9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Closing the Loop: Feedback Loops and Biases in Automated Decision-Making

feedback loops, bias, machine learning, dynamical systems theory, sequential decision-making

Extended Abstract

Motivation Automated decision-making processes that use machine learning algorithms have become widespread, but researchers have found that these systems often perpetuate or even introduce biases. Efforts have been made to mitigate these biases using fairness criteria [1–3]. However, these solutions are designed for stationary systems [4, 5]. They are often not effective in the long term [6, 7] due to the feedback loops created by the decision-making process [4, 8–17]. To design effective long-term bias mitigation techniques, an interdisciplinary approach is needed to understand the role of feedback loops in perpetuating and amplifying biases.

Contributions We rigorously analyze the ML-based decision-making pipeline and establish a classification of distinct types of feedback loops. We represent the typical ML-based decision-making pipeline as a block diagram (as is usual in dynamical systems theory), which is composed of different sub-systems: the individuals' sampling process s, the individual i's unobservable characteristics representing the decision-relevant construct θ, the observed features x and outcomes y, the ML model f (producing a prediction $\hat{y}$ for i), and the final decision d. The final decision can feed back into any of the other sub-systems, thus forming different types of feedback loops (see Fig. 1): A **sampling feedback loop** comprises the effects of the decision on the probability certain types of individuals enter the decision-making pipeline (e.g., apply for a loan). An **individual feedback loop** is present if the decision acts directly on the individual's characteristics. In contrast to the individual feedback loop, in a **feature feedback loop** the decision affects the *observable* characteristics of the individual (e.g., the credit score) rather than the actual ones (likelihood of repaying a loan). In an **ML model feedback loop**, the decision affects the ML model by modifying the training data set that will be used for future predictions (the outcome is realized and added to the training data set only for positive decisions). Finally, in an **outcome feedback loop**, the decision affects the outcome before it is realized and ultimately observed (e.g., a loan given at a higher interest rate increases the probability of defaulting). Notice that some feedback loops can be classified as **adversarial** whenever the decision feeds back into the system involving some strategic action of the affected individual(s).

Furthermore, we associate the different types of feedback loops with the biases they affect (see Table 1). Sampling and ML model feedback loops can change the representation of the training or evaluation sample dataset compared to the target population, thus leading to representation bias. An individual feedback loop can cause life bias by changing an individual's decision-relevant (though, often unobservable) attributes. In contrast, feature and outcome feedback loops act on the extraction and realization of those attributes, which can affect the measurement bias of the observable attributes. In general, we find that the existence of feedback loops in the ML-based decision-making pipeline can perpetuate, reinforce, or even reduce ML biases.

Case study We demonstrate the connection between feedback loops and biases with a unifying case study on recommender systems (RS). We consider the case of an online platform where an RS is used to provide content the users are interested in. We consider just one item that can be shown to users of two groups $a \in \{G1, G2\}$. In Figure 2, we provide one example for each of the five types of feedback loops to illustrate how they are associated with different biases. As can be seen in Fig. 2a, a **sampling feedback loop** phenomenon leads to the reduction of G2 users on the platform. Therefore, G2 is underrepresented on the platform in the long term, with just 8.9% of the platform users. Also, the interest of users remaining on the platform is not biased (see Fig. 2b). Fig. 2c shows that an **individual feedback loop** results in a polarization of interests on the platform. Namely, users with high initial interest are more likely to be recommended the item and, as a result of this, their interest further increases over time, and vice versa for users with low initial interest. As shown in Fig. 2d, a **feature feedback loop** reduces the measurement error (which can be seen by the reduced variance) and the measurement bias of G2's clicking history over time. Fig. 2e shows the results of a **ML model feedback loop**: the ML model improves quickly for G1 but not for G2 (as visualized with the reduction of prediction errors over time in Fig. 2f). From the perspective of platform users, an **outcome feedback loop** can result in a situation in which one keeps receiving recommendations due to having clicked on similar content in the past, despite not being interested in it – see outcome realizations and the shift of the prediction model in Fig. 2g. This increases the prediction error (w.r.t. the true interest) over time (see Fig. 2h).

Potential impact By rigorously analyzing the ML pipeline, we believe that our framework is a necessary first step toward understanding the exact role of the feedback loops in it. Providing a rigorous classification of feedback loops will enable a deeper understanding of the existing works in the ML literature and it will allow putting their results into the perspective of their assumptions (e.g., which types of feedback loops are considered and which are not). We believe that our framework will be helpful for the purposeful design of feedback loops [9, 14], and for the development of long-term bias and unfairness mitigation techniques [1–3].

References

[1] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, NIPS'16, pages 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[2] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098095. URL https://doi.org/10.1145/3097983.3098095.

