiently understood**. They arise:
  - Within algorithmic fairness itself
  - Across Responsible AI pillars
  - Dynamic environments may further amplify these tensions (an open challenge)
- The goal is to show that these tensions are **not one kind of problem**: some are provable, some are value choices we have not stated, some are empirical limits we have not moved, and that **each kind calls for a different response.**

Roadmap

- **Part 1: Within-fairness tensions**
  1. **Metric incompatibility:** Can different fairness criteria be satisfied simultaneously?
  2. **Granularity:** Whose fairness do we assess: groups or individuals?
  3. **Representation:** How do representation choices determine which harms become visible?
  4. **Composition:** Does fairness for individual attributes extend to their intersections?
  5. **Estimation:** Can small groups be assessed reliably?
- **Part 2: Cross-pillar tensions**
  6. **Fairness - Performance:** Must fairness reduce predictive performance?
  7. **Fairness – Data robustness:** Does fairness hold under ordinary data imperfections?
  8. **Fairness - Privacy:** Can we protect sensitive information under fairness?
  9. **Fairness - Adversarial robustness:** Fair for whom under adversarial perturbations?
- **Part 3: Not all tensions are equal – Implications for solutions**

Roadmap

- **Part 1: Within-fairness tensions**
  1. **Metric incompatibility:** Can different fairness criteria be satisfied simultaneously?
  2. **Granularity:** Whose fairness do we assess: groups or individuals?
  3. **Representation:** How do representation choices determine which harms become visible?
  4. **Composition:** Does fairness for individual attributes extend to their intersections?
  5. **Estimation:** Can small groups be assessed reliably?
- **Part 2: Cross-pillar tensions**
  6. **Fairness - Performance:** Must fairness reduce predictive performance?
  7. **Fairness – Data robustness:** Does fairness hold under ordinary data imperfections?
  8. **Fairness - Privacy:** Can we protect sensitive information under fairness?
  9. **Fairness - Adversarial robustness:** Fair for whom under adversarial perturbations?
- **Part 3: Not all tensions are equal – Implications for solutions**

Tension 1 – Metric incompatibility

Mathematical impossibility of group fairness

- **Some group-fairness criteria are mutually incompatible**
  - Chouldechova, Big Data 17: A calibrated classifier cannot simultaneously satisfy predictive parity and equalized odds when base rates differ across groups.
  - Kleinberg et al, ITCS 17: Calibration, balance for the positive class, and balance for the negative class cannot all hold simultaneously.
  - Pleiss et al, NeurIPS 17: Equalized odds remain incompatible with calibration.
- **Exact compatibility** is possible only under restrictive conditions, such as equal base rates or perfect predictions.

Responses in the field

Single-criterion optimization: Select the fairness definition that addresses the most relevant harm and optimize for it.
(Hardt et al, NeurIPS 16; Zafar et al, AISTATS 17; Agarwal et al, ICML 18)

Multi-criteria balancing and relaxation: Relax exact constraints to acceptable tolerances, or navigate trade-offs across multiple criteria simultaneously
(Pleiss et al, NeurIPS 17; Padh et al, UAI 21, Swati et al, CIKM 25).

➔ Exact satisfaction of all criteria is generally impossible: what to prioritise, how much to relax, and which point to select are normative, context-dependent decisions, not merely technical ones

AIML

Tension 2 – Granularity: Groups are not individuals

Parity in aggregate can coexist with inconsistent treatment of individuals

Whose fairness do we assess: groups or individuals?

Group fairness

- Are groups treated similarly?
- Detects disparities between protected groups.
- Can hide disparities within groups.

Individual fairness (Dwork et al, ITCS 12)

- Are similar individuals treated similarly?
- Detects inconsistent treatment at the person level.
- Requires a defensible notion of similarity.

Satisfying one perspective does not guarantee the other

Responses in the field

Choose one level (group or individual): Select the level that corresponds to the most relevant harm.
(Most fairness work focuses on groups.)

Address both levels: Jointly assess, and mitigate, group disparities and individual inconsistency.
(Speicher et al, KDD 18; Gorantla et al, AIES 23; Han et al, WWW 23).

➔ Neither perspective is sufficient: the chosen level determines which harms become visible and which remain hidden. The choice, one level or both, depends on context and should be stated.

AIML

Tension 3 – Representation choices shape fairness

Fairness can be shaped before learning begins by how people, attributes, and modalities are represented in data.

