Population size effects on strategic classification dynamics

Marta C. Couto
Informatics Institute, UvA
Amsterdam, The Netherlands
m.gomesdacunhacouto@uva.nl

Flavia Barsotti
ING Group & TU Delft (DIAM)
Amsterdam, The Netherlands
flavia.barsotti@ing.com;f.barsotti@tudelft.nl

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

ABSTRACT

Classification algorithms are increasingly used in high-stakes domains such as healthcare and finance. Individuals affected by these decisions can strategically adapt to obtain favourable outcomes. This may create data distribution shifts that require algorithm retraining. The resulting feedback loop between user adaptation and institutional updates fundamentally shapes both algorithmic accuracy and fairness; yet, these dynamics remain under-explored. We propose evolutionary game theory as a rigorous framework to study these multiagent dynamics. This approach provides insights into which strategies — such as levels of user honesty and algorithmic decision thresholds — persist in the long run. Our key contribution is to formally consider finite populations of users and institutions. We explore population size asymmetries, reflecting real-world settings where a large user population interacts with a comparatively small set of institutions, and different numbers of institutions translate different competitive landscapes. In alignment with previous results, our model reveals that when moderate institutions use algorithms susceptible to manipulation users either game the system or incur excessive costs to meet institutional standards. We also show, perhaps counter-intuitively, that unfavourable states for users become more prevalent when the institutions' population becomes smaller. The effect of population size is not always monotonic, revealing a new layer of complexity to strategic classification. Notably, the potential negative impact on users persists even under the highly idealised scenario of manipulation-proof classifiers, which generally reduces social costs and encourages honest user improvement. Overall, our findings indicate that population size asymmetries, ubiquitous in practice, play a critical and non-trivial role in shaping the dynamics of strategic classification, and should be considered when assessing algorithmic accuracy and fairness in the long run.

KEYWORDS

Multiagent dynamics; Evolutionary Game Theory; Stochastic Population Dynamics; Markov Chain; Explainability; Strategic Classification; Mixed-motive Games; Socio-technical Systems.

ACM Reference Format:
Marta C. Couto, Flavia Barsotti, and Fernando P. Santos. 2026. Population size effects on strategic classification dynamics. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/CKRS7125

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/CKRS7125

1 INTRODUCTION

Classification algorithms are widely used in multiple contexts and can often severely impact individuals' well-being [30], especially when considering high-stakes domains, such as healthcare, credit lending, and fraud detection. When individuals have knowledge about the classification algorithm, they may strategically adapt their features or actions to obtain favourable outcomes, as reported in prior works [7]. Users' adaptation is desirable when it leads to honest improvement and enables *compliant* algorithmic recourse [12]. However, some users may also "game" the system by manipulating data (e.g., providing false information) [1, 3, 22]. This is risky for institutions, as users' behaviour introduces data distribution shifts, affecting the accuracy of model predictions [11, 29, 36]. To reduce the impacts on algorithmic performance, institutions can impose additional requirements on applicants. In turn, this can result in an exaggerated burden on users [6, 18, 26]. The field of *strategic classification* studies this complex interaction [16].

In their seminal paper, Hardt et al. (2016) introduce the problem of strategic classification and the challenge of designing strategy-robust classifiers that maintain performance after users' adaptation [16]. Other works focus on how classifiers can incentivise users to honestly improve [24]. Interactions between users and institutions are framed as a game, assuming that: i) users perfectly know the classifiers and can *best-respond* to them, and ii) institutions use this assumption about users to build strategy-robust algorithms or create incentives for improvement. In real settings, however, users might not have full information about classifiers and their behavior can be hard to anticipate by institutions. More recent works relax those assumptions, introducing noisy and social learning in modelling users' behaviour [4–6, 15, 20, 37]. Other works study the effect of various levels of algorithmic transparency [1, 13, 32]. Interestingly, while transparency may leave room for strategic behaviour, full opaqueness can also be detrimental, even for institutions [13].

Most previous studies model the interaction as a single-step process: an institution deploys an algorithm and users respond once. This simplification might overlook the dynamical nature of the relation between algorithms and their users — one that typically unfolds over multiple rounds of continuous co-adaptation [11]. Only a few works take this into account [8, 31, 37]. Some of these take a population dynamics approach [8, 31]. By means of *replicator dynamics*, Saig & Rosenfeld (2025) study the long-term co-evolution of model performance and user population composition. They consider a single institution and a population of users (one-to-many) evolving through *natural selection*, assuming that fitness depends on how adequately the institution classifies users [31]. Instead, Couto et al. (2025) consider both a population of users and a population of institutions (many-to-many) [8]. This allows to account for (i) direct competition among institutions, and (ii) potential conflicts of interest between users and institutions, an

aspect relevant in practice when unqualified users can benefit from low algorithmic precision. However, that work uses a deterministic method by assuming infinite populations. Real populations are finite, which may introduce stochastic effects with consequences that are often unexpected for long-term outcomes [19, 27].

Here, we study the dynamics of strategic classification in finite populations. In particular, we pose the question: **What is the effect of having different population sizes, reflecting real-world settings where a large user population interacts with a comparatively small set of institutions**?