[3] Joachim Baumann, Anikó Hannák, and Christoph Heitz. Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2315–2326, New York, NY, USA, 2022. Association for Computing Machinery. doi: https://doi.org/10.1145/3531146.3534645. URL https://doi.org/10.1145/3531146.3534645.

[4] Alexandra Chouldechova and Aaron Roth. The Frontiers of Fairness in Machine Learning. pages 1–13, 2018. URL http://arxiv.org/abs/1810.08810.

[5] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, 3 2021. ISSN 2326-8298. doi: 10.1146/annurev-statistics-042720-125902. URL https://www.annualreviews.org/doi/10.1146/annurev-statistics-042720-125902.

[6] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed Impact of Fair Machine Learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158. PMLR, 2018. URL https://proceedings.mlr.press/v80/liu18c.html.

[7] Yi Sun, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. The Backfire Effects of Fairness Constraints. *ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments*, 2022. URL https://responsibledecisionmaking.github.io/assets/pdf/papers/44.pdf.

[8] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, Suresh Venkatasubramanian, and Christo Wilson. Runaway Feedback Loops in Predictive Policing. In *Proceedings of Machine Learning Research*, volume 81, pages 1–12, 2018. URL https://github.com/algofairness/.

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL http://www.fairmlbook.org.

[10] Yaowei Hu and Lu Zhang. Achieving Long-Term Fairness in Sequential Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9549–9557, 2022. doi: 10.1609/aaai.v36i9.21188. URL https://ojs.aaai.org/index.php/AAAI/article/view/21188.

[11] Alexandra Chouldechova and Aaron Roth. A Snapshot of the Frontiers of Fairness in Machine Learning. *Commun. ACM*, 63(5):82–89, 4 2020. ISSN 0001-0782. doi: 10.1145/3376898. URL https://doi.org/10.1145/3376898.

[12] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6), 7 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

[13] Xueru Zhang and Mingyan Liu. Fairness in Learning-Based Sequential Decision Algorithms: A Survey. *Studies in Systems, Decision and Control*, 325:525–555, 2021. ISSN 21984190. doi: 10.1007/978-3-030-60990-0_18.

[14] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020. doi: 10.1145/3351095.3372878.

[15] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dunner, and Moritz Hardt. Performative prediction. *37th International Conference on Machine Learning, ICML 2020*, PartF16814: 7555–7565, 2020.

[16] Lydia T. Liu, Adam Tauman Kalai, Ashia Wilson, Christian Borgs, Nika Haghtalab, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020. doi: 10.1145/3351095.3372861.

[17] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, 2:1389–1398, 2018. doi: 10.1145/3178876.3186044.

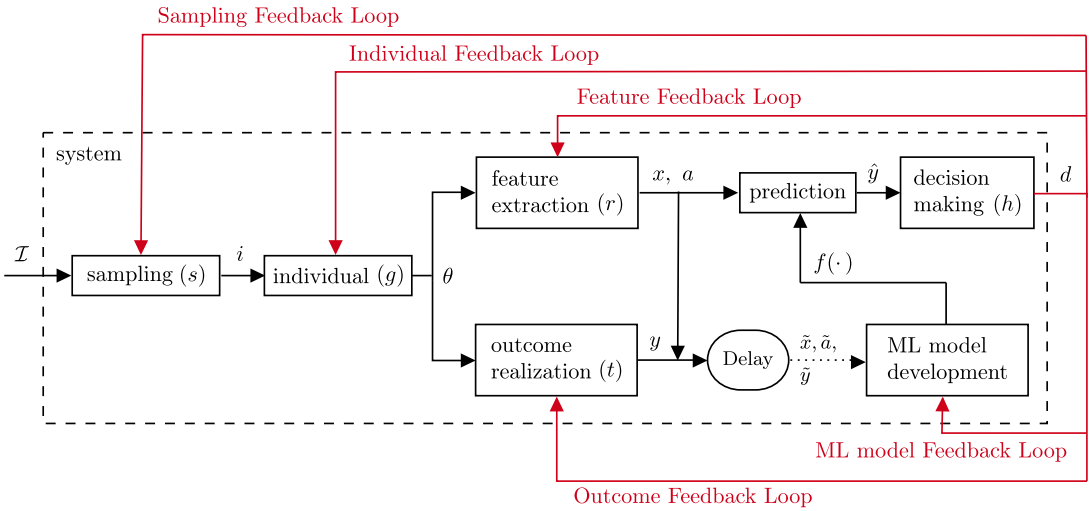
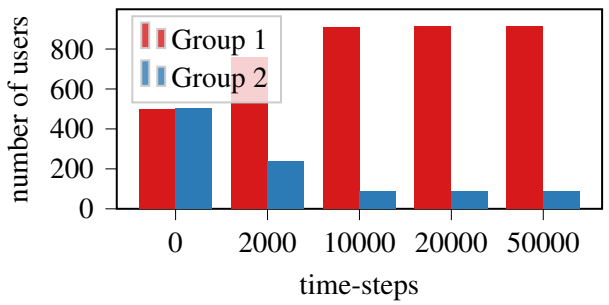