- **Which groups and harms does the chosen representation make visible?**
  - **Defining protected groups (**Le Quy et al, WIDM 22): Simplified categories can exclude identities or conceal differences
    - Gender, Race/Ethnicity, Age are often simplified because of data availability or technical convenience.
  - **Feature representation bias** (Mougan et al, AIES 23; Panagiotou et al, ICAIF 24): encoding choices and feature types can produce different fairness outcomes
    - One-hot and target encoding can produce different fairness outcomes, esp. for underrepresented categories
    - Methods can exhibit bias toward specific feature types, e.g. favouring numerical over categorical features
  - **Modality bias** (Swati et al, ECAF 24): Disparities can differ across text, images and tabular data; fusion can amplify or mitigate them.

Responses in the field

Make representation choices explicit: Define and document groups, attributes and modalities, including omissions and limitations.

Test representation sensitivity: Compare alternative group definitions, encodings, feature treatments, and fusion strategies.

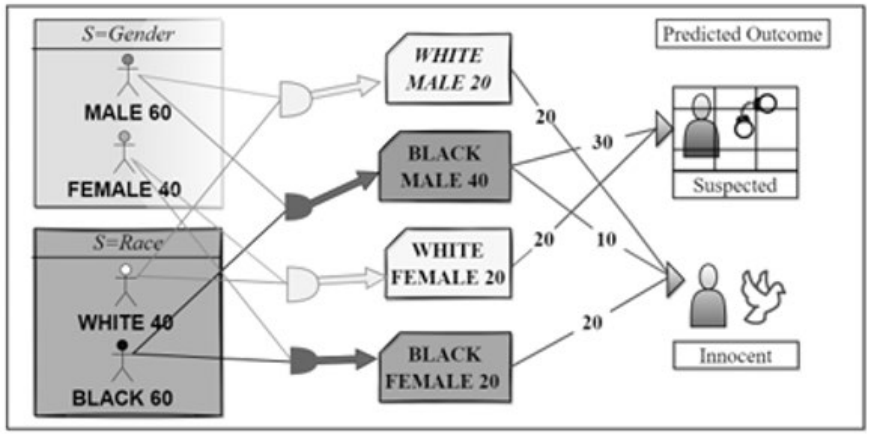
➔ Fairness conclusions are conditional on the chosen representation: metrics reveal only the groups and harms made visible by the data.

AIML

Tension 4 – Composition: Fairness does not compose

Marginal fairness can conceal severe intersectional harm

- **Does fairness for individual attributes extend to their intersections?**
  - Individuals belong to multiple groups simultaneously (e.g., black women over 50)
  - Fairness on each attribute separately does not guarantee fairness at their intersections.
  - Fairness gerrymandering (Kearns et al, ICML 18): A model can satisfy marginal fairness constraints while treating intersectional groups very differently.



Prediction distribution of a hypothetical drug trafficker detection model for different population (sub)groups (Roy et al, FAccT 23)

Marginal suspected-decision rates:

- MALE: $30/60 = 50\%$; FEMALE: $20/40 = 50\%$
- WHITE: $20/40 = 50\%$; BLACK: $30/60 = 50\%$

Intersectional suspected-decision rates:

- WHITE FEMALE: 20/20=100%;
- BLACK FEMALE: 0/20=0%;
- BLACK MALE: 30/40=75%;
- WHITE MALE: 0/20=0%

Responses in the field

Focus on intersections: Assess and address fairness violations for joint protected groups, not just individual attributes (Kearns et al, ICML 18; Roy et al, FAccT 23, Roy et al., DS 22, Swati et al, CIKM 25)

AIML

➔ Marginal fairness does not guarantee intersectional fairness, and the gap can be exploited.

Tension 5 – Estimation: Finer groups less reliable estimates

The number of subgroups grows exponentially while evidence per subgroup shrinks.