We consider that institutions can set a medium or high classification threshold, suggesting how strict they are in terms of requirements in the application process. In turn, users can invest in an honest or deceitful application, at a high and low cost, respectively, or even not make any investment. Figure 1 summarizes our model. We observe that when moderate institutions (medium classification threshold) use algorithms that are vulnerable to manipulation, users end up either gaming the system by misreporting personal information or incurring excessive costs to meet institutional standards. In this scenario, unfavourable states for users can be amplified when the institutional population is small, a result which is relevant and at the same time counter-intuitive. Yet, the effect of the population size is not monotonic, resulting from an intricate interaction of factors. Interestingly, our results highlight that negative impacts may persist even under a scenario with manipulation-proof classifiers, which should generally reduce social costs and encourage honest user improvement. Overall, our findings demonstrate that population size asymmetries, which are common in practice but often overlooked in the literature, have a critical impact on the long-term dynamics of strategic classification.

2 MODEL AND METHODS

2.1 The general game

We introduce a game between two types of players, the *institution* and the *user*. The institution aims to accurately classify the user to provide a service. As a running example, we will often assume that the institution can be a bank deploying an algorithm to decide whether to give a loan or not to a customer. However, examples of strategic classification exist in multiple domains such as hiring [10], recommender systems [23], collective action dilemmas [14] or even predictive policing [3].

We assume that the algorithm deployed by the bank takes as input user's self-reported features (or attributes) in the application process. On the other hand, the bank sets a threshold on the credit score to classify customers' applications as positive (i.e., getting the credit) or negative (i.e., not getting the credit).

We assume that an institution may choose to set its score threshold at two different levels, **Medium** (**M**) or **High** (**H**) (horizontal lines in Figure 1B). In other words, institutions can be moderate or harsh, respectively, where a harsh institution only accepts high quality applications, while a moderate institution has a less stringent set of requirements. Formally, we define the institutions' strategy set S_I as

$$S^I := \{S_i^I, i = 1, 2\} := \{\textbf{Medium}, \textbf{High}\}. \quad (1)$$

In turn, users' strategies correspond to adaptations in their feature space, which may affect their score. Let $z_1(\tau) \in [0, 1]$ be a (normalized) true feature value (e.g., real income), and $z_2(\tau) \in [0, 1]$ be a (normalized) observable feature value (e.g., declared income) at moment τ, where $\tau \in \{0, 1\}$. We can equate z_1 as the true score and z_2 as the predicted score by the institution. In Figure 1B, we represent z_1 on the x-axis and z_2 on the y-axis. The time $\tau = 0$ refers to the initial condition (before actions are taken), and $\tau = 1$ refers to the moment after actions are taken. We assume that initially the observed feature corresponds to the real one, that is, $z_1(0) = z_2(0)$ — in Figure 1B, individuals initially lie on the diagonal (black avatars). Only after actions are taken, these quantities may differ. Moreover, when the true and observable features do not match, i.e. $z_1(1) \neq z_2(1)$, it means that an individual provided false information.

Let θ be the success threshold linked to a user's true success or quality (dashed line in Figure 1B), such that, for any τ, if $z_1(\tau) \geq \theta$, the user is successful (e.g. can repay a loan); if $z_1(\tau) < \theta$, the user is not successful (e.g. is not able to repay a loan). Additionally, we distinguish users in terms of their true quality prior to any action, i.e. at $\tau = 0$. We denote the user type by u, where $u \in \{G, B\}$. If $z_1(0) \geq \theta$, the user is denoted by Good ($u = G$), while for $z_1(0) < \theta$, the user is denoted by Bad ($u = B$). Hence, in Figure 1B, the Bad user is located left of the dashed line, and the Good user is on the right (black avatars).

Users can choose from different actions when interacting with an institution (e.g., when applying for a bank loan). When providing personal information, users can possibly (i) disclose their features as they are (i.e. without any adaptation) at no cost, (ii) fake or misreport their features at a low cost, or (iii) improve their features and report them truthfully at a high cost. Here, we assume that each type of user can choose from two actions. As shown in [8], this simplification does not affect the results significantly. Formally, the set of strategies S^G and S^B of the Good and Bad users are, respectively,

$$S^G := \{S_i^G, i = 1, 2\} := \{\textbf{Not adapt}, \textbf{Adapt}\}, \quad (2)$$

$$S^B := \{S_i^B, i = 1, 2\} := \{\textbf{Fake}, \textbf{Improve}\}. \quad (3)$$

A Good user that does **Not adapt** (**N**), keeps their initial score, $z_1(0) = z_1(1) = z_2(1)$, and incurs no cost. When deciding to **Adapt** (**A**), the user increases their true score, $z_1(0) < z_1(1) = z_2(1)$, at some cost[1]. As for a Bad user, they can choose to **Fake** (**F**) or to **Improve** (**I**). To **Fake** means appearing as having a higher score than they truly have, $z_1(0) = z_1(1) < z_2(1)$, that is, to provide false or manipulated information at a low cost. To **Improve** means to increase one's true score, $z_1(0) < z_1(1) = z_2(1)$ at a high cost. In Figure 1B, the four different actions are depicted by the coloured arrows, which show the variation in the feature space (z_1, z_2) from $\tau = 0$ to $\tau = 1$.

