Transparency, Detection and Imitation in Strategic Classification

Flavia Barsotti[1,2,3], **Rüya Gökhan Koçer**[1] and **Fernando P. Santos**[4]
[1]ING Analytics, Amsterdam, The Netherlands
[2]Institute for Advanced Study, University of Amsterdam, The Netherlands
[3]Delft Institute of Applied Mathematics, TU Delft, Delft, The Netherlands
[4]Informatics Institute, University of Amsterdam, The Netherlands
{flavia.barsotti, ruya.kocer}@ing.com, 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. Our model identifies key aspects related to strategic adaptation and the challenges that an institution could face as it attempts to provide explanations. Resorting to an agent-based approach, our model scrutinizes: i) the impact of transparency in explanations, ii) the interaction between faking behavior and detection capacity and iii) the role of 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 behavioral imitation — as often happens across societies — can alleviate malicious adaptation and contribute to accuracy, even after transparent explanations.

1 Introduction

Building ethical algorithms is increasingly important in multiple domains, including the banking sector [Citron and Pasquale, 2014]. The expanding use of AI has been accompanied by ethical concerns emanating from both intended and unintended consequences of such use and applications, especially when AI-models impact humans in high-stake decisions. In this context, a fundamental aspect is the role of an ethically sound recourse, that in some cases may also be legally mandatory [Citron and Pasquale, 2014; Goodman and Flaxman, 2017; Wachter *et al.*, 2017].

While explanations are desirable, introducing recourse into decision-making guided by AI-applications poses challenges due to the ensuing adaptation process by humans. Consider the example of loan applications in banking: if costumers have access to the details of the algorithmic decision assigning a credit score, they might use that information to adapt and improve their condition in the future (e.g., on financial variables like savings, income, loan value) or try to game the algorithm (e.g., providing misleading or false information). It becomes fundamental to ensure that recourse and explanations are not misused and to prevent manipulation in the decision-making processes in order to avoid wrong decisions (e.g. through disinformation, gaming or faking behaviors) [Dalvi *et al.*, 2004; Barreno *et al.*, 2006; Hardt *et al.*, 2016; Akyol *et al.*, 2016; Bambauer and Zarsky, 2018; Hu *et al.*, 2019; Molnar, 2020]. Thus, from the perspective of an institution deploying algorithms, it is of utmost interest to overcome the tension between two ethical obligations: providing explanations to individuals without compromising the effectiveness and benefits of algorithms for society. This is a non-trivial problem which poses an ethical dilemma linking multiple dimensions: the algorithm, societal norms, individuals' adaptation, and the service at stake.

This paper proposes a mathematical framework to study the mentioned ethical dilemma by focusing on the interplay between the feedback shared by an Institution (e.g. a bank) and the strategic adaptation behaviors of Individuals subject to a generic classification model. We perform an analysis through an agent-based simulation approach. This enables us to assess the key issues related to strategic classification in the context of AI and their impact on the risks that the Institution could face resulting from the recourse provided. We use a binary classification model and study the effects of feedback, imitation and detection in countering faking behaviors via an illustrative example. We focus on: i) the role of noise in feedback/explanations provided by the Institution, ii) detection of gaming by Individuals and iii) adaptation through imitation of behaviors within a population. We find that the risks of transparent explanations regarding strategic manipulation (faking) are alleviated if effective methods to detect faking behaviors are in place and if individuals in a population are influenced by other individuals through imitation. Within the specific context of credit decision modeling in banking, the proposed framework represents a formal basis for the Institution to identify key factors to analyze and remedies to develop.[1]

[1]Supplementary Information available at: https://github.com/fp-santos/strategic-classification-imitation. An extended abstract of this paper appears in the Proc. of AAMAS'22 [Barsotti *et al.*, 2022].

1.1 Related Work

The problem we explore is closely related to the problem of adversarial [Dalvi *et al.*, 2004] and strategic classification [Hardt *et al.*, 2016], where the goal is to define a learning algorithm that is robust against the strategic adaption of individuals (so-called strategy-robust learning). The work by [Hardt *et al.*, 2016] framed the problem of strategic classification as a two-person sequential game between a Jury (Institution) and a Contestant (Individual). The authors assume that first the Jury trains a classifier, based on a sample of labeled individuals, and then Contestants adapt by changing their features at a cost. The authors prove that, if the Jury has some information about the Contestants' cost functions, and if such functions have certain properties (i.e., are separable) a Jury can define a classifier that is robust against strategic adaptation.