- **How much finer can we go?**

k binary protected attributes

2^k

possible intersections

Sparse evidence

Small or even empty subgroups

High uncertainty

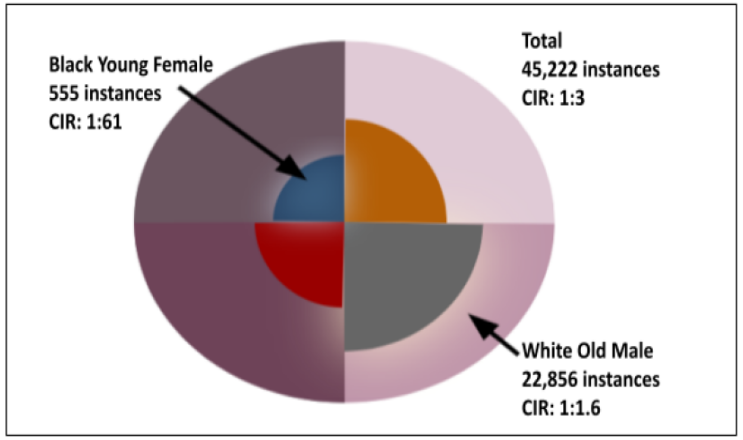
Unstable fairness estimates

Compounded imbalance

The positive class within a protected subgroup can become extremely rare

Undefined baseline

Fair compared to whom? worst subgroup (Ghosh et al, AIDBEI 21), overall population (Kearns et al, ICML 18), ..



Adult income prediction dataset (Roy et al, FAccT 23)
CIR = class imbalance ratio

Responses in the field

Quantify uncertainty: Report subgroup sizes, confidence intervals and estimation uncertainty. Account for multiple comparisons.
(Cherian & Candes, JMLR 24)

Strengthen the evidence: Use targeted data collection, and smoothed or regularized estimators for small subgroups
(Chen et al, NeurIPS 18; Foulds et al, ICDE 20)

AIM

➔ More granular auditing can reveal more relevant harms, but estimates degrade as groups shrink. The worst-subgroup baseline captures intersectional harm best but measures it least reliably.

Within-fairness tensions: Different kinds of problems

Tensions differ in kind, and so do the responses

UNAVOIDABLE: provable → choose and justify
CONTESTED: a value choice → make it explicit
EMPIRICAL: evidence or methods are the limit → can be improved

1	**Metric incompatibility**	Different metrics can impose incompatible requirements.	UNAVOIDABLE
2	**Granularity**	Group and individual perspectives capture different harms.	CONTESTED
3	**Representation**	Representation choices determine which groups and harms are visible.	CONTESTED EMPIRICAL
4	**Composition**	Marginal fairness does not guarantee intersectional fairness.	UNAVOIDABLE
5	**Estimation**	Sparse subgroups make intersectional estimates uncertain. Undefined baselines.	EMPIRICAL CONTESTED

→ Fairness is not one property to optimize. These tensions differ in kind, which determines what can be done about each.

AIML

Responsible AI is multidimensional

Responsible AI requires multiple objectives to be addressed throughout the system lifecycle.

- Fairness concerns who benefits, who is harmed and how outcomes are distributed.
- AI systems must also support
  - Privacy, Robustness, Transparency,
  - Human agency, Accountability,
  - Societal and environmental well-being,
  - …
- These requirements/pillars are interconnected, but improving one does not necessarily improve the others.

Part 1 examined tensions within fairness. Part 2 changes the level of analysis: fairness now interacts with other Responsible AI pillars.

High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, 2019.

! The seven requirements provide an organising framework, they do not necessarily exhaust all Responsible-AI objectives.

Roadmap

- **Part 1: Within-fairness tensions**
  1. **Metric incompatibility:** Can different fairness criteria be satisfied simultaneously?
  2. **Granularity:** Whose fairness do we assess: groups or individuals?
  3. **Representation:** How do representation choices determine which harms become visible?
  4. **Composition:** Does fairness for individual attributes extend to their intersections?
  5. **Estimation:** Can small groups be assessed reliably?
- **Part 2: Cross-pillar tensions**
  6. **Fairness - Performance:** Must fairness reduce predictive performance?
  7. **Fairness – Data robustness:** Does fairness hold under ordinary data imperfections?
  8. **Fairness - Privacy:** Can we protect sensitive information under fairness?
  9. **Fairness - Adversarial robustness:** Fair for whom under adversarial perturbations?
- **Part 3: Not all tensions are equal – Implications for solutions**

