Quantifying the Trade-Offs between Dimensions of Trustworthy AI - An Empirical Study on Fairness, Explainability, Privacy, and Robustness

Nils Kemmerzell (✉) and Annika Schreiner

Friedrich-Alexander-Universität Erlangen-Nürnberg
{nils.kemmerzell,annika.schreiner}@fau.de

The following tables present the extended experiment results from our paper, including all metric scores, dimension scores and model accuracy scores.
Significance on all tables is denoted as follows: *: $p \leq 0.05$.
Abbreviations in the results tables are used as described in Table 1.

Table 1. Abbreviations used in the result tables

Long Form	Abbreviation
Baseline	BL
Weighting	W
Constraint Optimization with $\alpha = 1$	α_1
Constraint Optimization with $\alpha = 100$	α_{100}
Patch Gaussian	PG
Membership Inference Attack	MInf
Model Extraction Attack	ModExt
Performance drop on shifted dataset, simulated through random perturbations	(Shift)

Table 2. Extended Trade-offs Explainability

	UTKFace			CelebA			FairFace			LFWA+		
Name	BL	λ_{cos}=0.1	λ_{cos}=1	BL	λ_{cos}=0.1	λ_{cos}=1	BL	λ_{cos}=0.1	λ_{cos}=1	BL	λ_{cos}=0.1	λ_{cos}=1
Explainability Score	6.21	8.51*	8.39*	5.45	7.87*	6.68*	5.99	7.29*	6.00	5.42	7.85*	7.68*
Robustness	9.65	9.99*	9.99*	9.67	9.99*	9.99*	8.93	9.99*	9.99*	9.85	9.99*	9.99*
Faithfulness	2.35	10.00*	10.00*	1.16	10.00*	5.06*	2.10	7.17*	2.92	1.42	10.00*	10.00*
Complexity	6.28	6.59*	6.59*	8.04	7.83*	7.08*	6.22	6.64	6.66	8.51	8.41	8.33
Randomisation	6.55	7.46*	6.96	2.92	3.66*	4.60*	6.72	5.34	4.41*	1.90	3.00*	2.40
Fairness Score	9.73	9.63	9.34*	9.48	9.51	8.36*	9.69	8.72*	8.85*	6.59	7.39	7.27
Accuracy Difference	9.81	9.75	9.38*	9.77	9.83	9.66	9.78	9.52	9.29	6.43	7.50	7.14
Precision Difference	9.76	9.61	9.26	9.29	9.36	8.90	9.70	9.05	9.06	6.23	6.94	6.95
Recall Difference	9.65	9.72	9.56	9.75	9.68	8.02*	9.66	8.88	9.05*	7.38	8.33	7.86
FPR Difference	9.73	9.56	9.15	9.28	9.34	7.94*	9.67	8.21*	8.47	5.48	6.67	6.43
DemP Difference	9.79	9.69	9.64	9.51	9.51	7.98*	9.79	8.58*	8.97	8.81	8.69	9.05
EOd Difference	9.60	9.45	9.03*	9.28	9.34	7.64*	9.56	8.10*	8.26*	5.24	6.19	6.19
Privacy Score	5.85	6.03*	6.11*	5.26	5.29	7.39*	7.20	8.45	8.98*	6.73	6.40*	6.57
Roc-auc MInf 1	9.84	9.97	9.94	9.89	10.00*	10.00*	10.00	9.97*	9.98	10.00	10.00	10.00
Roc-auc MInf 2	9.69	9.98*	9.98*	9.74	9.69	9.94*	9.88	9.95	9.99	8.62	8.30*	8.10*
Accuracy ModExt 1	1.94	2.10*	2.23*	0.80	0.82	4.81*	4.76	7.01	8.01*	4.71	4.43*	4.89
Accuracy ModExt 2	1.94	2.09	2.28*	0.60	0.66*	4.80*	4.15	6.86	7.93*	3.57	2.86	3.29
Robustness Score	6.66	6.78	6.95	6.47	6.33	7.04	6.91	8.29*	8.83*	4.41	5.73	6.10*
Accuracy (Shift)	8.16	8.77	8.12	7.61	8.43	6.74	8.06	8.35	8.66	0.75	3.93*	4.11*
Roc-auc (Shift)	9.34	9.40	8.99	9.25	9.57	6.20*	8.82	7.94	8.32	4.14	8.64*	7.55
Brier Loss	8.71	9.08	8.67	8.05	8.75*	8.17	8.46	9.20*	9.48*	3.34	7.82*	7.27*
Accuracy (FGSM)	2.98	1.86	4.05	3.45	0.64*	6.41	4.47	7.59*	8.55*	6.11	2.27*	5.52
Accuracy (PGD)	0.92	1.59*	1.91*	0.50	0.61*	4.85*	1.81	6.70*	8.02*	2.29	1.75*	2.21
Loss Sensitivity	9.83	9.97*	9.96*	9.94	9.98*	9.89	9.82	9.96*	9.96*	9.82	9.96*	9.94*
Accuracy Score	9.54	9.21*	9.04*	9.74	9.68*	7.49*	9.09	6.65*	5.99*	8.86	9.12*	8.89

