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

Nicolò Pagan[1,*,†], Joachim Baumann[1,2,*,†], Ezzat Elokda[3], Giulia De Pasquale[3], Saverio Bolognani[3] and Anikó Hannák[1]

[1]University of Zurich, Zurich, Switzerland
[2]Zurich University of Applied Sciences, Zurich, Switzerland
[3]ETH Zurich, Zurich, Switzerland

Abstract

Prediction-based decision-making systems are increasingly used in various domains, but they are vulnerable to feedback loops that exacerbate existing biases over time. While researchers started proposing longer-term solutions to prevent adverse outcomes (such as bias towards certain groups), these interventions largely depend on ad hoc modeling assumptions and a rigorous theoretical understanding of the feedback dynamics in ML-based decision-making systems is currently missing. We use dynamical systems theory to analyze the ML-based decision-making pipeline, classify feedback loops, and show which specific types of ML biases are affected by each type of feedback loop. We encourage readers to consult the more complete manuscript [1].

Keywords

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

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 understand and mitigate these biases using fairness criteria [2]. However, these solutions are designed for stationary systems [3, 4]. Even though researchers recently started studying the long-term effects of sequential decision-making algorithms (e.g., [5–7], see [8] for a recent survey), the proposed simulation-based solutions are drawn on ad hoc models which prevent a comparison of their underlying assumptions and a deep interpretation of the driving factors, i.e., what causes the feedback loops and which components of the system are involved. As a result, to date, we lack a comprehensive classification and theoretical understanding of these feedback loops, and how they relate to the amplification of different types of bias.

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

[*] Corresponding authors.

[†] These authors contributed equally.

nicolo.pagan@uzh.ch (N. Pagan); baumann@ifi.uzh.ch (J. Baumann); elokdae@ethz.ch (E. Elokda); degiulia@ethz.ch (G. De Pasquale); bsaverio@ethz.ch (S. Bolognani); hannak@ifi.uzh.ch (A. Hannák)
0000-0003-0071-7856 (N. Pagan); 0000-0003-2019-4829 (J. Baumann); 0000-0002-2415-6494 (E. Elokda); 0000-0002-8587-967X (G. De Pasquale); 0000-0002-7935-1385 (S. Bolognani); 0000-0002-0612-6320 (A. Hannák)

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). To validate this terminology, we reviewed and classified 24 recent relevant papers (see Table 1) – where 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 historical 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.

Table 1
Feedback loops in the algorithmic fairness literature and their relation to biases.

Feedback loop	**Literature**		**ML biases**
	non-adversarial	*adversarial*	
Sampling Feedback Loop	[9–11]	–	Representation bias
Individual Feedback Loop	[12, 13]	[5, 7, 14–17]	Historical bias
Feature Feedback Loop	[5, 6, 11, 18–21]	[5, 14–16, 22–25]	Measurement bias
ML Model Feedback Loop	[21, 26–29]	–	Representation bias
Outcome Feedback Loop	[25]	–	Measurement bias

Potential impact By rigorously analyzing the ML pipeline, we believe that our framework is a necessary preliminary step towards (i) understanding the exact role of the feedback loops and (ii) shifting the research focus from short-sighted solutions that aim to identify and correct existing biases to a more forward-looking approach that seeks to anticipate and prevent biases