Our work is related with [Kleinberg and Raghavan, 2020] in explicitly distinguishing gaming from improvement. The authors consider the interaction between an Evaluator (Institution) and an Agent (Individual). The agents can explicitly decide to place effort in improving features that contribute to their success (self-improvement) or features that are arbitrary. The goal is to understand whether the evaluator can design rules that induce agents to improve rather than gaming. The authors show that improving rules exist when agents cannot easily transfer effort from improvement to gaming. Here we focus on a setting where individuals find cheaper to game than to improve, in the absence of extra mechanisms. Introducing the possibility that gaming can be detected allows us to interpolate between a scenario where successfully transferring effort from improving to gaming is always possible (no detection) to a scenario where it is impossible (effective detection).

Also disentangling gaming from improving, Miller *et al.* distinguish causal features (i.e., that can change the true label of individuals) and non-causal features [Miller *et al.*, 2020]. By considering real and observable features, we consider, respectively, a causal and non-causal feature. We do not perform any causal analysis; we focus, instead, on the distance between the causal and non-causal feature to explicitly study detection mechanisms that depend on such distance.

Recent works also consider imitation in strategic classification [Bechavod *et al.*, 2021; Ghalme *et al.*, 2021]. In [Bechavod *et al.*, 2021], the authors formulate the problem of strategic classification as a principal-agent game in a population composed of multiple groups. Notably, the model proposed assumes that the classifier used by institutions is not fully known to individuals, who try to guess it through information gathered from peers. The focus of [Bechavod *et al.*, 2021] is on how improvement occurs in multiple sub-groups, when individuals, depending on their group, can differ in their improvement costs and pool of peers to imitate. Differently from our setting, Bechavod *et al.* assume that individuals imitate information about the threshold and not directly actions. Agents are still assumed to be fully rational: they construct an estimate for the deployed rule, and then adapt through empirical risk minimization. Imitation is also considered in [Ghalme *et al.*, 2021], where agents use imitation to infer information about the classifier used by the Institution and best respond based on the information obtained. In our model, we assume that individuals directly imitate the behavior of others (behavioral imitation) rather than copying information available to infer the classification threshold. Here individuals do not necessarily best-respond to the (incomplete) knowledge they have about the classifier used by the institution: some agents (imitators) resort to adaptation through a combination of best-response and social learning. In both real and artificial systems, imitation and social learning are pointed as major drivers of behavior adaptation and norms formation [Sen and Airiau, 2007; Banerjee, 1992; Bikhchandani *et al.*, 1992]. Transparency in strategic classification is also studied in [Akyol *et al.*, 2016].

Also related with our work is a recent call to consider noisy adaptation by individuals in strategic classification [Jagadeesan *et al.*, 2021]. As highlighted above, we consider a model that deviates from the classic rational best-response model. We assume that individuals have imperfect information and adapt influenced by the behaviors of others in a population. We interpret noise to be a measure of transparency controlled by the Institution, yet the same formalism can be interpreted as a noisy response by individuals.

Finally, our paper is linked with literature on strategyproof regression and classification [Dekel *et al.*, 2010; Meir *et al.*, 2012; Krishnaswamy *et al.*, 2021] — aiming at designing estimators that performs well, when agents may misreport labeled examples — and *trust manipulation* — where the goal is to design scalable trust mechanisms that are robust to manipulation by strategic agents [Yu *et al.*, 2013].

Building on top of the previous literature, our paper contributes along three main directions:

1. We propose a **mathematical framework** to evaluate the impacts of strategic classification under arbitrary levels of transparency, while providing a basis to apply multi-agent simulations to study when faking behaviors are expected to be detrimental.
2. We study the role of **detection** mechanisms and the role of behavior **imitation** for different levels of feedback **transparency**.
3. We provide results suggesting that **detection and imitation alleviate the risks of strategic adaptation** to algorithmic decisions.

2 Model

Let us assume an environment where two types of agents exist: **Institution** and **Individual**. The goal of the Institution is to accurately classify individuals in order to provide them some service (e.g., a bank deciding to offer a loan based on credit score, or a college deciding to admit a student). At time t, a generic Individual i is characterized by a (normalized) real feature value $x_1(i,t) \in [0,1]$ (e.g., real income) and a (normalized) observable feature value $x_2(i,t) \in [0,1]$ (e.g., declared income, possibly cheating). We assume that the Institution does not have perfect access to the information regarding the individuals' real states. A mismatch between the real and the observable feature can result from individuals providing erroneous information (*faking*) or Institutions implementing not accurate scoring methods. We assume that:

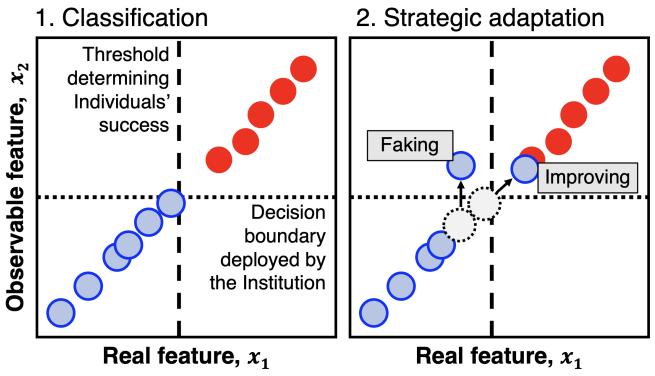

Figure 1: We consider a population of N Individuals being classified by one Institution. Circles represent Individuals – red/full disks are classified as positive and blue/circles are classified as negative. Individuals are located in a feature space, where the horizontal axis represents their real feature (x_1) and the vertical axis represents the feature observable by the Institution and used in classification (x_2). The institution aims at accurately classifying individuals in order to provide a service (e.g., concede a loan). Classification is based on the observable feature and actual success is based on the real feature. After an initial classification step (left panel), individuals can adapt their real feature (improve) or just adapt their observable feature (fake), as shown in the right panel.

i) the probability of success by individuals (e.g., repay a loan) only depends on the real feature x_1; ii) the classifier set by the Institution is only function of the observable feature x_2.

The model works as follows: The Institution runs the internal classification model and then shares some feedback with Individuals. The accuracy level of the feedback provided is tuned by a specific parameter in our model, allowing to assess scenarios ranging from full transparency (i.e., Individuals have full information about the exact position of the Institution decision boundary) to full obscurity (i.e., individuals can only guess where the decision boundary lies). With such knowledge, Individuals are allowed to act upon the way they were classified: Individuals classified as negative can change their features at a cost. Importantly, they can decide to pay a higher cost to improve (i.e., increasing the real feature x_1) or fake (i.e., only improving the observable feature x_2). This is reflected in the evaluation metrics of interest to the institution — widespread faking behaviors are likely to increase the number of False Positives (FP). In the case of a credit scoring model, FP represents Individuals classified as non-risky while actually being risky (e.g. not able to repay). This baseline setting is depicted in Fig. 1. Depending on the specific domain and modelling approach, the classification problem solved by the Institution can be arbitrarily complex. From an ethical perspective, we are interested in mitigating the risks that inaccurate feedback could create and the interplay between these effects and strategic adaptation. We assume that the expected probability of success by individuals (e.g., repaying a loan or having success in college) grows with x_1 and the classification by the Institution consists in setting a threshold θ defining the values of x_2 above which individuals are classified as positive (i.e., granted the loan or admitted to college).

Probability of success. Let us consider a generic Individual i having a (normalized) real feature $x_1(i,t) \in [0,1]$, and a (normalized) observable feature $x_2(i,t) \in [0,1]$ at time t. We denote with $\rho_i(x_1(i,t)) \in [0,1]$ the *probability of success* of Individual i at time t which depends on the real feature $x_1(i,t)$ as follows

$$\rho_i(x_1(i,t)) = \begin{cases} \frac{1}{1+e^{1/\epsilon(0.5-x_1(i,t))}}, & \text{if } \epsilon > 0 \\ H(x_1(i,t)-0.5), & \text{if } \epsilon = 0, \end{cases} \quad (1)$$

with $\epsilon \in \Re_0^+$ capturing the noise between the real feature and the probability of success. Function $\rho_i(x_1(i,t))$ was also called the *true* or *target* classifier [Hardt *et al.*, 2016]. Function $H(\cdot)$ is the Heaviside step function defined as

$$H(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases} \quad (2)$$

Based on the value of the noise parameter ϵ, observe that as ϵ becomes very large ($\epsilon \to \infty$), the probability of success of individuals gets closer to a value of 0.5 as limiting behavior, regardless of their real feature $x_1(i,t)$, for all Individuals i.

Conversely, for $\epsilon \to 0$ all individuals with $x_1 \geq 0.5$ will successfully repay the loan (or, in general, be successful). That is why in the special case of $\epsilon = 0$ we assume a Heaviside functional form.