Table 3. Extended Trade-offs Fairness

	UTKFace				CelebA				FairFace				LFWA+			
Name	BL	W	α_1	α_{100}	BL	W	α_1	α_{100}	BL	W	α_1	α_{100}	BL	W	α_1	α_{100}
Explainability Score	6.21	5.97*	5.74*	5.59*	5.45	5.36	5.39	5.38	5.99	6.10*	6.00	5.97	5.42	5.58	5.22*	5.57
Robustness	9.65	9.65	9.52	9.52	9.67	9.64	9.63	9.75*	8.93	8.96	8.94	9.31	9.85	9.85	9.78	9.98*
Faithfulness	2.35	1.59	0.69*	0.30*	1.16	0.87	0.87	0.54	2.10	2.32	2.01	1.57	1.42	1.96	0.38*	0.43*
Complexity	6.28	6.33	6.25	6.37	8.04	8.12	7.72*	7.44*	6.22	6.21	6.16	6.09	8.51	8.35	7.61*	6.69*
Randomisation	6.55	6.32*	6.51	6.15	2.92	2.80	3.33*	3.79*	6.72	6.90	6.89	6.92	1.90	2.15	3.10*	5.17*
Fairness Score	9.73	9.58	9.61	9.65	9.48	9.39	9.37	9.45	9.69	9.57	9.64	9.64	6.59	6.70	6.72	9.87*
Accuracy Difference	9.81	9.63	9.68*	9.67	9.77	9.74	9.69	9.87*	9.78	9.71	9.82	9.80	6.43	6.90	7.38	9.88*
Precision Difference	9.76	9.45	9.63	9.63	9.29	9.19	9.14	9.30	9.70	9.57	9.74	9.55	6.23	6.84	6.51	9.95*
Recall Difference	9.65	9.82	9.57	9.70	9.75	9.68	9.75	9.52*	9.66	9.52	9.48	9.79	7.38	7.14	8.10	9.76*
FPR Difference	9.73	9.40	9.62	9.61	9.28	9.16	9.12	9.29	9.67	9.53	9.68	9.50	5.48	6.19	5.71	10.00*
DemP Difference	9.79	9.77	9.73	9.83	9.51	9.42*	9.43	9.41	9.79	9.69	9.66	9.70	8.81	8.10	7.62	9.88
EOd Difference	9.60	9.40	9.41	9.49	9.28	9.16	9.12	9.28	9.56	9.39	9.48	9.49	5.24	5.00	5.00	9.76*
Privacy Score	5.85	5.94	5.95	5.96*	5.26	5.26	5.20*	5.27	7.20	7.13*	7.13*	7.10*	6.73	7.07*	6.72	9.01
Roc-auc MInf 1	9.84	9.84	9.76	9.85	9.89	9.90	9.70*	9.91	10.00	9.96	9.98	9.99	10.00	10.00	9.77	9.64*
Roc-auc MInf 2	9.69	9.76	9.80	9.58	9.74	9.67*	9.65	9.67	9.88	9.89	9.81	9.86	8.62	8.84	8.70	8.50
Accuracy ModExt 1	1.94	2.09	2.15	2.22*	0.80	0.82	0.82	0.83	4.76	4.66*	4.64*	4.58	4.71	5.86*	4.96	9.11*
Accuracy ModExt 2	1.94	2.05	2.12	2.20*	0.60	0.65*	0.63*	0.66*	4.15	4.02	4.09	3.95*	3.57	3.57	3.43	8.79
Robustness Score	6.66	6.63	6.44	4.81*	6.47	6.24	6.24	5.66*	6.91	6.91	6.75	6.67*	4.41	5.26	4.85	8.17*
Accuracy (Shift)	8.16	7.45*	6.88*	2.19*	7.61	6.47	6.58	4.54*	8.06	8.12	7.94	7.08*	0.75	3.50	2.25	8.00*
Roc-auc (Shift)	9.34	9.07	8.78*	5.71*	9.25	9.20	9.06	8.49*	8.82	8.89	8.68	8.23*	4.14	5.20	3.98	5.37
Brier Loss	8.71	8.22*	7.77*	4.27*	8.05	7.42	7.39	6.18*	8.46	8.58	8.14*	7.85*	3.34	4.20	4.08	7.46*
Accuracy (FGSM)	2.98	4.17	4.26	5.59*	3.45	3.89	4.03	4.31*	4.47	4.39	4.20	5.17	6.11	6.27	6.36*	9.41*
Accuracy (PGD)	0.92	1.00	1.12	1.35	0.50	0.52	0.50	0.56	1.81	1.86	1.94*	2.08*	2.29	2.57	2.71	8.82*
Loss Sensitivity	9.83	9.87	9.81	9.75	9.94	9.91*	9.88	9.90	9.82	9.64*	9.60*	9.64*	9.82	9.85	9.74	9.96
Accuracy Score	9.54	9.50	9.44	9.32	9.74	9.74	9.76	9.68*	9.09	9.07	9.03*	8.96*	8.86	8.71	8.64	5.59*

