Effort-aware Fairness: Incorporating a Philosophy-informed, Human-centered Notion of Effort into Algorithmic Fairness Metrics

Tin Trung Nguyen[1,*], Jiannan Xu[1,*], Zora Che[1], Phuong-Anh Nguyen-Le[1], Rushil Dandamudi[1], Donald Braman[2], Furong Huang[1], Hal Daumé III[1], Zubin Jelveh[1]

UNIVERSITY OF MARYLAND 1 THE GEORGE WASHINGTON UNIVERSITY WASHINGTON, DC 2

*: Equal contribution

TRAILS Trustworthy AI in Law & Society

Motivation

• Standard fairness metrics (e.g., demographic parity) ignore individuals' actual efforts—people may reach similar outcomes for different reasons.

• Existing approaches emphasize **counterfactual effort or recourse**, not **real effort over time** [1].

• **Force-based definitions of effort** from philosophy [2] remain largely unexplored in AI fairness research.

Contributions

• Propose a **philosophy-informed framework** for **Effort-aware Fairness (EaF)** based on the concept of **Force**, capturing the temporal trajectory and inertia of predictive features.

• Conduct a **pre-registered experiment** showing that people prioritize **feature trajectories** over aggregate values in fairness judgments.

• Develop **computational pipelines** to evaluate **EaF** in **criminal justice** and **personal finance** settings.

Paper

Scan here to check the paper!

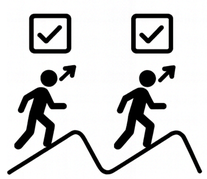
Metrics

• Inspired by Newton's Second Law [3] : $F = m \times a$.

• Define Effort as a Force-based formulation:
Effort $=$ inertia $\times$ acceleration.

• Inertia represents disadvantages beyond an individual's control (e.g., childhood poverty, disability).

• Acceleration captures the rate of change in cumulative input features (e.g., cumulative income growth since a reference year).



Effort-awre Individual Fairness (EaIF)

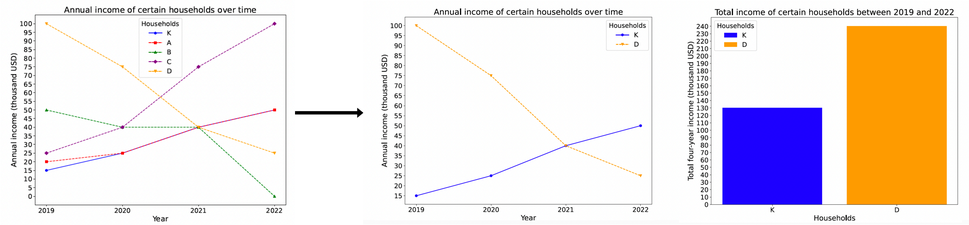
Effort-awre Group Fairness (EaGF)

Experiments

• **Pre-registered human study** designed to test three research questions (RQ1–RQ3) on how people perceive fairness based on effort trajectories in the criminal justice (CLUE) and personal finance (SHED) contexts.

• **Stage 1 (Input similarity):** Participants compared pairs of individuals to judge similarity based on **temporal trajectories** vs **static aggregate values**, addressing whether **trajectory distance** correlates more strongly with perceived input distance (**RQ1**) and whether this depends on the **Effort feature** type (income vs. arrests) (**RQ3**).

• **Stage 2 (Outcome similarity):** Participants evaluated whether individuals with **similar effort trajectories** should receive **similar outcomes**, testing whether **trajectory distance** correlates more strongly with perceived output distance (**RQ2**) and how this varies by **Effort feature** (**RQ3**).



Results

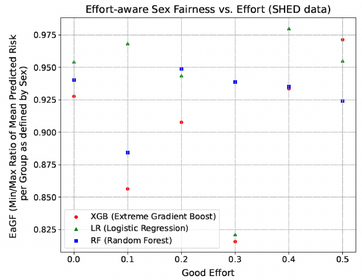
Human Experiment

Participants' fairness judgments were strongly influenced by **temporal trajectories**—individuals with better improvement paths were seen as more deserving, even when current feature values were identical.

Metric Operationalization

Model	EaIF Equal weights	EaIF Human study weights
XGBoost	0.80	0.79
Logistic regression	0.88	0.87
Random forest	0.91	0.90

EaIF Metrics: Random Forest was most fair, followed by Logistic Regression and XGBoost.



EaGF Metrics: Parity declined in higher-effort bins—revealing disparities even among high-effort individuals.

Key Takeaways:

• Accounting for **effort trajectories** and inertia reshapes fairness assessments.

• Effort-aware metrics expose **hidden biases** affecting those who exert high effort yet face structural disadvantages.

[1] Wen Huang, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fairness through equality of effort. In Companion Proceedings of the Web Conference 2020, pages 743–751 [2] Massin, O. (2017). Towards a definition of efforts. Motivation Science, 3(3), 230. [3] Young, H. D., Freedman, R. A., Sandin, T. R., & Ford, A. L. (1996). University physics (Vol. 9). Reading, MA: Addison-Wesley.