Can Algorithms be Explained Without Compromising Efficiency? The Benefits of Detection and Imitation in Strategic Classification

Extended Abstract

Flavia Barsotti
ING Analytics & IAS, University of Amsterdam, The Netherlands
flavia.barsotti@ing.com,f.barsotti@uva.nl

Rüya Gökhan Koçer
ING Analytics
Amsterdam, The Netherlands
ruya.kocer@ing.com

Fernando P. Santos
Informatics Institute, University of Amsterdam, The Netherlands
f.p.santos@uva.nl

ABSTRACT

Given the ubiquity of AI-based decisions that affect individuals' lives, providing transparent explanations about algorithms is ethically sound and often legally mandatory. How do individuals strategically adapt following explanations? What are the consequences of adaptation for algorithmic accuracy? We simulate the interplay between explanations shared by an Institution (e.g. a bank) and the dynamics of strategic adaptation by Individuals reacting to such feedback. Resorting to an agent-based approach, our model scrutinizes the role of: i) transparency in explanations, ii) detection capacity and iii) behavior imitation. We find that the risks of transparent explanations are alleviated if effective methods to detect faking behaviors are in place. Furthermore, we observe that social learning and imitation — as often observed across societies — is likely to alleviate the impacts of (malicious) adaptation.

KEYWORDS

Strategic classification; Adaptive agents; Transparency; Imitation

ACM Reference Format:
Flavia Barsotti, Rüya Gökhan Koçer, and Fernando P. Santos. 2022. Can Algorithms be Explained Without Compromising Efficiency? The Benefits of Detection and Imitation in Strategic Classification: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

The expanding use of AI-applications in various decision-making processes has been accompanied by ethical concerns regarding both intended and unintended consequences of such applications [6, 15]. In the context of consequential algorithmic decisions, people seek information on the reasoning behind automated decision-making processes and the requirements to legitimately attain a desired outcome. The expectation of recourse is not only ethically sound but may also be legally mandatory [6, 10, 18]. In the case of a classification outcome, the explanations provided by the Institution should clarify the reasons behind a particular decision. This will help Individuals to understand their situation and adapt in future steps [17]. While explanations are legitimate and desirable, introducing recourse also poses some challenges: it is relevant to ensure that the recipe is not misused to, e.g., manipulate the decision making processes and generate wrong decisions (e.g. through disinformation, gaming or faking behaviours) [1, 3, 7, 11].

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

From the perspective of an institution, it is of utmost interest to understand how to provide explanations without compromising the original purpose of the algorithm. This is a non-trivial problem that requires considering the nature of the algorithm, societal norms, individuals adaptation processes, and the issue at stake.

The goal of this study is to introduce a formal framework to analyze this problem and illustrate opportunities and pitfalls that need to be considered when deploying algorithms. This becomes particularly relevant when algorithms are applied to strategic users who can learn from private information and from each other.

Related Work. The problem we explore in the present paper is related to the problem of adversarial [7] or strategic classification [11], where the goal is to define a learning algorithm that is robust against the strategic adaption of individuals. Our work is related with [12], where an explicit distinction between gaming and improving is established. The role of social learning and information sharing about the classifiers used by institutions, and the impacts of this process in strategic classification, was addressed in [9]. The role of transparency in strategic classification is discussed in [1]. This topic also relates with strategyproof regression and classification, where the goal is to design estimators that perform well, given that agents may misreport labeled examples to influence final classification decisions in their favor [8, 13, 16].

2 MODEL

Let us assume a population with two types of agents: **Institution** and **Individuals**. The goal of the Institution is to accurately classify individuals to provide them a service (e.g., granting or not a loan).

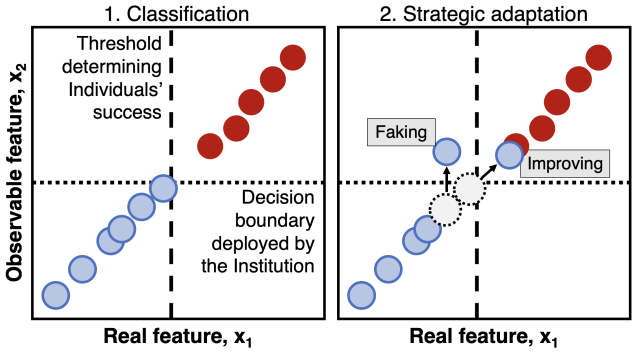

Figure 1: Overview of the model and feature space.