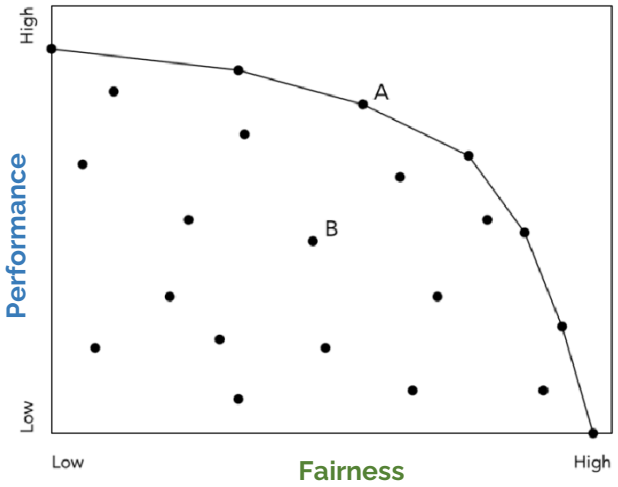
Tension 6 – Fairness vs Performance

Fairness-aware learning must decide how predictive performance and fairness are jointly optimized

- **How do we jointly optimize fairness and performance?**
- **Common approach: linear scalarization**

$$\theta^*(\lambda) = \arg\min_\theta \ (\mathcal{L}_{perf}(\theta) + \lambda \mathcal{L}_{fair}(\theta))$$

  - A fixed λ selects one fairness-performance compromise and encodes their relative priority
  - Varying λ yields a set of candidate frontier points
  - Linear scalarization may miss Pareto-optimal points in non-convex regions of the frontier.
- **Beyond linear scalarization**
  - **Chebyshev scalarization:** Can recover Pareto-optimal points in non-convex regions (Wei & Niethammer, ASA Data Sci. J. 22)
  - **Fairness constraints:** Select the best-performing model satisfying a fairness tolerance ε (Padala & Gujar, IJCAI 20)
  - **Adaptive selection:** Change priorities across tasks or during training (Roy & Ntoutsi, ECML PKDD 22)



- Each point is a feasible model.
- The frontier is the set where no objective can improve without another getting worse.
- Moving toward it is better optimisation; moving along it is a trade-off.

➔ Methods differ in how they search for and select feasible compromises; they do not establish whether the frontier itself can be moved.

AIML

Tension 6 cont. – Fairness vs Performance

Is the fairness-performance trade-off intrinsic?

- **Yes, within a fixed prediction setting** (Zhao & Gordon, JMLR 22)
  - For a fixed data distribution, fairness criterion and performance measure:
    - some fairness requirements impose an unavoidable loss in predictive performance.
    - no better learning algorithm can move beyond this frontier while the setting remains fixed
- **Not necessarily, if the setting can change** (Dutta et al., ICML 20)
  - The observed trade-off may be a symptom of data inequality or informativeness between groups (e.g., noisier representations for the unprivileged group, historic differences in representation and opportunity)
  - More informative features for the disadvantaged group can raise its class separability and move the frontier
  - The trade-off is a property of the data, not of fairness itself

➔A trade-off can be unavoidable within a fixed setting without being intrinsic to fairness itself: ask whether better optimisation can reach the frontier, or better data can move it.

Tension 6 cont. – Fairness conflicts in multi-task learning

Shared representations can transfer both knowledge and bias

- **([Dutta et al., ICML 20]) showed that data can move the frontier; the learning setup can too.**
  - Multi-task learning: Multiple tasks update shared model parameters.
- **Two forms of cross-task conflicts can arise:**
  - Accuracy-gradient conflicts → negative transfer: joint learning reduces a task's predictive performance.
  - Fairness-gradient conflict → bias transfer: joint learning increases a task's fairness violation.

(a) TAG (b) Recon

Conflict: angle > 90° between task gradients. Fairness-loss gradient conflicts persist in TAG (task grouping) and Recon (gradient correction) on ACS-PUMS. - negative transfer correction alone does not resolve fairness conflicts

- **Possible response: FairBranch** (Roy et al, IJCNN 24)
  - Group tasks according to parameter similarity and create separate branches for different task groups.
  - Correct fairness-gradient conflicts within compatible branches.

➔ The fairness–performance balance depends not only on objective weighting, but also on the architecture: which tasks share parameters and whether their updates are compatible.

