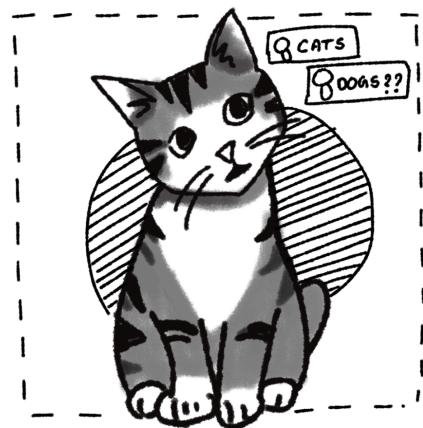
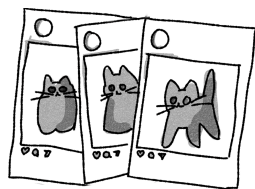


Test-Time Collective Action



proxy-based perturbations for correcting algorithmic harms

Imagine a photo app with an image classifier that labels cats as **dog** far more often than the inverse.



Cats are systematically misclassified.

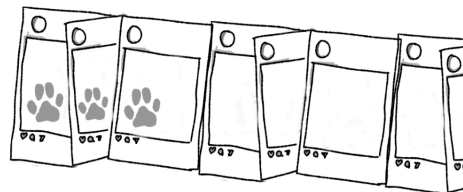
This represents a **quality-of-service harm** (Shelby et al., 2023) where the system performs well for most users but disproportionately under-performs for particular groups.

Who fixes these issues currently? Usually, the platform, by changing the data, the training, or the outputs (Caton & Haas, 2024). What happens when platforms doesn't want to, or cannot fix these issues?

2

What is Algorithmic Collective Action (ACA)?

Hardt et al. (2023) introduce algorithmic collective action (ACA) where the collective executes an algorithmic strategy, instructing each participant how to modify their own data to achieve a collective goal.



α · the collective

everyone else

goal

Plant a signal: The users want to teach the model that images with paw marks belong to the "cat" class.

3

How does ACA work?

1. Users agree on a shared signal (In this case, a tiny paw mark).

2. Every cat photo they share carries it.

3. The platform retrains on that data. (if it does)

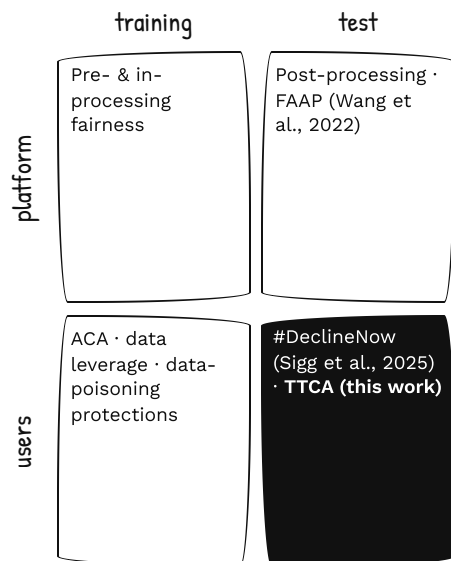
4. The model may learn: paw mark → cat.



the catch is step 3. Users can't make the platform retrain. Scraped data can't be revised, and trainers can wait for a model that resists it (Radiya-Dixit et al., 2022).

4

Where is this work situated in terms of current literature?



5

TTCA - Test time collective action

Correcting a frozen model from outside the training loop

1. pool queries

Each member spends up to B queries asking the platform to label shared public images.

2. build a proxy

Train 5 proxy models on the platform's answers (KnockoffNets). They inherit its cat → dog mistakes.

3. create one perturbation per class

On the proxies, optimise a perturbation δ_{cat} that pushes every cat photo toward "cat".

4. apply perturbation to misclassified images

Members add δ_{cat} to their cat photos before uploading. No retraining, no more queries.

Works effectively when used for corrective action aimed at fixing misclassified images

6

Experimental Results

CIFAR-10, CIFAR-100, FairFace · We use bottom 10% performing classes as misclassified class.

1. subgroup accuracy rises with query budget and transfers across architectures

Even at the smallest budget the subgroup accuracy beats the unperturbed baseline. Also a smaller collective proxy can be used to generate perturbation for larger platform model.

2. It's cheaper to pool data and generate proxy than individually trying to correct the errors.

when users pool their data, they only spend the query budget once and can reuse the generated perturbations, which leads to around 100x-300x savings on total queries used by the collective.

7

Limitations

→ **Some tasks are the harm.** Gender recognition from faces can erase trans and non-binary people (Keyes 2018; Scheuerman et al. 2019). Making it more accurate doesn't fix that the harm itself in this case alternative option would be to use refusal or opt-out (Zong & Matias 2024).

→ The perturbation only fits the model it was built on. A platform update can break it and platforms could also deploy defences against it.

Full references are on the **back of the sheet after opening the zine.**

Note: There are a lot of simplifications in this zine and we would highly recommend reading our paper and the original works mentioned here if want to know more about these works. All illustrations were drawn by Meghana Bhanghe. Claude Design was used to generate template and formatting for the zine

want to read our paper? →

meghanabhanghe@hey.com

8



p2 ·

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; García, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *AIES*, 723–741. ACM.

Caton, S.; and Haas, C. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7).

p3-4 ·

Hardt, M.; Mazumdar, E.; Mendler-Dünner, C.; and Zrnic, T. 2023. Algorithmic Collective Action in Machine Learning. In *ICML*, Proceedings of Machine Learning Research, 12570–12586. PMLR.

Radiya-Dixit, E.; Hong, S.; Carlini, N.; and Tramèr, F. 2022. Data Poisoning Won't Save You From Facial Recognition. In *ICLR*. OpenReview.net.

p5 ·

Vincent, N.; Li, H.; Tilly, N.; Chancellor, S.; and Hecht, B. J. 2021. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. In *FAccT*, 215–227. ACM.

Wang, Z.; Dong, X.; Xue, H.; Zhang, Z.; Chiu, W.; Wei, T.; and Ren, K. 2022. Fairness-aware Adversarial Perturbation Towards Bias Mitigation for Deployed Deep Models. In *CVPR*, 10369–10378. IEEE.

Sigg, D.; Hardt, M.; and Mendler-Dünner, C. 2025. Decline Now: A Combinatorial Model for Algorithmic Collective Action. In *CHI*, 912:1–912:17. ACM.

p6 ·

Bhange, M.; Aïvodji, U.; and Creager, E. 2026. Test-Time Collective Action: Proxy-Based Perturbations for Correcting Algorithmic Harms. In *AIES*.

Orekondy, T.; Schiele, B.; and Fritz, M. 2019. Knockoff Nets: Stealing Functionality of Black-Box Models. In *CVPR*, 4954–4963. Computer Vision Foundation / IEEE.

p7 ·

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario. (CIFAR-10/100)

Kärkkäinen, K.; and Joo, J. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *WACV*, 1547–1557. IEEE.

p8 ·

Keyes, O. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum. Comput. Interact.*, 2(CSCW): 88:1–88:22.

Scheuerman, M. K.; Paul, J. M.; and Brubaker, J. R. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW): 144:1–144:33.

Zong, J.; and Matias, J. N. 2024. Data Refusal from Below: A Framework for Understanding, Evaluating, and Envisioning Refusal as Design. *ACM J. Responsib. Comput.*, 1(1).