[1]We could interpret **Adapt** as truthful adaptation (equivalent to **Improve**) or not (equivalent to **Fake**). For our purposes, it only matters whether we observe some costly adaptation (compared to none at no cost) because, regardless of its type, the impact on the institution is the same. For clarity, here we assume this adaptation to be truthful, but we denote it by **Adapt** to eliminate ambiguity about the type of user we refer to.

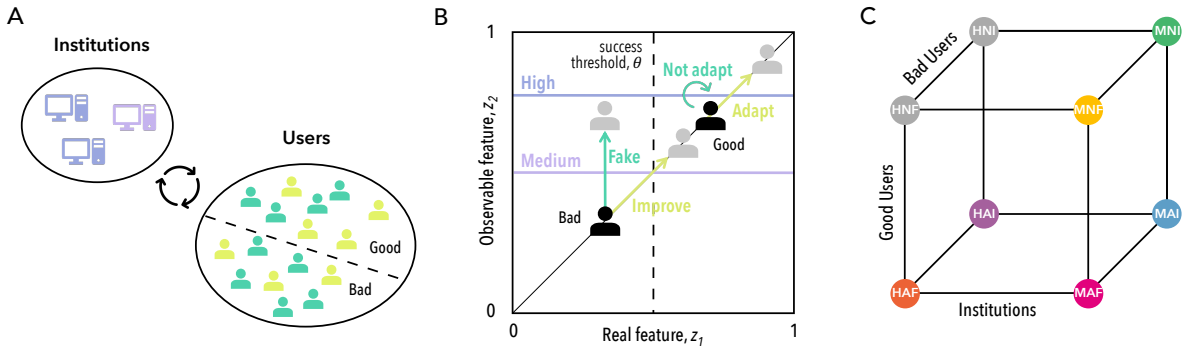

Figure 1: Model illustration. (A) Co-adaptation dynamics of strategic classification. A relatively small population of institutions (using classification algorithms) interacts with a large population of users (being classified). The population of users is composed of two sub-populations corresponding to Good and Bad users, having high and low credit scores, respectively. The strategies that each individual can adopt co-evolve based on their relative average success. (B) Feature space and strategies. The horizontal axis represents the user's real feature (z_1); the vertical axis represents the user's feature observable by the institution (z_2) and used in the algorithmic classification. For simplicity, we depict 1-dimensional features. We represent the institution's strategies (decision boundary) by coloured horizontal lines, Medium and High. The dashed vertical line denotes the success threshold θ determining a user's true success. A Good user has their initial real feature above the success threshold, i.e. $z_1(0) > \theta$. On the contrary, a Bad user has their initial real feature below the success threshold, i.e. $z_1(0) < \theta$. We represent the users' strategies by coloured arrows, which express a change in the feature space. A Good user can decide to Adapt (paying a cost to improve their condition) or Not adapt (remain as they are at no cost). A Bad user can decide to Improve (paying a cost to truthfully improve their initial condition, i.e., to change their real features) or Fake (paying a lower cost to manipulate their initial condition, i.e., to change only their observable features). The black avatars refer to the initial condition ($\tau = 0$), whereas the grey avatars refer to the final condition ($\tau = 1$), i.e. before and after adaptation. (C) We represent the state space $\{0, 1, \ldots, N^I\} \times \{0, 1, \ldots, N^G\} \times \{0, 1, \ldots, N^B\}$ by a cube, each dimension corresponding to a different type of player. The vertices (coloured circles) denote the homogeneous states, from the set S in Equation (13), or {(H)igh, (M)edium} × {(A)dapt, (N)ot Adapt} × {(F)ake, (I)mprove}, where each sub-population is composed of a single strategy. All other points of the state space denote non-homogeneous states (not explicitly shown for simplicity).

The interaction between the institutions and the users is represented by a general game with the following scheme

	Good		Bad	
	Not adapt	**Adapt**	**Fake**	**Improve**
Medium	$\mathcal{I}^G_{11}, \mathcal{U}^G_{11}$	$\mathcal{I}^G_{12}, \mathcal{U}^G_{12}$	$\mathcal{I}^B_{11}, \mathcal{U}^B_{11}$	$\mathcal{I}^B_{12}, \mathcal{U}^B_{12}$
High	$\mathcal{I}^G_{21}, \mathcal{U}^G_{21}$	$\mathcal{I}^G_{22}, \mathcal{U}^G_{22}$	$\mathcal{I}^B_{21}, \mathcal{U}^B_{21}$	$\mathcal{I}^B_{22}, \mathcal{U}^B_{22}$

, (4)

where the rows correspond to the institution's decisions, and the columns correspond to each user type's decisions. $\mathcal{I}^G_{ij}$ denotes the payoff that the institution obtains when playing strategy S^I_i against a Good user playing strategy S^G_j, and $\mathcal{U}^G_{ij}$ denotes the payoff that a Good user obtains when playing strategy S^G_j against an institution playing strategy S^I_i, where $i = 1, 2$ and $j = 1, 2$. The remaining payoffs are analogously defined.