The Institution sets the classification threshold θ for the model assigning a score $S_i(x_2(i,t))$ to each Individual i, which defines the binary classification outcome $\Theta_i \in \{0,1\}$. We assume that the Institution provides an explanation and offers a recourse to Individuals for which $\Theta_i = 0$, i.e. those receiving a score below the threshold:

$$\Theta_i(S_i(x_2(i,t)); \theta) = \begin{cases} 1, & \text{if } S_i(x_2(i,t)) \geq \theta \\ 0, & \text{if } S_i(x_2(i,t)) < \theta. \end{cases} \quad (3)$$

Given the partial information available to Individuals, we assume that a generic Individual i could try to infer the classification threshold set by the Institution

Inference on the classification threshold. A generic Individual i, with real feature $x_1(i,t)$ and observable feature $x_2(i,t)$ at time t, infers the classification threshold θ_i used by the Institution via the estimate $\hat{\theta}_i$ defined as

$$\hat{\theta}_i = \max(S_i(x_2(i,t)), N \sim (\theta, \sigma)), \quad (4)$$

where $N \sim (\theta, \sigma)$ represents a value sampled from a Normal distribution with mean θ and standard deviation σ. Parameter σ controls for the accuracy of the feedback provided by the Institution. For the sake of simplicity, here we assume that the scoring function value equals the value of the normalized observable feature $x_2(i,t)$, e.g. $S_i(x_2(i,t)) := x_2(i,t)$. The information regarding $x_2(i,t)$ is known to the Individual.

Observe that a high value of σ means that individuals cannot do better than randomly guessing the threshold used by the Institution. On the contrary, a low value of σ implies that individuals have exact information about the real threshold θ and consider strategic adaptation accordingly.

Faking information and probability of detection. After information is provided and individuals form their perception of the threshold used by the Institution, $\hat{\theta}_i$, individuals take their adaptation decisions. Let us use

$$f(i,t) = x_2(i,t) - x_1(i,t), \quad f(i,t) \in [0,1], \tag{5}$$

as the amount of fake information provided at time t by Individual i, measured as the difference between the observable feature and the real feature for the same individual. We denote with $d[f(i,t)]$ given by

$$d[f(i,t)] = f(i,t)^{1/\phi}, \quad \phi \geq 0, \tag{6}$$

the *probability of detection* for Individual i at time t. Since $f(i,t) \in [0,1]$, we introduce parameter ϕ as a measure of *detection effectiveness*. If $\phi = 1$, we assume a linear dependence of the detection probability on the amount of fake information; if $\phi \to +\infty$ detection never fails and faking is always identified; if $\phi \to 0$ detection always fails. The shape of this function is represented in Fig. 3. Note that we restrict our analysis to $x_1(i,t) \leq x_2(i,t)$ as $x_1(i,t) > x_2(i,t)$ would imply that individuals are hiding their true potential to be classified as positive by the algorithm and would be trivial for individuals to set, at no cost, their observable feature (x_2) to match the real feature (x_1).

Based on the explanation provided, individuals adapt. This decision depends on: i) utility maximization comprising the costs and benefits associated to each choice, and ii) behavior imitation. Next we introduce the cost functions and the adjustment mechanism based on utility maximization.

Cost functions. Let us denote with c_ι, c_f, respectively, the cost of improving and faking regarding the information provided; c_d denotes the cost of being detected. In order to study the impacts of the different costs, we define the following cost functions:

$$c_f = k \cdot c_\iota, \quad c_d = (1+k) \cdot c_\iota, \quad k \in [0,1], c_\iota \geq 0. \tag{7}$$

The cost of faking, c_f, results from, e.g., the effort of cheating on an exam, committing plagiarism or declaring non-existent income. In all cases, it is clear that there is some cost involved that is naturally much lower than the cost of honestly improving. The cost of being detected, c_d, is imputed when an individual fakes and is detected, e.g., resulting in suspension from college, a fine or an audit. In principle costs can vary independently (as studied in online Supplementary Information), however, by controlling k we can interpolate between extreme scenarios in terms of challenge degree to the Institution: i) Individuals are unlikely to improve (e.g. $k = 0$, assuming faking is cheap and detection cost is low) and ii) Individuals are likely to improve (e.g. $k = 1$, assuming faking is as expensive as improvement and detection costs are high).

Utility. Let us consider the set of Individuals i for which $\Theta_i(S_i(x_2(i,t)); \theta) = 0$, e.g. who received a negative classification. For each Individual i in this set, let us consider the estimate about the decision threshold $\hat{\theta}_i$ in Eq. (4) and the detection mechanism with probability of detection $d[f(i,t)]$ in Eq. (6). Based on the feedback received from the bank, these individuals define at time $t+1$ the new pair of features, potentially different from the original one. We assume that individuals decide the 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 $u(i,t+1)$, given as:

$$u(i,t+1) = (1 - d[f(i,t+1)])\hat{\Theta}_i(S_i(x_2(i,t+1)); \hat{\theta}_i)b$$
$$- \Delta_1(i,t+1)c_\iota - f(i,t+1)c_f - d[f(i,t+1)]c_d, \tag{8}$$

with $\Delta_1(i,t+1)$ indicating the variation in the information regarding the real feature

$$\Delta_1(i,t+1) := (x_1(i,t+1) - x_1(i,t)). \tag{9}$$

Notice that $\Theta_i(S_i(x_2(i,t)); \theta)$ provides the output of the bank's binary classification model; $\hat{\Theta}_i(S_i(x_2(i,t+1)); \hat{\theta}_i)$ indicates the estimate on the expected classification done by Individual i; $f(i,t+1)$ is given in Eq. (5) and $d[f(i,t+1)]$ refers to Eq. (6). Parameter $b \geq 0$ indicates the benefit of receiving a "positive" classification, i.e. $\Theta(\cdot) = 1$, while c_ι, c_f, denote, respectively, the cost of improving or faking and c_d the cost incurred after being detected.