AIML

Tension 7 – Fairness vs Data robustness

Data limitations are rarely group-neutral

- **Does fairness hold under ordinary data imperfections?**

Data limitation	Potential group effect
Representation imbalance	Fewer observations for minority and intersectional groups
Class imbalance	Rare outcomes may be learned poorly within particular groups (Roy et al., DS 22)
Feature noise	Identical noise produces different group losses, driven by differences between group distributions (Khani & Liang, ICML 20)
Missing values	Missingness rates and patterns vary systematically by group; impute-then-classify worsens the achievable fairness–accuracy frontier (Feng et al., NeurIPS 23)
Label noise	Instance-dependent label noise can make fairness-aware methods more biased than fairness-unaware ones (Wu et al., CLeaR 22)
Noisy protected attributes	Fairness for noisy group labels does not guarantee fairness for the true groups (Wang et al., NeurIPS 20)

➔ Data quantity and quality jointly determine whose performance and whose fairness remains reliable.

AIML

Tension 7 cont. – Fairness vs Data robustness

Fairness interventions designed for clean data may not remain valid under realistic imperfections

- **Possible responses**
  - **Evaluate under perturbation:** stress-test performance and fairness across plausible corruptions, not only on the clean dataset
  - **Report degradation separately for groups and intersections**
  - **Preserve informative missingness:** encode the missing pattern as a feature instead of imputing it away
  - **Model uncertainty in labels and group membership:** treat group labels as uncertain rather than exact

➔ Fairness should be treated as a robustness property: conclusions must remain credible under ordinary data imperfections (before any adversary is involved – see Tension 9).
➔Unlike the other cross-pillar tensions, this is not a trade-off: better data improves fairness and performance together.

AIML

Tension 8 – Fairness vs Privacy

Fairness assessment requires the group data that privacy would minimise

- **Each objective creates a cost for the other:** auditing for fairness expands what must be collected and retained; protecting privacy removes the evidence fairness depends on.

Fairness needs visibility (*Andrus et al., FAccT 21*)

- Auditing and mitigation require protected attributes linked to outcomes
- Granular information can reveal intersectional harms.

Privacy limits exposure

- Sensitive attributes create risks of disclosure, inference and misuse.
- Granular data can increase re-identification risks.

Responses in the field

Govern protected attributes: Use purpose limitation, restricted access, documentation, retention limits and deletion rules.

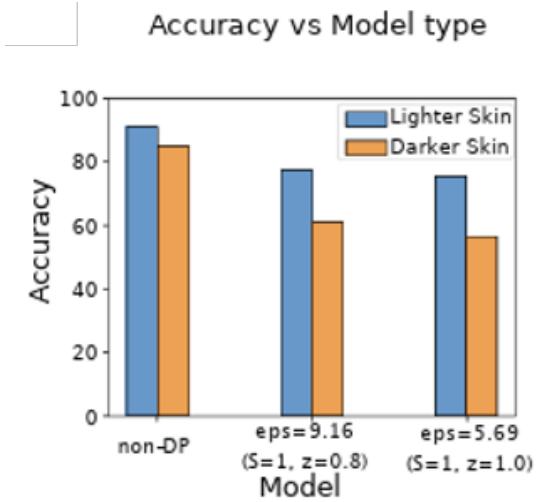
Enable controlled auditing: Use trusted auditors, secure computation and privacy-preserving aggregate statistics.

→ The goal is not demographic unawareness, but controlled access to demographic information for legitimate fairness assessment.
→ But access is only half the problem: **the mechanisms that protect privacy are not themselves group-neutral.**

AIML

Tension 8 cont. – Privacy and fairness impose costs on each other

- **Does equal privacy protection imply equal effects across groups?**
- **Observation: unequal group effects** (Bagdasaryan et al, NeurIPS 19)
  - Differentially private learning can reduce accuracy disproportionately for underrepresented groups.
  - Privacy-preserving noise can also make small-group fairness estimates less reliable.
- **Why can this happen?**
  - **Gradient clipping:** limits large gradients, potentially reducing the influence of underrepresented or difficult examples.
  - **Noise addition:** can obscure the weaker statistical signal available for smaller groups.
- **The reverse also holds: fairness constraints increase privacy risk** (Chang & Shokri, EuroS&P 21)
  - Fair models memorise unprivileged subgroups, raising their exposure to membership inference



Under differential privacy both groups lose accuracy, but the darker-skin group loses more, widening the gap as stronger privacy is enforced (smaller eps)

→ Privacy and fairness impose costs on each other, and both costs concentrate on the smallest groups.