Table 4. Extended Trade-offs Privacy (*Note:* Only converging models)

	UTKFace			CelebA				FairFace			
Metric	Baseline	$\epsilon = 10$	$\epsilon = 100$	Baseline	$\epsilon = 1$	$\epsilon = 10$	$\epsilon = 100$	Baseline	$\epsilon = 1$	$\epsilon = 10$	$\epsilon = 100$
Explainability Score	6.21	3.56*	3.48*	5.45	3.36*	3.45*	3.35*	5.99	3.40*	3.56*	3.46*
Robustness	9.65	0.00*	0.00*	9.67	0.00*	0.00*	0.00*	8.93	0.00*	0.00*	0.00*
Faithfulness	2.35	0.65*	0.36*	1.16	0.00*	0.14*	0.30*	2.10	0.01*	0.45*	0.83*
Complexity	6.28	5.52*	5.63*	8.04	5.98*	5.91*	6.28*	6.22	5.69*	5.45*	5.69*
Randomisation	6.55	8.08*	7.94*	2.92	7.46*	7.77*	6.84*	6.72	7.89*	8.35*	7.32*
Fairness Score	9.73	9.27	9.40*	9.48	7.55*	8.34*	8.55*	9.69	9.29	8.86	9.20*
Accuracy Difference	9.81	9.64	9.61	9.77	9.37	9.42*	9.75	9.78	9.80	9.88	9.76
Precision Difference	9.76	9.71	9.52	9.29	7.63*	8.18*	8.36*	9.70	9.72	9.68	9.47
Recall Difference	9.65	8.89	9.22	9.75	7.77*	8.83*	8.60*	9.66	9.13	8.30	8.95*
FPR Difference	9.73	9.39	9.39	9.28	6.70*	7.68*	8.11*	9.67	9.07	8.54	9.14
DemP Difference	9.79	9.17	9.54	9.51	7.23*	8.26*	8.36*	9.79	9.10	8.42	9.05*
EOd Difference	9.60	8.81	9.15*	9.28	6.60*	7.68*	8.11*	9.56	8.90	8.30	8.81*
Privacy Score	5.85	7.29*	6.95*	5.26	6.55*	5.80*	5.63*	7.20	9.16*	8.82*	8.05*
Roc-auc MInf 1	9.84	9.96	9.96*	9.89	10.00*	9.99*	10.00*	10.00	9.96*	9.93*	9.93*
Roc-auc MInf 2	9.69	9.93*	10.00*	9.74	9.94*	9.83	9.75	9.88	9.98	9.98	9.96
Accuracy ModExt 1	1.94	4.65*	3.88*	0.80	3.08*	1.67*	1.38*	4.76	8.35*	7.77*	6.23*
Accuracy ModExt 2	1.94	4.62*	3.94*	0.60	3.17*	1.69*	1.40*	4.15	8.33*	7.60*	6.08*
Robustness Score	6.66	7.28	6.42	6.47	3.23*	2.19*	4.21*	6.91	8.89*	8.71*	7.39
Accuracy (Shift)	8.16	7.82	6.48*	7.61	0.55*	0.21*	3.31*	8.06	9.36*	9.14	7.62
Roc-auc (Shift)	9.34	7.32*	6.50*	9.25	2.48*	0.79*	5.65*	8.82	8.90	8.53	6.90*
Brier Loss	8.71	8.22	6.93*	8.05	0.97*	0.61*	4.17*	8.46	9.63*	9.32	7.65*
Accuracy (FGSM)	2.98	7.34*	6.37*	3.45	4.91	2.30	1.90	4.47	9.43*	8.77*	7.85*
Accuracy (PGD)	0.92	4.60*	3.94*	0.50	3.41*	2.27*	1.67*	1.81	8.33*	7.59*	5.99*
Loss Sensitivity	9.83	8.40*	8.28*	9.94	7.04*	6.94*	8.56*	9.82	7.71*	8.91	8.33*
Accuracy Score	9.54	7.70*	8.03*	9.74	8.23*	8.89*	9.12*	9.09	5.83*	6.21*	7.00*