Each term of the expected utility function can be explained as follows. We assume that individuals receive a benefit b if they are classified as positive and are not detected to be faking, $(1-d[f(i,t+1)]) \cdot \hat{\Theta}_i(S_i(x_2(i,t+1)); \hat{\theta}_i) \cdot b$; individuals pay an improvement cost c_ι for each unit of improvement in the information regarding their real feature, $\Delta_1(i,t+1) \cdot c_\iota$; individuals pay a faking cost c_f for each unit of change in their observable feature alone, $f(i,t+1) \cdot c_f$; and, finally, individuals pay a detection cost c_d in the case of faking and effective detection, $d[f(i,t)] \cdot c_d$. Note that, if faking does not occur, the last term of the utility function is null, i.e. individuals expect to never pay any detection cost since $d[f(i,t)] = 0$.

Remark. If 1) there is no detection mechanism ($d[f(i,t)] = 0$) and 2) faking can be implemented at no cost ($c_f = 0$), the expected utility in Eq. (8) simplifies to:

$$u(i,t+1) = \hat{\Theta}_i(S_i(x_2(i,t+1)); \hat{\theta}_i) \cdot b - \Delta_1(i,t+1) \cdot c_\iota.$$

In this scenario, utility increases with $x_2(i,t+1)$ and decreases with $(x_1(i,t+1) - x_1(i,t))$: individuals maximize the expected utility by increasing x_2 without modifying x_1 — that is, faking without improving. For the Institution (e.g., a bank), this means that Individuals that were initially accurately classified as positive are expected to repay the loan whereas those that decided to adapt are expected to fail in repaying the loan and constitute FP. In the context of strategic adaptation modelling, we introduce: 1) utility maximization and implement it through a standard optimization process in a 2-D (x_1, x_2) space, 2) the possibility of imitation.

Imitation. Let us assume that a set I of individuals in the population decide to imitate (is influenced by) the behavior of others individuals in the population. Let us denote with $\vec{u}^*_m(i)$ the vector resulting from utility maximization by individual i

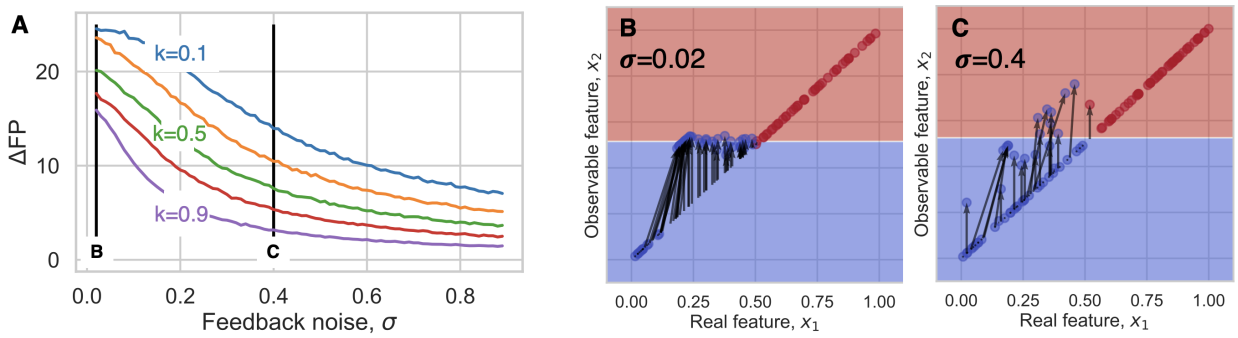

Figure 2: Panel **A** reports the difference in the number of False Positives after and before strategic adaptation (ΔFP). This occurs for different cost scenarios $k = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The cost functions are introduced in Eq. (7). Panels **B** and **C** elaborate on the impacts of strategic classification for two values of feedback noise: **B**) transparent feedback (low noise, $\sigma = 0.02$) and **C**) noisy feedback (high noise, $\sigma = 0.4$). We represent individuals' strategic adaptation along both their real feature ($x_1(i,t)$, horizontal axis) and the observable feature ($x_2(i,t)$, vertical axis). This provides extra insight on why transparency results in higher ΔFP. Parameters: $N = 100$, $b = 1.0$, $c_\iota = 3.0$, $\epsilon = 0$, $\phi = 1/3$, $\alpha = 0$, $k = 0.5$ (**B** and **C**). Results in **A** are an average over 10^3 runs starting from random initial conditions.