AIML

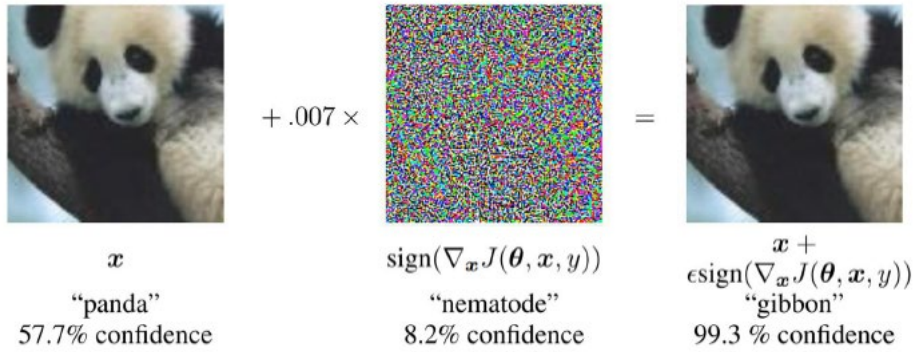
Tension 8 cont. – Is the privacy–fairness conflict intrinsic?

- **Exact compatibility is not guaranteed** (Cummings et al, UMAP 19)
  - Even with full access to the data distribution, there are cases where differential privacy rules out exact equality of false positives and false negatives.
- **Approximate joint guarantees are possible** (Cummings et al, UMAP 19; Jagielski et al, ICML 19)
  - Relax exact equality to a bounded fairness violation.
  - Learn models that provide differential privacy while approximately satisfying fairness constraints.
- **Responses in the field**
  - Specify acceptable privacy and fairness tolerances rather than demanding exact satisfaction.
  - Make the three-way trade-off with accuracy explicit (and justify these choices)

→ Privacy and fairness are not universally incompatible; compatibility depends on their definitions and accepted tolerances.

Tension 9 – Fairness vs Adversarial Robustness

Fairness under normal conditions does not guarantee fairness under attack



$x' = x + \delta \quad with \quad \|\delta\|_p \leq \varepsilon$

A deliberately constructed, bounded perturbation changes the model's prediction while the image remains visually similar.

Two traditionally separate objectives

Fairness

Are outcomes or errors distributed equitably across groups?

Adversarial robustness

Do predictions remain correct under permitted adversarial perturbations?

- **Bidirectional effects**

Adversarial robustness → Fairness

Adversarial training and attacks may produce unequal clean and robust performance across protected groups (Nanda et al, FAccT 21; Ramanaik et al., BIAS 24).

Fairness → Adversarial robustness

Fairness interventions may change decision boundaries and increase adversarial vulnerability in some settings (Tran et al., IJCAI 24).

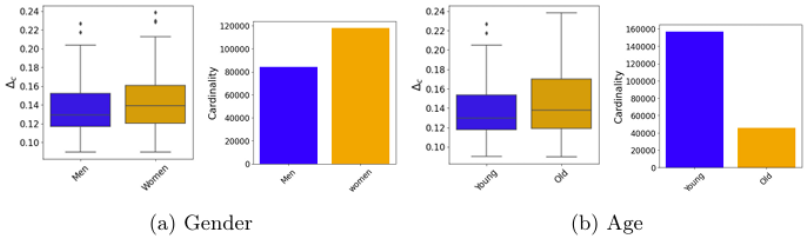
➔ Optimizing fairness or robustness in isolation can create unintended costs for the other.