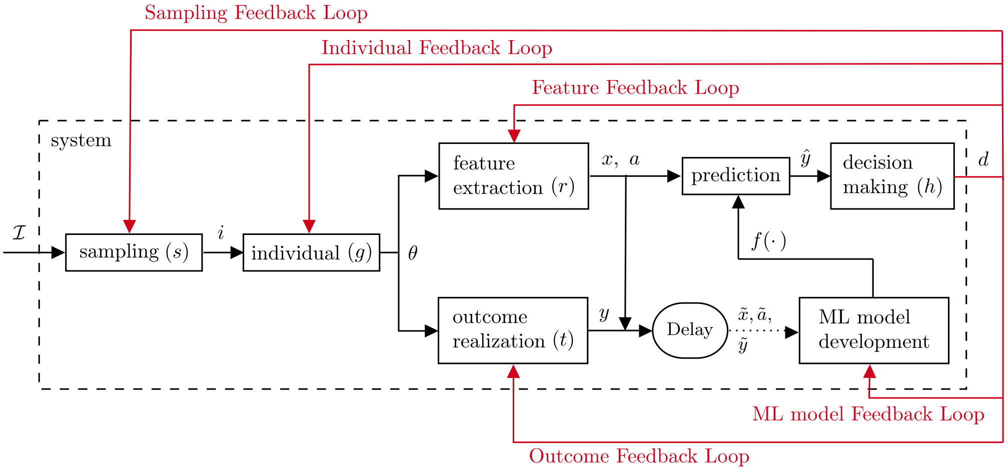

Figure 1: The ML-based decision-making pipeline as a closed-loop system in which different feedback loops can emerge. The pipeline is described in more detail in Appendix A along with the notation used.

in the long term. First, providing a rigorous classification of feedback loops will pave the way for a systematic review of 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). Second, with the help of additional tools, e.g., dynamical systems and control theory, it will be possible to fully exploit the potential of our framework in the purposeful design of feedback loops, and for the development of effective long-term unfairness mitigation techniques.

Acknowledgments

We want to thank Kenny Joseph, Florian Dörfler, Sarah Dean, and the members of the Social Computing Group at the University of Zurich (Corinna Hertweck, Stefania Ionescu, Aleksandra Urman, Leonore Röseler, Azza Bouleimen, and Desheng Hu) for their feedback on an earlier version of this manuscript. This work was supported by the University of Zurich, ETH Zurich, and NCCR Automation, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 180545).

References

[1] N. Pagan, J. Baumann, E. Elokda, G. De Pasquale, S. Bolognani, A. Hannák, A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems (2023). arXiv:2305.06055.

[2] J. Baumann, A. Castelnovo, R. Crupi, N. Inverardi, D. Regoli, Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias, in: 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3593013.3594058.

[3] A. Chouldechova, A. Roth, The Frontiers of Fairness in Machine Learning (2018) 1–13. URL: http://arxiv.org/abs/1810.08810.

[4] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 8 (2021) 141–163. doi:10.1146/annurev-statistics-042720-125902.

[5] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, Y. 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 (2020) 525–534. doi:10.1145/3351095.3372878.

[6] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, M. Hardt, Delayed Impact of Fair Machine Learning, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 3150–3158. URL: https://proceedings.mlr.press/v80/liu18c.html.

[7] L. Hu, Y. 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 (2018) 1389–1398. doi:10.1145/3178876.3186044.

[8] X. Zhang, M. Liu, Fairness in Learning-Based Sequential Decision Algorithms: A Survey, Studies in Systems, Decision and Control 325 (2021) 525–555. doi:10.1007/978-3-030-60990-0{_}18.

[9] T. Hashimoto, M. Srivastava, H. Namkoong, P. Liang, Fairness Without Demographics in Repeated Loss Minimization, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 1929–1938. URL: https://proceedings.mlr.press/v80/hashimoto18a.html.

[10] X. Zhang, M. M. Khalili, C. Tekin, M. Liu, Group retention when using machine learning in sequential decision making: The interplay between user dynamics and fairness, Advances in Neural Information Processing Systems 32 (2019).

[11] X. Zhang, M. M. Khalili, M. Liu, Long-Term Impacts of Fair Machine Learning, Ergonomics in Design 28 (2020) 7–11. doi:10.1177/1064804619884160.

[12] W. S. Rossi, J. W. Polderman, P. Frasca, The closed loop between opinion formation and personalised recommendations, IEEE Transactions on Control of Network Systems (2021) 1. doi:10.1109/TCNS.2021.3105616.

[13] N. Perra, L. E. C. Rocha, Modelling opinion dynamics in the age of algorithmic personalisation, Scientific reports 9 (2019) 1–11. doi:https://doi.org/10.1038/s41598-019-43830-2.

[14] H. Heidari, V. Nanda, K. P. Gummadi, On the Long-term Impact of Algorithmic Decision Policies: Effort unfairness and feature segregation through social learning, in: Proceedings of the 36th International Conference on Machine Learning, volume 97, 2019, pp. 2692–2701. doi:https://proceedings.mlr.press/v97/heidari19a.html.