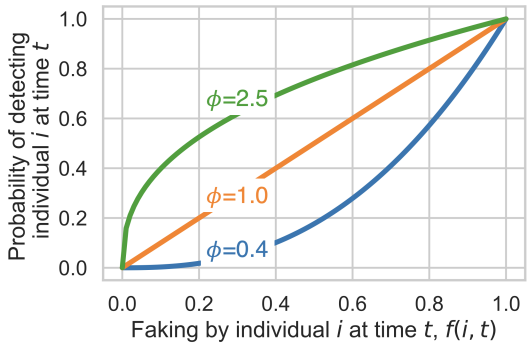

Figure 3: Detection mechanism. The probability that an individual is detected depends on the extent of the faking behavior (recall Eq. (5): $f(i,t) = x_1(i,t) - x_2(i,t)$) and detection effectiveness ϕ. This plot refers to Eq. (6) and shows that, in general, detection probability increases with $f(i,t)$ and ϕ.

and $\underline{\vec{u}}_P$ the vector resulting from the average behavior of a randomly observed pool P of individuals in the population. We assume that imitators in I adapt their behavior by setting

$$\vec{x}(i,t+1) = \vec{x}(i,t) + (1-\alpha)\cdot \vec{u}^*_m(i) + \alpha \cdot \underline{\vec{u}}_P, \quad \alpha \in [0,1], \tag{10}$$

where parameter α works as imitation strength in the adaptation process of imitators. Vector $\underline{\vec{u}}_P$ is taken as the mean adaptation vector of $P \in \{0, 1, ..., N-I\}$ individuals randomly sampled from the pool of $N - I$ individuals that are first-movers and act without imitating others.

Limiting the set of imitators and individuals to be imitated assumes that, in a population, not everyone has the same visibility and likelihood to be influenced by others. Section 3 discusses the results of the simulation study performed by means of the proposed analytical setting.

3 Results

Transparent feedback and faking behaviors. We explore how providing transparent feedback may impact the strategic adaptation of Individuals. The plot depicted in Fig. 2 shows the riskiness potentially associated to transparency/accuracy of the feedback shared by the Institution. Providing exact information about the classification threshold θ used by the Institution for the internal classification model increases the chance that individuals decide to react to the feedback by providing a new value of their observable feature $x_2(i,t+1)$ higher than the previous one. We assume that Individuals can modify the observable feature with or without modifying the real feature — faking results from the mismatch between the observable and real feature e.g. $x_2(i,t) - x_1(i,t)$. Here we assume that the Institution does not retrain the classifier $\Theta(\cdot)$ in $(t, t+1]$. As a consequence, more individuals can be erroneously classified as positive (i.e., the number of FP increases).

Panel **A** of Fig. 2 highlights that increasing σ, the level of noise in feedback/recourse, decreases the difference in FP before and after strategic adaptation (ΔFP). This occurs for different costs scenarios $k = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Recall that Eq. (7) introduces the cost functions as $c_f = k \cdot c_\iota$ and $c_d = (1+k) \cdot c_\iota$: increasing k makes it harder to fake and more costly to be detected.

Given a certain level of feedback noise, namely fixing $\sigma = \{0.02, 0.4\}$, Panels **B** and **C** further explore the drivers behind "why" transparency results in a higher number of FP. When feedback is transparent and individuals have the information regarding the decision boundary of the Institution, Individuals close to the boundary increase their observable feature just by the amount needed to be classified as positive; Individuals far away from the boundary do not attempt to fake and will be again classified as negative by the Institution (Panel **B**). Conversely, if feedback is noisy and individuals have uncertainty about the real decision threshold θ of the Institution, they attempt to maximize utility according to their random guess $\hat{\theta}_i$. Individuals overestimating the threshold can be dissuaded from spending hight costs to increase their observable feature; Individuals increasing their observable feature, yet undershooting the real decision threshold, will still be classified as negative. Overall, this results in a lower ΔFP (Panel **C**).