Note that by construction, the institution does not know whether it faces a Good or Bad user — that is what the institution is trying to predict in the first place. Thus, the institution's strategy holds for both types of players (i.e. the classification algorithm deployed by the institution applies to all users simultaneously). Moreover, unlike previous work [16], our setup avoids the strong assumption that users know the classifiers perfectly. Instead, we consider a simultaneous game, where users decide between being "generally" honest or not (for Bad users), and adapting or not (for Good users), with fixed costs for each action. This setup provides a general formulation of our problem. Next, we define two particular cases.

2.2 The baseline scenario: imperfect classifiers

We now build a baseline case of the game introduced in Section 2.1. Before assigning payoffs to the game, it is useful to reason in terms of classification outcomes and formalise it by the following matrix

	Good		Bad	
	Not adapt	**Adapt**	**Fake**	**Improve**
M	TP	TP	FP	TP
H	FN	TP	TN	FN

. (5)

We start from the point of view of a **High** institution. A **High** institution sets its threshold so high that it only accepts Good users who **Adapt**, yielding a true positive (TP) (Figure 1B). A Good user that does **Not adapt** is classified as negative (FN). At the same time, a **High** institution can accurately classify faking behaviour from a Bad user (TN). Still, it would decline a Bad user that improves (FN). An institution with a **Medium** threshold always accepts Good users (TP). It also accepts Bad users, yielding a false positive (FP) if users **Fake** and a TP if they **Improve**.

We can now assign actual payoffs to each outcome. We assume that an institution earns a payoff of ρ in case of a TP. For example,

ρ can denote a bank's gain from a successful lending operation, associated with the loan's interest rate. In the case of a false positive FP, we assume the institution has a loss measured by λ. When the institution rejects a user's application and thus the user is classified as negative, we assume the institution gets a null payoff. As for the user, we assume they obtain a benefit b when they are accepted (classified as positive), both in the case of a TP or a FP. Moreover, the user pays a cost c_I for adapting or improving and a cost c_F for faking. Hence, we write the payoff matrices of the game described in Equation (5) as

	Good		Bad	
	Not adapt	**Adapt**	**Fake**	**Improve**
M	ρ, b	$\rho, b - c_I$	$-\lambda, b - c_F$	$\rho, b - c_I$
H	0, 0	$\rho, b - c_I$	$0, -c_F$	$0, -c_I$

(6)

where all parameters ρ, λ, b, c_I and c_F take real positive values. As stated before, we consider the cost of improvement higher than the cost of faking, i.e., $c_I > c_F$. Additionally, we assume that (i) $b > c_I$ (the cost of improving is never greater than the benefit of receiving a loan, irrespective of being a Good or Bad user) and (ii) $\lambda > \rho$ (the loss for the institution associated with a FP is greater than the benefit of a TP).

2.3 The robust scenario: manipulation-proof classifiers

We now consider a scenario where **Medium** institutions are *robust* against faking behaviour. For that, we simply change one entry of the previous matrix in Equation (5). We do so assuming that when a **Medium** institution meets a **Fake** user, it classifies them as negative, yielding a TN. The new outcome matrix becomes

	Good		Bad	
	Not adapt	**Adapt**	**Fake**	**Improve**
M	TP	TP	***TN***	TP
H	FN	TP	TN	FN

(7)

where the only difference to the baseline scenario is in bold. Similarly, in terms of payoffs, we write

	Good		Bad	
	Not adapt	**Adapt**	**Fake**	**Improve**
M	ρ, b	$\rho, b - c_I$	$0, -c_F$	$\rho, b - c_I$
H	0, 0	$\rho, b - c_I$	$0, -c_F$	$0, -c_I$

(8)

We keep the same assumptions for the parameters as above.

2.4 Evolutionary dynamics

The previous section defines the games of interest. Here, we describe the behavioural dynamics that models how individuals adapt their strategies over time. We apply tools from evolutionary game theory, a natural and rigorous framework for modelling strategic dynamics [17]. This allows us to identify which strategies can emerge and persist in the long run.

We consider finite populations of players: one population of institutions of size N^I and two populations of users, Good and Bad, of sizes N^G and N^B, respectively. We denote the total population size by $N \equiv N^I + N^G + N^B$.

We define the state of the population by a vector $\mathbf{s} = (n^I, n^G, n^B)$, where n^I is the number of institutions playing strategy S_1^I, n^G is the number of Good users playing strategy S_1^G, and n^B is the number of Bad users playing strategy S_1^B. Therefore, the state space of the system is the lattice $\{0, 1, \ldots, N^I\} \times \{0, 1, \ldots, N^G\} \times \{0, 1, \ldots, N^B\}$ (illustrated with a cube in Figure 1C). We assume that the populations' sizes are constant; therefore, $\mathbf{s}$ fully describes the state of the system, as the number of institutions playing strategy S_2^I is simply given by $N^I - n^I$. Likewise for the populations of users.