AIML

Tension 9 cont. – Adversarial robustness is not group-neutral

Aggregate robustness can conceal unequal group vulnerability

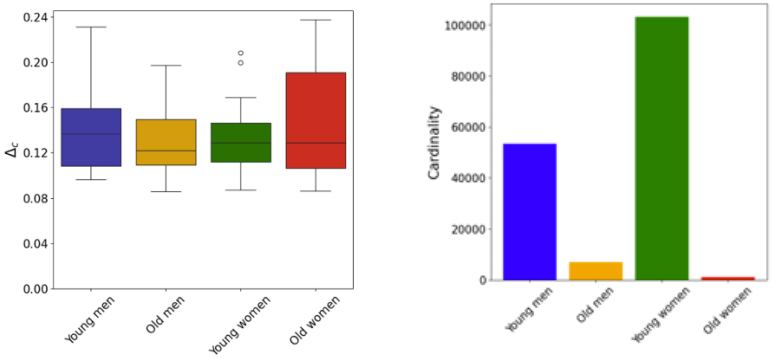
- **Adversarial vulnerability can differ across groups**
  - **Across protected groups** (Nanda et al., FAccT 21)
    - Finds unequal susceptibility across race and gender groups
    - Some demographic groups are systematically closer to the decision boundary.
  - **Across intersectional groups – VAE robustness** (Ramanaik et al., BIAS 24)
    - Older women exhibit higher and more variable adversarial damage than other age-gender groups
    - Subgroup size contributes to vulnerability but does not fully explain it.
    - Adversarial perturbation shifts the embedding, so the reconstruction is pulled toward surrounding majority-group embeddings (old faces reconstructed as young)



(a) Gender (b) Age

Marginal groups: gender and age

Higher Δc → lower robustness



(c) $\beta = 10$ (d) Subgroup cardinality

Intersectional groups: age x gender

➔ Adversarial robustness can vary across protected groups and their intersections, and these disparities remain hidden by aggregate evaluation.

AIML

Tension 9 cont. – Can fairness and robustness be jointly addressed?

Fairness and adversarial robustness interact in both directions, motivating joint evaluation and design

- **Responses in the field**
  - **Joint evaluation:** measure clean and adversarial performance for every group and intersection, not in aggregation (Nanda et al., FAccT 21; Ramanaik et al., BIAS 24)
  - **Joint adversarial training:** defend against perturbations targeting both predictive accuracy and group fairness (Chai et al., ICML 25)
  - **Margin-preserving mitigation:** bound the loss so fairness constraints do not shrink the decision margin (Ramp loss -Tran et al., IJCAI 24)

➔ Fairness and robustness are not intrinsically incompatible: depending on the robustness notion and learning setup, strengthening one may harm or help the other, so their effects should be evaluated jointly across groups.

Cross-pillar tensions: different kinds of problems

Fairness interacts with other Responsible AI pillars

UNAVOIDABLE: provable → choose and justify
CONTESTED: a value choice → make it explicit
EMPIRICAL: evidence or methods are the limit → can be improved

6	**Fairness - Performance**	A frontier exists in a fixed setting, but better data and methods can move it	UNAVOIDABLE (in a fixed setting) EMPIRICAL (if setting can change)
7	**Fairness – Data robustness**	Not a conflict of objectives: poor data degrades both performance and fairness, unevenly across groups.	NOT A TRADE-OFF
8	**Fairness - Privacy**	Exact compatibility is not guaranteed; bounded violations of fairness are possible.	UNAVOIDABLE (on exactness) CONTESTED (at tolerance)
9	**Fairness – Adversarial robustness**	Direction and magnitude depend on the robustness notion and the learning setup	EMPIRICAL

➔ Cross-pillar tensions differ in kind: some are provable, some depend on tolerances, and one is not a trade-off at all

AIML

Roadmap

- **Part 1: Within-fairness tensions**
  1. **Metric incompatibility:** Can different fairness criteria be satisfied simultaneously?
  2. **Granularity:** Whose fairness do we assess: groups or individuals?
  3. **Representation:** How do representation choices determine which harms become visible?
  4. **Composition:** Does fairness for individual attributes extend to their intersections?
  5. **Estimation:** Can small groups be assessed reliably?
- **Part 2: Cross-pillar tensions**
  6. **Fairness - Performance:** Must fairness reduce predictive performance?
  7. **Fairness – Data robustness:** Does fairness hold under ordinary data imperfections?
  8. **Fairness - Privacy:** Can we protect sensitive information under fairness?
  9. **Fairness - Adversarial robustness:** Fair for whom under adversarial perturbations?
- **Part 3: Not all tensions are equal – Implications for solutions**

What kind of problem is fairness?