At time t, an Individual i is characterized by a (normalized) real feature value $x_1(i,t) \in [0,1]$ and a (normalized) observable feature

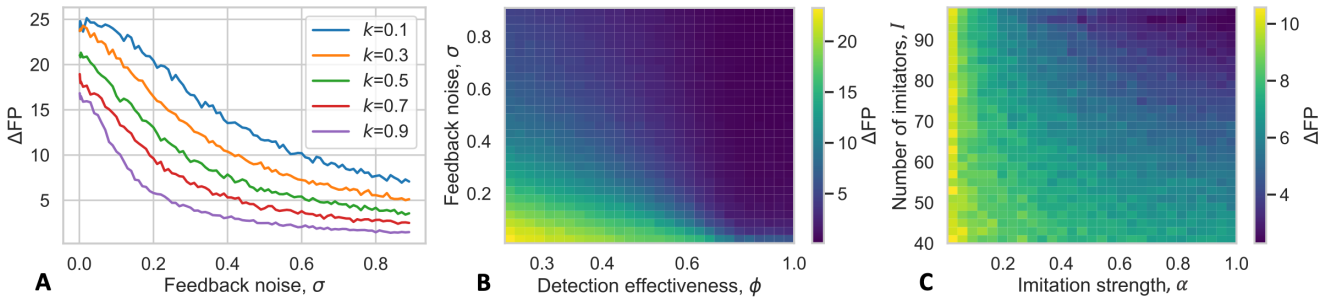

Figure 2: A) We observe that the number of False positives added after strategic adaptation (ΔFP) is higher if accurate information about decisions is provided (low σ). We set $c_f = k.c_i$ and $c_d = (1+k)c_i$, so that, by controlling parameter k, we can interpolate between scenarios where Individuals are likely ($k = 1.0$) or unlikely ($k = 0.0$) to improve. The risks associated with strategic adaptation are alleviated (i.e, ΔFP is reduced) if B) effective detection mechanisms are in place (high ϕ) and C) Individuals imitate the behavior of others (high α). Parameters (when not specified): $N = 100$, $b = 1.0$, $c_i = 3.0$, $k = 0.5$, $\alpha = 0$, $\sigma = 0.001$, $\phi = 0.7$.

value $x_2(i,t) \in [0,1]$. The Institution does not have perfect access to the the individuals' real state (x_1). A mismatch between the real and the observable feature can result from individuals providing erroneous information (*faking*) or Institutions implementing scoring methods that are not accurate. We assume that: i) the probability of success by individuals (e.g., to repay a loan) only depends on the real feature x_1; ii) the classifier set by the company is only a function of the observable feature x_2. This setting is summarized in Fig. 1. In the simplified scenario discussed in this work, we assume that an individual is successful if $x_1(i,t) > 0.5$.

The institution sets the classification threshold θ for the specific model assigning a score $S_i(x_2(i,t))$ to each Individual i. This defines the binary classification outcome $\Theta_i \in \{0,1\}$: $\Theta_i = 1$ if $S_i(x_2(i,t)) > \theta$ and $\Theta_i = 0$ otherwise. The institution provides an explanation and offers a recourse to Individuals for which $\Theta_i = 0$, i.e. those classified as "negative". Individual i infers the classification threshold θ_i via the estimate $\hat{\theta}_i$ defined as

$$\hat{\theta}_i = \max(S_i(x_2(i,t)), N \sim (\theta, \sigma)), \tag{1}$$

where $N \sim (\theta, \sigma)$ represents a value sampled from a Normal distribution with mean θ and standard deviation σ. Parameter σ controls the accuracy of the feedback provided.

Based on the feedback received from the bank, individuals adapt at time $t+1$ their features. Individuals decision upon the new vector of features at time $t+1$, namely $\vec{x}(i,t+1) = (x_1(i,t+1), x_2(i,t+1))$, by maximizing the (expected) utility function $u(i,t+1)$ defined as

$$u(i,t+1) = (1 - d[f(i,t+1)]) \cdot \hat{\Theta}_i(S_i(x_2(i,t+1)), \hat{\theta}_i) \cdot b$$
$$- \Delta_1(i,t+1) \cdot c_i - f(i,t+1)c_f - d[f(i,t+1)] \cdot c_d, \tag{2}$$