Let $f_i^I(\mathbf{s})$ denote the expected payoff or fitness obtained by an institution when implementing strategy S_i^I. We assume well-mixed populations of players, such that any institution can interact with any user with the same likelihood. Thus, we can write the expected payoff $f_i^I(\mathbf{s})$ as

$$f_i^I(\mathbf{s}) = \sum_{u \in \{G,B\}} \frac{N^u}{N^G + N^B} \cdot \frac{\mathcal{I}_{i1}^u \, n^u + \mathcal{I}_{i2}^u \, (N^u - n^u)}{N^u}. \tag{9}$$

Observe that, for $u = G$, the ratio

$$p_G \equiv N^G/(N^G + N^B) \in [0, 1] \tag{10}$$

represents the proportion of Good users with respect to the population of users.

Similarly, let $f_j^u(\mathbf{s})$ denote the expected payoff obtained by a u-type user implementing strategy S_j^u, where $u = G, B$. We have

$$f_j^u(\mathbf{s}) = \frac{\mathcal{U}_{1j}^u \, n^I + \mathcal{U}_{2j}^u \, (N^I - n^I)}{N^I}. \tag{11}$$

To model the co-adaptation of the three populations, we consider a pairwise imitation process, where successful strategies (with relatively higher payoffs) are more likely to be imitated and thus spread throughout the populations. This represents a social learning mechanism that goes as follows. In each time step, one individual i is randomly selected from the whole population of players. This individual, whether an institution or a user, may revise its current strategy. For that, another individual j, of the same kind, is randomly selected as a potential role model. Individual i switches to individual j's strategy with probability P_{ij} [34, 35]

$$P_{ij} = \frac{1}{1 + e^{-\beta(f_j - f_i)}}, \tag{12}$$

where f_i and f_j are the expected payoffs of individuals i and j, respectively, and $\beta (\geq 0)$ is the selection strength. Parameter β determines how important the payoff differences are to the imitation process: if $\beta = 0$, players change their strategy with probability 0.5, regardless of the payoffs; if β is high, player i is very likely to imitate player j if j's payoff is higher than i's. Hence, β controls how noisy the imitation learning is. Since the strategy update only depends on the current state of the system, this process can be described by a Markov chain [21].

We also allow for mutations: with small probability μ, individuals switch to a random strategy instead of imitating others. Mutations prevent absorbing states by reintroducing absent strategies even in homogeneous populations (i.e., populations with a single strategy present). This ensures the system forms an ergodic Markov chain that converges to a unique stationary distribution, describing the long-run probability (or relative time) of the system being in each

state [2, 25]. In particular, the system has $(N^I+1)(N^G+1)(N^B+1)$ states. When considering, for example, population sizes of 100 individuals, that yields around 10^6 states. To avoid the difficulties associated with such a large state space, we resort to the widely used small mutation limit method [19]. In the small mutation limit, we assume that mutations are very rare. Consequently, the time between mutations allows the system to reach a homogeneous state and remain there long enough until a new mutation occurs (and possibly steers the system back to an interior state). This yields the time spent in non-homogeneous (interior) states as negligible. Thus, this approximation enables the reduction of the full state space to the 8 homogeneous states, which correspond to the corners of the state space depicted by the cube in Figure 1C. The set of homogeneous or monomorphic states S is

$$S = \{(0,0,0), (N^I,0,0), (0,N^G,0), (0,0,N^B), (N^I,N^G,0), (N^I,0,N^B), (0,N^G,N^B), (N^I,N^G,N^B)\} \tag{13}$$

We also use the notation S = {**HAI, MAI, HNI, HAF, MNI, MAF, HNF, MNF**} to highlight the strategy composition of the population at each homogeneous state. For example, the state $(0,0,0)$ is also denoted by **HAI**, where only the strategies **High** (**H**), **Adapt** (**A**) and **Improve** (**I**) are present in the respective sub-populations.

To define the reduced Markov chain, we need to calculate several transition probabilities. First, the transition probability $T^I_{\pm}(\mathbf{s})$ that the number of institutions playing S^I_1 increases or decreases by one at state $\mathbf{s} = (n^I, n^G, n^B)$ is given by

$$T^I_{\pm}(n^I,n^G,n^B) = \frac{N^I - n^I}{N^I}\frac{n^I}{N^I-1}\frac{1}{1+e^{\mp\beta(f^I_1(n^G,n^B)-f^I_2(n^G,n^B))}}. \tag{14}$$

Similarly, the transition probability $T^u_{\pm}(\mathbf{s})$ (where $u = G, B$) that the number of u-type users playing S^u_1 increases or decreases by one is given by

$$T^u_{\pm}(n^I,n^G,n^B) = \frac{N^u - n^u}{N^u}\frac{n^u}{N^u-1}\frac{1}{1+e^{\mp\beta(f^u_1(n^I)-f^u_2(n^I))}}. \tag{15}$$