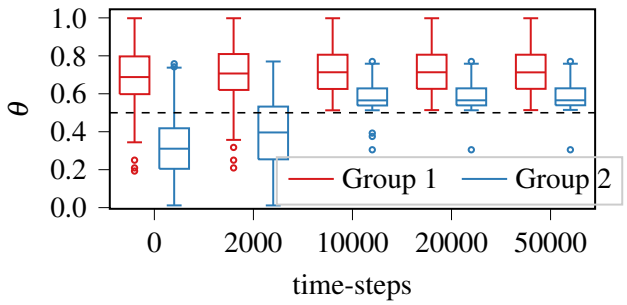
Figure 1: The ML-based decision-making pipeline as a closed-loop system in which different feedback loops can emerge. At the beginning of the pipeline, an individual i is sampled from the world (i.e., the environment) $\mathscr{I}$, through the function $s : \mathscr{I} \to i$. Let i be the individual's identity and let $g : i \to \theta$ be a function that returns the individual's attributes relevant for prediction. The features x, extracted through the function $r : \theta \to x, a$, and the outcome y (also called label or target), realized through the function $t : \theta \to y$, are imperfect proxies that can be measured. For instance, y can represent whether or not an individual repays a granted loan and x is a set of features that are used by the decision-maker to predict the repayment probability $\hat{y}$. For each sampled individual, the final decision d (e.g., whether to grant the loan or not) is informed by the prediction $\hat{y}$, which is produced based on the observed features x to approximate y using a learned function $f : x \to \hat{y}$. Once the outcome is observed, i.e., after one time-unit of delay, the past time's feature label pair $(\tilde{x}, \tilde{y})$ can end up as a sample in the dataset (X, Y) that is used to (re)train and (re)evaluate an ML model. In fully-automated decision-making systems, the decision rule h is solely based on the prediction ($h : \hat{y} \to d$), usually taking the form of a simple threshold rule, e.g., $d = 1$ if and only if $\hat{y} \geq \bar{y}$. The symbol a indicates the sensitive attribute of the individual (e.g., race or gender) and can possibly also be incorporated in the features x. Notice that d does not always directly follow from $\hat{y}$. For example, efforts to ensure group fairness usually consider the group membership a.

Table 1: Feedback loops and the ML biases they affect

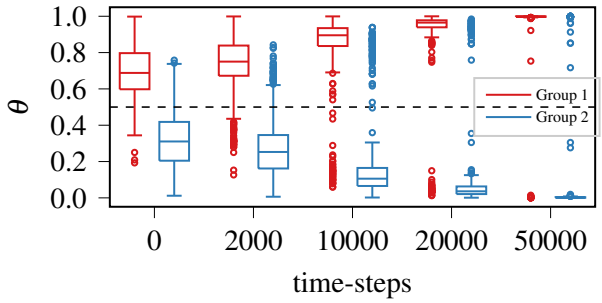
Feedback loop	**ML bias**
Sampling, ML model	Representation bias
Individual	Life bias
Feature, Outcome	Measurement bias



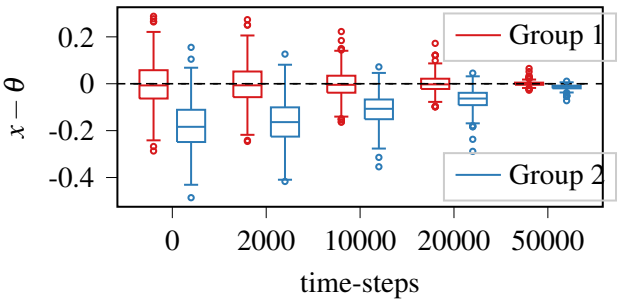
(a) Platform user cardinalities under a **sampling FL**: n_{G2} decreases from 504 to 89 individuals after 10,000 time-steps. This distribution persists in future time-steps, suggesting that it is a (locally) stable equilibrium point of the dynamical system.