with $\Delta_1(i,t+1) := (x_1(i,t+1) - x_1(i,t))$ indicating the difference in the real feature. $\hat{\Theta}_i(S_i(x_2(i,t+1)), \hat{\theta}_i)$ indicates the estimate on the expected classification done by Individual i; Parameter $b \geq 0$ indicates the benefit of receiving a "positive" classification, e.g. $\Theta(\cdot) = 1$; We use $f(i,t)$, defined as $f(i,t) = x_2(i,t) - x_1(i,t)$, to denote the amount of fake information provided at time t by Individual i. Parameters c_i and c_f denote, respectively, the cost of improving and faking. The detection probability for Individual i at time t is given by

$$d[f(i,t)] = f(i,t)^{1/\phi}, \quad \phi \geq 0. \tag{3}$$

Since $f(i,t) \in [0,1]$, parameter ϕ is a measure of *detection effectiveness*. If $\phi = 1$, we assume a linear dependence of the detection probability on the amount of fake information; if $\phi = +\infty$ detection never fails and faking is always identified; if $\phi = 0$ detection always fails. We consider c_d as the cost for an Individual of being detected after faking.

Regarding imitation, we assume that a set I of individuals imitate (is influenced by) the behavior of others. We use $\vec{u}_m^*$ to denote the vector resulting from utility maximization and $\underline{\vec{u}}_P$ the vector resulting from the average behavior of a randomly observed pool P of individuals. Imitators I adapt their behavior by setting

$$\vec{x}(t+1) = \vec{x}(t) + (1-\alpha) \cdot \vec{u}_m^* + \alpha \cdot \underline{\vec{u}}_P, \quad \alpha \in [0,1], \tag{4}$$

where parameter α works as imitation strength in the adaptation process of imitators. Vector $\underline{\vec{u}}_P$ is the mean adaptation vector of $P \in [0, N-I]$ individuals randomly sampled from the pool of $N-I$ individuals that are first-movers and act without imitating others. Results deriving from this model are summarized in Fig. 2.

3 CONCLUSION

Here we formalize the interaction between multiple stakeholders (Individuals and Institution) and the Individuals social embedding (through imitation) in the context of a classification problem. Our work contributes to a recent trend in designing ethical multiagent systems taking into account their broader sociotechnical context [5, 14]. We find that the risks of transparent explanations are alleviated if effective methods to detect faking behaviors are in place and individuals imitate the behavior of others — as often observed across societies [2, 4]. This points out at least two different directions for further research to facilitate the ethical use of AI: understanding the normative factors shaping the imitation patterns in a society; developing techniques to improve the capacity to spot fraudulent behaviour associated with strategic classification.

ACKNOWLEDGMENTS

This research was supported by the Innovation Center for AI (ICAI).

REFERENCES

[1] Emrah Akyol, Cedric Langbort, and Tamer Basar. 2016. Price of transparency in strategic machine learning. *arXiv preprint arXiv:1610.08210* (2016).

[2] Abhijit V Banerjee. 1992. A simple model of herd behavior. *The Quarterly Journal of Economics* 107, 3 (1992), 797–817.

[3] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. 2006. Can machine learning be secure?. In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*. 16–25.

[4] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* 100, 5 (1992), 992–1026.

[5] Amit K Chopra and Munindar P Singh. 2018. Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 48–53.

[6] Danielle Keats Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Washington Law Review* 89 (2014), 1.

[7] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. 2004. Adversarial classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 99–108.

[8] Boi Faltings and Goran Radanovic. 2017. Game theory for data science: Eliciting truthful information. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11, 2 (2017), 1–151.

[9] Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. 2021. Strategic Classification in the Dark. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 3672–3681. https://proceedings.mlr.press/v139/ghalme21a.html

[10] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57.

[11] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. 111–122.

[12] Jon Kleinberg and Manish Raghavan. 2020. How Do Classifiers Induce Agents to Invest Effort Strategically? *ACM Transactions on Economics and Computation (TEAC)* 8, 4 (2020), 1–23.

[13] Reshef Meir, Ariel D Procaccia, and Jeffrey S Rosenschein. 2012. Algorithms for strategyproof classification. *Artificial Intelligence* 186 (2012), 123–156.

[14] Pradeep K Murukannaiah, Nirav Ajmeri, Catholijn M Jonker, and Munindar P Singh. 2020. New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'20)*. 1706–1710.

[15] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[16] Javier Perote and Juan Perote-Pena. 2004. Strategy-proof estimators for simple regression. *Mathematical Social Sciences* 47, 2 (2004), 153–176.

[17] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 284–293.

[18] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31 (2017), 841.