The fixation probability denotes the probability that a single mutant of a given strategy completely takes over a resident population of the opposite strategy [27, 33]. Let us denote the fixation probabilities of strategies S^I_1, S^G_1 and S^B_1 in the respective populations as ρ^I_1, ρ^G_1 and ρ^B_1. We can write the fixation probabilities as [21, 35]

$$\begin{aligned}
\rho^I_1(n^G,n^B) &= \frac{1}{1+\sum_{k=1}^{N^I}\prod_{m=1}^{k}\frac{T^I_-(m,n^G,n^B)}{T^I_+(m,n^G,n^B)}},\\
\rho^G_1(n^I,n^B) &= \frac{1}{1+\sum_{k=1}^{N^G}\prod_{m=1}^{k}\frac{T^G_-(n^I,m,n^B)}{T^G_+(n^I,m,n^B)}},\\
\rho^B_1(n^I,n^G) &= \frac{1}{1+\sum_{k=1}^{N^B}\prod_{m=1}^{k}\frac{T^B_-(n^I,n^G,m)}{T^B_+(n^I,n^G,m)}},
\end{aligned} \tag{16}$$

where here n^I, n^G, and n^B can only take values of 0 or the respective population size, because in the small mutation limit, we assume that while fixation takes place in one population, the other two populations remain fixed and homogeneous. To write the fixation probabilities of strategies S^I_2, S^G_2 and S^B_2 — ρ^I_2, ρ^G_2 and ρ^B_2, respectively — we simply switch the signs "+" and "−" in the transition probabilities T in Equation (16). Each of these fixation probabilities is associated with one direction at an edge of the cubic state space (Figure 1C).

Finally, we write the transition matrix Λ of the reduced Markov chain. Each entry $\Lambda_{ss'}$ corresponds to the probability of going from state s to state s', with $s, s' \in S$ in Equation (13) and is given by

$$\Lambda_{ss'} = \begin{cases}
\frac{N^I}{N}\rho^I_1(n^G,n^B), & \text{if } s=(0,n^G,n^B) \text{ and } s'=(N^I,n^G,n^B)\\
\frac{N^I}{N}\rho^I_2(n^G,n^B), & \text{if } s=(N^I,n^G,n^B) \text{ and } s'=(0,n^G,n^B)\\
\frac{N^G}{N}\rho^G_1(n^I,n^B), & \text{if } s=(n^I,0,n^B) \text{ and } s'=(n^I,N^G,n^B)\\
\frac{N^G}{N}\rho^G_2(n^I,n^B), & \text{if } s=(n^I,N^G,n^B) \text{ and } s'=(n^I,0,n^B)\\
\frac{N^B}{N}\rho^B_1(n^I,n^G), & \text{if } s=(n^I,n^G,0) \text{ and } s'=(n^I,n^G,N^B)\\
\frac{N^B}{N}\rho^B_2(n^I,n^G), & \text{if } s=(n^I,n^G,N^B) \text{ and } s'=(n^I,n^G,0)\\
0, & \text{otherwise, except } s=s'\\
1-\sum_{s''=1;s\neq s''}^{8}\Lambda_{ss''} & \text{if } s=s'.
\end{cases} \tag{17}$$

The factors multiplying the fixation probabilities represent the likelihood that the single mutant appearing at a given monomorphic state is of the type driving the respective transition.

We obtain the stationary distribution vector $\mathbf{v}$ over the 8 states by solving the eigenvector equation $\mathbf{v} = \mathbf{v}\Lambda$. Again, the stationary distribution gives the prevalence or relative time spent in each state.

3 RESULTS AND DISCUSSION

We start by looking at the most likely directions of adaptation. For each edge of the state space, we calculate the direction in which the fixation probability is greater. For example, if $\rho^I_1(0,0) > \rho^I_2(0,0)$, we say that **Medium** institutions (S^I_1) are favoured to replace **High** institutions (S^I_2), when $n^G = n^B = 0$ and indicate this with an arrow pointing from state **HAI** (purple circle) to **MAI** (blue circle) in Figure 2. A dashed edge indicates equal fixation probabilities in both directions.

For the imperfect classifier scenario introduced in Section 2.2, the only state for which all three adjacent edges point in its direction for the entire parameter space is **HAF** (orange circle) (Figure 2 left). This makes **HAF** a highly robust state (where institutions have **High** thresholds, Good users **Adapt** and Bad users **Fake**). From the point of view of the users, this is not very desirable — Good users pay excessive costs to be accepted and Bad users are rejected. Furthermore, all adjacent edges point towards the state **MNF** (yellow circle), under the condition $p_G > p^*_G \equiv \lambda/(\lambda+\rho)$, with p_G in Equation (10). This means that also a state with **Medium** institutions, where Good users do **Not adapt** and Bad users **Fake**, is robust, if there are sufficiently many Good users. This state is now more favourable to the users because Good users are not required to incur the adaptation cost to be accepted, and Bad users are also accepted. However, this is less beneficial to the institutions as they bear the costs of false positives (λ). This scenario adequately captures the fundamental problem of strategic classification — it is hard to achieve states that are advantageous to both institutions and users, that is, where algorithmic performance and fairness coexist.