Table 5. Extended Trade-offs Robustness

	UTKFace			CelebA			FairFace			LFWA+		
Name	BL	Augmix	PG	BL	Augmix	PG	BL	Augmix	PG	BL	Augmix	PG
Explainability Score	6.21	5.78*	7.68*	5.45	5.78*	7.09*	5.99	6.26*	7.76*	5.42	5.59	7.06*
Robustness	9.65	9.69	9.85*	9.67	9.67	9.88*	8.93	9.43*	9.60*	9.85	9.76*	9.90*
Faithfulness	2.35	0.24*	7.35*	1.16	1.94*	7.00*	2.10	2.26	7.82*	1.42	1.69	7.29*
Complexity	6.28	6.01*	6.06*	8.04	7.61*	7.89*	6.22	6.18	6.08*	8.51	7.99*	8.02*
Randomisation	6.55	7.17*	7.45*	2.92	3.89*	3.60*	6.72	7.17	7.55*	1.90	2.93*	3.03*
Fairness Score	9.73	9.76	9.70	9.48	9.33	9.41*	9.69	9.82	9.68	6.59	8.40*	7.01
Accuracy Difference	9.81	9.75	9.71	9.77	9.79	9.70	9.78	9.90	9.72	6.43	8.45*	6.55
Precision Difference	9.76	9.84	9.61	9.29	9.13	9.20*	9.70	9.82	9.84	6.23	8.71*	7.05
Recall Difference	9.65	9.62	9.83	9.75	9.52	9.77	9.66	9.80	9.47	7.38	7.62	5.95
FPR Difference	9.73	9.85	9.59	9.28	9.11	9.16*	9.67	9.81	9.82	5.48	8.81*	7.14
DemP Difference	9.79	9.87	9.88	9.51	9.31	9.47	9.79	9.86	9.75	8.81	9.17	9.40
EOd Difference	9.60	9.62	9.59	9.28	9.11	9.16*	9.56	9.76*	9.47	5.24	7.62*	5.95
Privacy Score	5.85	5.88	5.82	5.26	5.27	5.23	7.20	6.98*	7.06*	6.73	6.58	6.63
Roc-auc MInf 1	9.84	9.70	9.58*	9.89	9.99	9.86	10.00	10.00	9.97	10.00	10.00	10.00
Roc-auc MInf 2	9.69	9.88*	9.65	9.74	9.68	9.65	9.88	9.88	9.81	8.62	8.26	8.62
Accuracy ModExt 1	1.94	1.97	2.03	0.80	0.79	0.80	4.76	4.33*	4.59	4.71	4.79	4.64
Accuracy ModExt 2	1.94	1.98	2.02	0.60	0.62	0.61	4.15	3.69*	3.85*	3.57	3.29	3.25
Robustness Score	6.66	7.02	6.17	6.47	6.80	6.49	6.91	7.03	6.80	4.41	6.60*	6.26
Accuracy (Shift)	8.16	8.79*	7.32*	7.61	9.19*	8.67*	8.06	8.79*	7.79	0.75	6.61*	6.25*
Roc-auc (Shift)	9.34	9.69*	9.05	9.25	9.76*	9.67*	8.82	9.42*	8.90	4.14	8.03*	7.84
Brier Loss	8.71	9.26*	8.14*	8.05	9.31*	8.94*	8.46	9.13*	8.46	3.34	6.88*	7.58*
Accuracy (FGSM)	2.98	3.68	1.71	3.45	2.13	1.20*	4.47	3.58	4.21	6.11	5.62	3.21*
Accuracy (PGD)	0.92	0.80	0.92	0.50	0.44	0.50	1.81	1.55*	1.86	2.29	2.61	2.82*
Loss Sensitivity	9.83	9.89	9.88	9.94	9.96	9.96	9.82	9.72*	9.60*	9.82	9.85	9.88*
Accuracy Score	9.54	9.60	9.54	9.74	9.76*	9.76*	9.09	9.22*	9.07	8.86	8.70	8.59*