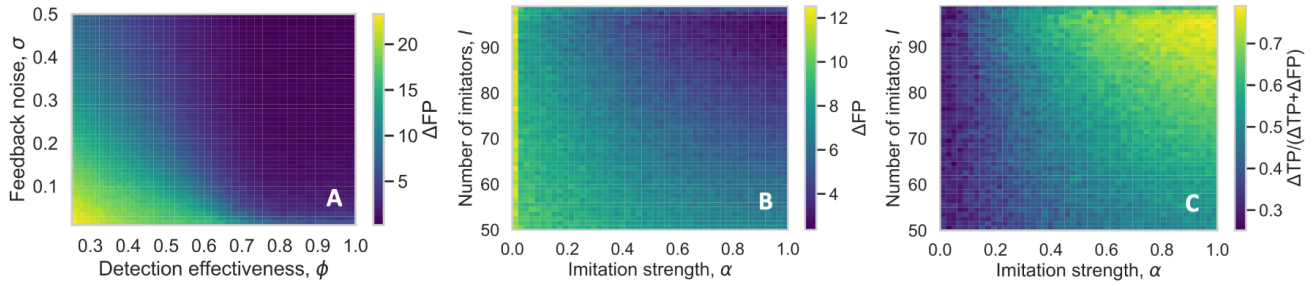

Figure 4: **A**) The role of detection effectiveness in mitigating the risks of transparent feedback. We report the difference between the number of False Positives after and before the strategic adaptation, i.e. ΔFP. **B**) The role of imitation. Increasing the number of imitators (I, vertical axis) and imitation strength (α, horizontal axis) reduces ΔFP. **C**) Imitation also contributes to increase the variation in True Positives (ΔTP) after strategic adaptation, relative to the variation in False Positives, i.e., $\Delta TP/(\Delta TP + \Delta FP)$. Parameters used: $N = 100$, $b = 1.0$, $c_\iota = 3.0$, $\epsilon = 0$, $k = 0.5$, $\alpha = 0$ (Panel **A**) and $\sigma = 0.001$, $\phi = 0.7$ (Panels **B** and **C**). Average over 400 runs.

Transparent feedback is ethically desirable as part of a responsible corporate practice: a recourse should be provided by the Institution to the Individuals. Thus, our analysis investigates how to identify mechanisms to mitigate the risks associated with sharing transparent feedback. One of these mechanisms is detecting faking behaviors, which can be performed with varying levels of effectiveness. As defined in Eq. (6), we implement detection effectiveness via the scaling factor ϕ that relates detection probability to the magnitude of the faking behavior (Fig. 3).

Fig. 4**A** reports the difference in the number of FP after and before strategic adaptation: the risks of transparent feedback can be mitigated if the Institution implements *effective (faking) detection mechanisms*. The proposed modelling approach captures this via the detection probability $d[f(i,t)]$ given in Eq. (6) and further explained in Fig. 3: reducing feedback noise (vertical axis) can increase the number of False Positives FP; however, such risk is contingent on the effectiveness of the detection mechanisms used. If detection is effective (high ϕ, horizontal axis), providing transparent feedback does not increase FP.

Utility maximization, strategic adaptation and imitation. Individuals can resort to utility maximization and social learning alike, in order to adapt [Banerjee, 1992; Bikhchandani *et al.*, 1992]. In Fig. 4**B** and **C** we assume that a set I of individuals (Imitators) are going to be influenced by the behavior of others in the population. As such, they will place some weight on the adaptation process of the individuals they observe. We assume that imitators in the population adapt with an imitation strength α (Eq. 10).

We let parameters (I, α) vary in Fig. 4**B** and **C**. We can observe that a larger pool of imitators (high I, vertical axis) and high imitation strength (high α, horizontal axis) reduce the increase in the number of False Positives after strategic adaptation. As an intuition, this might happen given that imitation has a different impact on those close and further away from the decision boundary: Individuals that are closer to the boundary (and who can easily increase x_2 to be classified as positive) alter their observable feature to a lesser extent by imitating those that are too far away from the decision boundary and that do not even attempt to adapt. On the other hand, individuals that are far away from the boundary might imitate those that are closer to the boundary (thus adapting by slightly increasing x_2), leading to an overall increase in faking behaviors. Because these individuals are too far from the decision threshold, however, even if they fake through imitation they will not adapt enough to be classified as positive. In online Supplementary Information we show that this conclusion extends to other values of ϕ, ϵ and more complex scenarios.