We now see how the manipulation-proof scenario introduced in Section 2.3 changes these dynamics (Figure 2 right). First, one of the edges that previously pointed to the state **HAF** (starting from **MAF**) is now neutral, which makes this state less robust compared

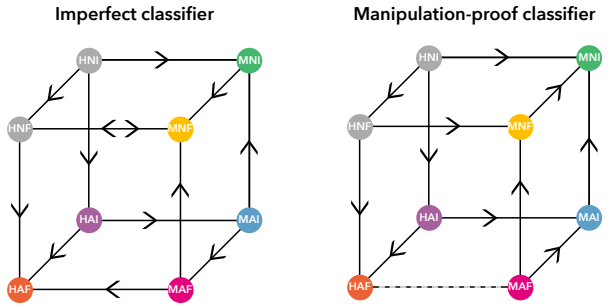

Figure 2: Transitions of the stochastic dynamics. We study two scenarios: The first one (left) corresponds to the *imperfect classifier*, illustrating the problem of strategic classification. The second scenario (right) assumes that it is possible to build a *manipulation-proof classifier* without an extremely high decision threshold. The figure shows the most probable directions of adaptation: the arrows indicate, for each edge of the state space, the direction of the larger fixation probability; the dashed line represents a neutral fixation probability, where transitions in both directions are equally likely. All directions are independent of the exact parameter values except for the transition between states MNF and HNF: the transition from HNF to MNF (rightwards) is more likely than the transition from MNF to HNF (leftwards) if $p_G > \lambda/(\lambda+\rho)$, with p_G in Equation (10). The opposite is true if $p_G < \lambda/(\lambda+\rho)$.

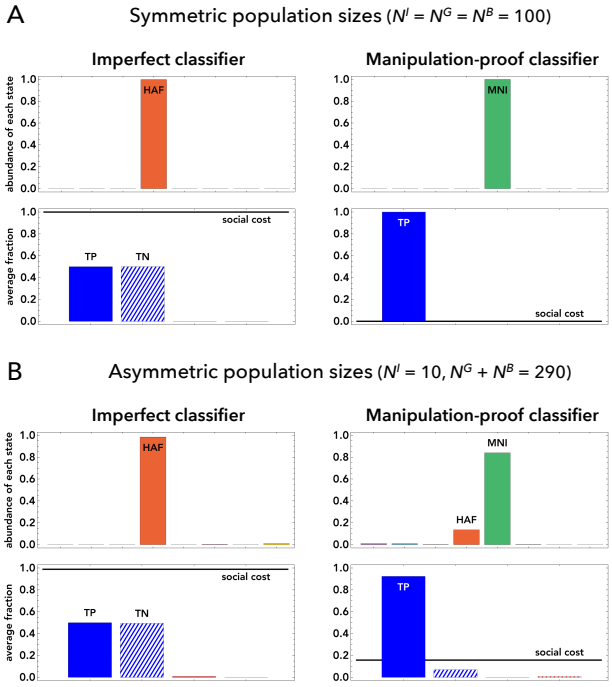

Figure 3: Stochastic dynamics in the small mutation limit for $p_G < p_G^* \equiv \lambda/(\lambda+\rho)$, p_G in Equation (10): stationary distribution over all homogeneous states (above panels) and the respective average metrics, algorithmic performance and social cost (below panels). The algorithmic performance is given by the fraction of TP, TN, FP and FN cases (in bars); the social cost imposed on users is given by the fraction of Good users that Adapt (black line). (A) Symmetric population sizes, $N_I = N_G = N_B = 100$. (B) Asymmetric population sizes, $N_I = 10, N_G = N_B = 200$. A small population of institutions makes the state HAF (red) more prevalent for the manipulation-proof scenario. Consequently, we see an increase in the social cost, while the performance is not much affected. Parameters: $\rho = 15, \lambda = 50, b = 50, c_F = 1, c_I = 5, \beta = 0.02, N = 300, p_G = 0.5$.

to the previous scenario. Second, the state **MNI** (green circle) becomes highly robust, because Bad users are now incentivised to **Improve** (the previous **MNI→MNF** transition flips its direction). Additionally, the state **MNF** is unconditionally favoured in relation to **HNF** (contrary to the previous scenario in Figure 2 left). This also contributes to transitions that ultimately favour state **MNI**. The state **MNI** is collectively beneficial — the institutions have a **Medium** threshold, the Good users are therefore accepted without unjustified costs, and Bad users honestly improve. Naturally, this is a highly idealised scenario, where we assume that **Medium** institutions are capable of being robust against manipulation while simultaneously not having excessively high decision boundaries. This scenario serves to show how a simple change in the incentive structure significantly alters the overall dynamics.

We confirm these observations by calculating the stationary distributions over the 8 states for each scenario, for particular parameter values (Figure 3A). First, we consider that all populations have the same size ($N_I = N_G = N_B = 100$), which makes $p_G = 0.5$. For $\rho = 15$ and $\lambda = 50$, then we have $p_G < p_G^*$. The state **HAF** is the only prevalent state when classifiers are imperfect, whereas state **MNI** is the only prevalent state when classifiers are robust to manipulation. This result is consistent with the results obtained with infinite populations in a previous work [8]. Furthermore, for a more explicit comparison of the consequences of each scenario, we plot algorithmic performance (defined by the average fraction of TP, TN, FP and FN cases), and the social cost (defined by the average fraction of Good users that **Adapt**). We see that for both scenarios, the algorithmic performance is maximal (there are only true cases). However, in the imperfect classifier scenario, the social cost is also maximal, because Good players always **Adapt** to the **High** institutions.