[15] J. Kleinberg, M. Raghavan, How Do Classifiers Induce Agents to Invest Effort Strategically?, ACM Transactions on Economics and Computation 8 (2020). doi:10.1145/3417742.

[16] L. T. Liu, A. T. Kalai, A. Wilson, C. Borgs, N. Haghtalab, J. Chayes, The disparate equilibria of algorithmic decision making when individuals invest rationally, FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020) 381–391.

doi:10.1145/3351095.3372861.

[17] X. Zhang, R. Tu, Y. Liu, M. Liu, H. Kjellström, K. Zhang, C. Zhang, How do fair decisions fare in long-term qualification?, Advances in Neural Information Processing Systems 2020-Decem (2020) 1–13.

[18] A. J. B. Chaney, B. M. Stewart, B. E. Engelhardt, How algorithmic confounding in recommendation systems increases homogeneity and decreases utility (2018) 224–232. doi:10.1145/3240323.3240370.

[19] Y. Sun, A. Cuesta-Infante, K. 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.

[20] Y. Sun, Algorithmic Fairness in Sequential Decision Making, Ph.D. thesis, 2022.

[21] A. Sinha, D. F. Gleich, K. Ramani, Deconvolving Feedback Loops in Recommender Systems, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/962e56a8a0b0420d87272a682bfd1e53-Paper.pdf.

[22] L. Hu, N. Immorlica, J. W. Vaughan, The Disparate Effects of Strategic Manipulation, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 259–268. doi:10.1145/3287560.3287597.

[23] S. Tsirtsis, B. Tabibian, M. Khajehnejad, A. Singla, B. Schölkopf, M. Gomez-Rodriguez, Optimal Decision Making Under Strategic Behavior (2019). URL: http://arxiv.org/abs/1905.09239.

[24] S. Milli, J. Miller, A. D. Dragan, M. Hardt, The social cost of strategic classification, FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (2019) 230–239. doi:10.1145/3287560.3287576.

[25] J. C. Perdomo, T. Zrnic, C. Mendler-Dunner, M. Hardt, Performative prediction, 37th International Conference on Machine Learning, ICML 2020 PartF16814 (2020) 7555–7565.

[26] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, S. Venkatasubramanian, C. Wilson, Runaway Feedback Loops in Predictive Policing, in: Proceedings of Machine Learning Research, volume 81, 2018, pp. 1–12. URL: https://github.com/algofairness/.

[27] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, S. Venkatasubramanian, M. Mohri, K. Sridharan, Decision making with limited feedback: Error bounds for predictive policing and recidivism prediction, Proceedings of Machine Learning Research 83 (2018) 1–9.

[28] Y. Bechavod, K. Ligett, A. Roth, B. Waggoner, Z. S. Wu, Equal opportunity in online classification with partial feedback, Advances in Neural Information Processing Systems 32 (2019).

[29] H. Elzayn, M. Kearns, S. Jabbari, S. Neel, Z. Schutzman, C. Jung, A. Roth, Fair algorithms for learning in allocation problems, FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (2019) 170–179. doi:10.1145/3287560.3287571.

A. Notation for the ML-based decision-making pipeline in Fig. 1

At the beginning of the pipeline, an individual i is sampled from the world (i.e., the environment) $\mathscr{I}$, which represents a signal entering in the sampling function block $s : \mathscr{I} \to i$. Let i be the individual's identity – i.e., its index in the population – and let $g : i \to \theta$ be a function that returns the individual's attributes. More precisely, θ denotes the construct that is relevant for the 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 (for example, the credit score, as widely used in the US) that are used by the decision-maker to predict the repayment probability $\hat{y}$ in order to decide whether to grant the loan or not. For each sampled individual, the final decision d 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. More precisely, the training, evaluation, prediction, or decision-making can use the information on the individual group memberships. Notice that d does not always directly follow from $\hat{y}$. Efforts to ensure group fairness usually take the group membership a into account, e.g., to avoid disparate impact. Similarly, in non-automated decision-making systems, human decision-makers might consider any external, environmental information z, resulting in a more complex decision rule $h : f, x, a, \hat{y}, z \to d$.