4 Conclusion

We propose a framework to explore the interplay between explanations and strategic adaptation by Individuals within the context of a generic classification model. Given the increased use of AI in multiple high-stake domains, this approach helps assessing key aspects related to strategic classification and their implications in terms of ethical AI and risk management. By considering interactions among multiple stakeholders (Individuals and Institution) and Individuals' social embedding (through imitation), this work contributes to a recent scientific trend in designing ethical multiagent systems taking into account their broader socio-technical context [Murukannaiah *et al.*, 2020; Chopra and Singh, 2018]. We present an illustrative example on credit applications and a critical analysis of the dilemma therein. Based on the simulation study, our results highlight that behavior imitation (prevalent in society) and fraud detection capacity are two key factors to shape the dynamics in this context. This suggests 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 behavior; studying how biases impact imitation processes in strategic adaptation to algorithms [Santos *et al.*, 2021]; or even relating the faking/improving dilemma with cooperation and fairness dilemmas, whose mechanisms have been extensively studied in evolutionary game theory and human-AI interactions [Santos *et al.*, 2019; Shirado and Christakis, 2020; Domingos *et al.*, 2021].

Acknowledgments

This research was supported by the Innovation Center for AI (ICAI). Disclaimer: The views expressed in this paper are solely those of the authors and do not necessarily represent the views of their current or previous employers.

References

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

[Bambauer and Zarsky, 2018] Jane Bambauer and Tal Zarsky. The algorithm game. *Notre Dame L. Rev.*, 94:1, 2018.

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

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

[Barsotti *et al.*, 2022] Flavia Barsotti, Rüya Gökhan Koçer, and Fernando P Santos. Can algorithms be explained without compromising efficiency? The benefits of detection and imitation in strategic classification. In *Proc. of AAMAS 2022*, pages 1536–1538, 2022.

[Bechavod *et al.*, 2021] Yahav Bechavod, Chara Podimata, Zhiwei Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. *arXiv preprint arXiv:2103.01028*, 2021.

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

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

[Citron and Pasquale, 2014] Danielle Keats Citron and Frank Pasquale. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89:1, 2014.

[Dalvi *et al.*, 2004] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proc. of the 10th ACM SIGKDD*, pages 99–108, 2004.

[Dekel *et al.*, 2010] Ofer Dekel, Felix Fischer, and Ariel D Procaccia. Incentive compatible regression learning. *J. Comput. Syst. Sci.*, 76(8):759–777, 2010.

[Domingos *et al.*, 2021] Elias Fernández Domingos, Jelena Grujić, Juan C Burguillo, Francisco C Santos, and Tom Lenaerts. Modeling behavioral experiments on uncertainty and cooperation with population-based reinforcement learning. *Simul. Model. Pract. Theory.*, 109:102299, 2021.

[Ghalme *et al.*, 2021] Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *Proceedings of ICML 2021*, volume 139, pages 3672–3681. PMLR, 18–24 Jul 2021.

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

[Hardt *et al.*, 2016] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proc. of ITCS '16*, pages 111–122, 2016.

[Hu *et al.*, 2019] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proc. of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

[Jagadeesan *et al.*, 2021] Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. Alternative microfoundations for strategic classification. In *Proc. of ICML 2021*, pages 4687–4697. PMLR, 2021.

[Kleinberg and Raghavan, 2020] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

[Krishnaswamy *et al.*, 2021] Anilesh K. Krishnaswamy, Haoming Li, David Rein, Hanrui Zhang, and Vincent Conitzer. Classification with strategically withheld data. *Proceedings of AAAI 2021*, 35(6):5514–5522, May 2021.

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

[Miller *et al.*, 2020] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proc. of ICML 2020*, pages 6917–6926, 2020.

[Molnar, 2020] Christoph Molnar. *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/, 2020.

[Murukannaiah *et al.*, 2020] Pradeep K Murukannaiah, Nirav Ajmeri, Catholijn M Jonker, and Munindar P Singh. New foundations of ethical multiagent systems. In *Proceedings of AAMAS 2020*, pages 1706–1710, 2020.

[Santos *et al.*, 2019] Fernando P Santos, Jorge M Pacheco, Ana Paiva, and Francisco C Santos. Evolution of collective fairness in hybrid populations of humans and agents. In *Proceedings of AAAI 2019*, pages 6146–6153, 2019.

[Santos *et al.*, 2021] Fernando P Santos, Simon A Levin, and Vítor V Vasconcelos. Biased perceptions explain collective action deadlocks and suggest new mechanisms to prompt cooperation. *iScience*, 24(4):102375, 2021.

[Sen and Airiau, 2007] Sandip Sen and Stephane Airiau. Emergence of norms through social learning. In *Proceedings of IJCAI 2007*, volume 1507, page 1512, 2007.

[Shirado and Christakis, 2020] Hirokazu Shirado and Nicholas A Christakis. Network engineering using autonomous agents increases cooperation in human groups. *iScience*, 23(9):101438, 2020.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

[Yu *et al.*, 2013] Han Yu, Zhiqi Shen, Cyril Leung, Chunyan Miao, and Victor R Lesser. A survey of multi-agent trust management systems. *IEEE Access*, 1:35–50, 2013.