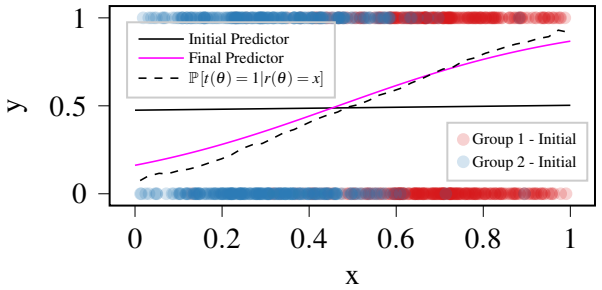
(b) Biased interests of platform users under a **Sampling FL**: since only those given $d = 1$ stay on the platform, the sample of active users becomes less representative over time, i.e., only interested users (those with high values for θ) stay on the platform.



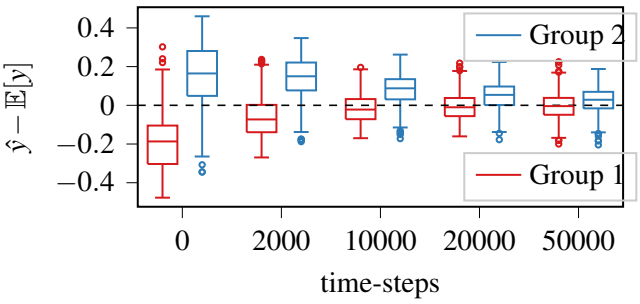
(c) Interests of platform users under an **individual FL**: Due to the difference in the two groups' initial distributions of θ, the polarization increases life bias. Namely, it results in even bigger group-level disparities with a very high θ for group 1 users and a very low θ for group 2 users, on average.



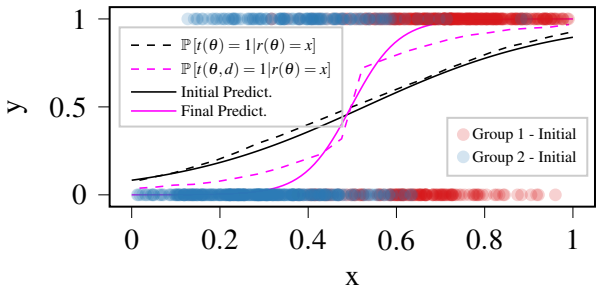
(d) Measurement error $(x-\theta)$ under a **feature FL**: this results in a reduction of the measurement error (which can be seen by the reduced variance) and the measurement bias of G2 over time, i.e., $x-\theta \sim 0$ after 50,000 time-steps.



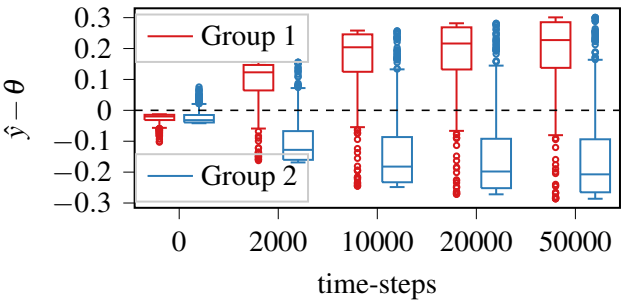
(e) Initial distribution of (X,Y), initial/final predictors, and outcome realization t under an **ML model FL**: Starting with an initial prediction of $\hat{y}=0.5$ for all, the model is retrained with newly observed data and becomes more accurate over time.



(f) Prediction error $(\hat{y}-\mathbb{E}[y])$ under an **ML model FL**: The prediction error quickly approaches 0 for G1, but the LR algorithm continues to perform poorly for G2 in the short to medium term.



(g) Initial distribution of (X,Y), initial/final predictors, and initial/final outcome realization t under an **outcome FL**: This introduces a measurement bias on the realized outcome y for both groups G1 and G2, meaning that the realized outcomes $t(\theta,d)$ are more extreme than they would be if there were no outcome feedback loop (see dashed lines).



(h) Prediction error with respect to the true, unobserved individual characteristics $(\hat{y}-\theta)$ under an **outcome FL**: The prediction error $\hat{y}-\theta$ diverges from 0 (as $\hat{y}$ predicts the realized outcome y and not θ) until it reaches a stable equilibrium point after approximately 10,000 time-steps (at approximately 0.2 and -0.2 for G1 and G2).

Figure 2: Dynamic effects of different feedback loops (FL) acting on an RS pipeline for an online platform. For simplicity, we consider one relevant item (e.g., a specific video) and denote a user's interest in this item with $\theta \in [0,1]$, where a larger θ corresponds to a higher interest. The realized outcome y denotes whether a user shows interest (e.g., clicks on the relevant item in question) or not. The platform's RS predicts a user's interest $\hat{y}=f(x)$, where the feature $x \in [0,1]$ represents a user's past clicking behavior. For this simple example, x is the percentage of recommended relevant items that the user has clicked on in the past and thus serves as a proxy of the user's interest in the relevant item.