Tensions sorted by what makes them hard

UNAVOIDABLE: provable → choose and justify
CONTESTED: a value choice → make it explicit
EMPIRICAL: evidence or methods are the limit → can be improved

1	**Metric incompatibility**	UNAVOIDABLE	Choose a criterion and justify it
4	**Composition**	UNAVOIDABLE	Look at intersections, marginal constraints are not enough
8	**Fairness - Privacy**	UNAVOIDABLE (on exactness) CONTESTED (at tolerance)	Specify tolerances, not exact guarantees
2	**Granularity**	CONTESTED	State which harm you are targeting
3	**Representation**	CONTESTED & EMPIRICAL	Document choices, test sensitivity
5	**Estimation**	EMPIRICAL & CONTESTED	Report subgroup sizes and uncertainty
6	**Fairness – Performance**	UNAVOIDABLE (in a fixed setting) EMPIRICAL (if setting can change)	Ask whether better data moves the frontier
9	**Fairness – Adversarial robustness**	EMPIRICAL	Evaluate both per group and intersection
7	**Fairness – Data robustness**	NOT A TRADE-OFF	Fix the data. Both improve together

➔ The impossibility results force a choice (they do not necessarily disappear when relaxed). Everything else is a value choice we have not stated, or an empirical limit we have not moved.

AIML

The tensions interact and compound

Addressing each tension in isolation is not just incomplete — it can make others worse

- **Finer subgroups make every tension harder**
  - The finer the groups, the more harm becomes visible and the harder every other tension becomes. Composition (T4), estimation (T5), privacy (T8) and adversarial vulnerability (T9) all worsen as groups shrink.
- **Adversarial robustness (T9) interacts with composition (T4)**
  - Disproportionate vulnerability at intersections
- **Fairness constraints raise privacy exposure (T8)**
  - Protecting a subgroup through fairness constraints increases its membership-inference risk
- **…**

➔ These interactions have been characterised at most pairwise. Their combined effect remains unclear.

AIML

Dynamic environments may further amplify these tensions

Fairness at training time does not imply fairness at deployment

- Most fairness research assumes a **static world**: fixed data, fixed model, one-shot evaluation
- **The real world is dynamic:**
  - Distributional shift: group compositions and base rates evolve over time
  - Feedback loops: deployed models change the data they observe
  - Temporal degradation: fairness results at training time may not transfer to deployment
- Dynamic environments may not only amplify existing tensions — they may introduce new ones
- **Open questions:**
  - How do we study fairness in dynamic systems (D'Amour et al., FAccT 20; Iosifidis & Ntoutsi, CIKM 19)?
  - How do intersectional fairness violations evolve under distribution shift?
  - What new tensions emerge in dynamic settings (Heidari et al., ICML 19)
  - How do we design systems that remain fair as societal notions of fairness evolve?

➔Dynamic settings are were all these tensions need to be re-examined.

Fairness is not one property to optimize

And its tensions are not one kind of problem to solve

- **Where the tensions are UNAVOIDABLE, a choice is forced:**
  - Which harm to prioritize when criteria conflict (T1)
  - Whether to require fairness at intersections, not just marginals (T4)
  - How much fairness to relax to keep privacy guarantees (T8)
- **Where they are CONTESTED, the choice is ours to state:**
  - Which groups to make visible, through representation and evaluation scope (T2, T3)
- **Where they are EMPIRICAL, the choice can be revised:**
  - How much accuracy to trade for fairness (T6) - the frontier can move
  - Whose robustness to protect under adversarial conditions (T9) - direction depends on the setup
  - How reliably we can measure small subgroups (T5) - better data and estimators help
- **The field has been making (some of) these choices implicitly, inside algorithms, datasets, and evaluation protocols.**

➔ Moving forward: Report the choices, justify the normative ones, and distinguish proven impossibilities from measured limits.

AIML

Fairness is not one property to optimize

And its tensions are not one kind of problem to solve

Thank you for your attention! I look forward to the exchange

Prof. Dr. Eirini Ntoutsi

Chair for Open Source Intelligence
Artificial Intelligence & Machine Learning (AIML) Lab
Institute for Data Security INF7
Research Institute CODE & UniBw Munich

Email: eirini.ntoutsi@unibw.de
www; https://www.unibw.de/aiml || https://aiml-research.github.io/

STELAR
HEGEMON Holistic Evaluation of Generative Foundation Models in a Security Context
DZdA Deutsches Zentrum für digitale Aufgaben
MAMMOth Multi-attribute, Multimodal Bias Mitigation in AI Systems
cyberagentur
Stiftung Innovation in der Hochschullehre
LernMINT
NoBIAS
BIAS
Funded by the European Union
Niedersachsen
DFG Deutsche Forschungsgemeinschaft
Alexander von HUMBOLDT STIFTUNG
AIML
VolkswagenStiftung