We then consider a more realistic setting where the population of institutions is much smaller than the populations of users: we keep the total population size $N = 300$, and consider only 10 institutions. In Figure 3B, we see that the imperfect scenario almost does not change. However, for the manipulation-proof scenario, the state **HAF** now emerges about 14% of the time. This has no effect on the algorithmic performance, but it increases the social cost. Notably, the increase of a state which is detrimental to the users turns up even in the idealized scenario.

We now study what happens for $p_G > p_G^*$ (Figure 4). First, we see that state **MNF** is now much more prevalent. When the populations' sizes are symmetric (Figure 4A), the state **MNF** is even more predominant than **HAF**. While this is beneficial for the users, decreasing social cost substantially, it also increases the fraction

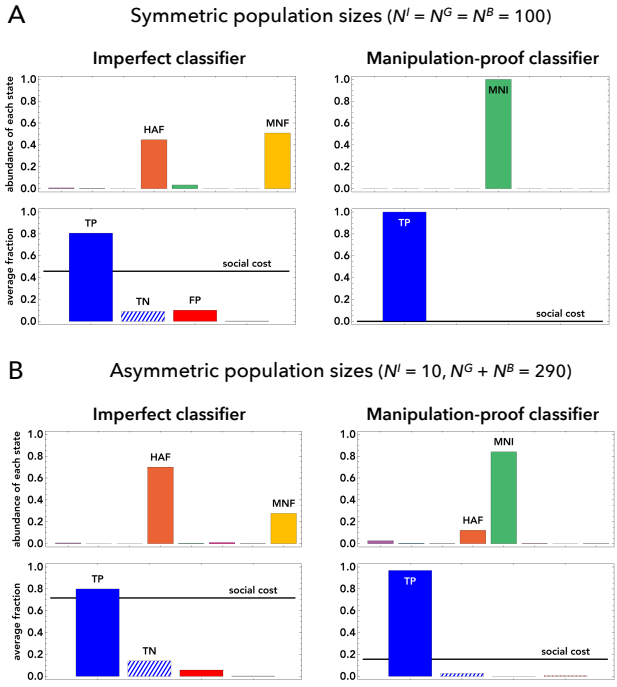

Figure 4: Stochastic dynamics in the small mutation limit for $p_G > p_G^*$. This figure is analogous to Figure 3, except that it stands for the high proportion of Good players (p_G) regime. Parameters: $\rho = 15, \lambda = 50, b = 50, c_F = 1, c_I = 5, \beta = 0.02, N = 300, p_G = 0.8$.

of false positives, partly compromising algorithmic performance. Regarding asymmetric population sizes case (Figure 4B), we again see a significant increase in the prevalence of state **HAF** and, consequently, an increase in social cost. As expected, there are no significant changes in the manipulation-proof scenario for this parameter regime, since the replacement directions on the right panel of Figure 2 are independent of the parameter values.

To better understand these outcomes, we first note that the impact of population sizes results from the interplay of several factors. On the one hand, as the institutions' population becomes smaller, a single institution can more easily fixate on the resident population, since fewer strategy switches are needed for one individual's strategy to completely replace the other, regardless of the specific strategies involved. On the other hand, the fewer individuals in a population, the fewer mutations are produced in that population. Specifically, the probability that a mutation occurs in a certain player type (institution, Good or Bad user) corresponds to the proportion of such type in the entire population — see Equation (17). Therefore, while a small population can facilitate transitions, it also curbs mutations that could trigger them. Thus, these two opposing contributions can override each other, rendering the effect of population sizes highly non-trivial.

Let us focus on the imperfect scenario in the left panels of Figures 2 and 4. We verify that the transition $\Lambda_{\mathbf{MNF\ HNF}}$ increases from the symmetric to the asymmetric population case, but not the opposite transition $\Lambda_{\mathbf{HNF\ MNF}}$. From the **HNF** state (grey circle), the transition to **HAF** is very likely. This can explain why the prevalence of state **HAF** increases at the cost of decreasing **MNF**.

For the manipulation-proof scenario (right panels of Figures 2 and 4), the mechanism is similar, but it affects different pathways. The transition $\Lambda_{\mathbf{MNI\ HNI}}$ increases from the symmetric to the asymmetric population case, but not the transition in the opposite direction. From the **HNI** state (grey circle), there are two likely pathways to **HAF**, through state **HNF** or state **HAI**. We observe the same mechanism for the $p_G < p_G^*$ regime (Figure 3A to B, right panels).

As the impact of population size results from the interplay of various factors and therefore can exhibit counter-intuitive behaviour, we examine a broader range of population sizes. Figure 5A shows the prevalence of each state across varying population sizes (while keeping the total population size constant). Indeed, we verify that for the imperfect classifier scenario (left), the effect of population size is non-monotonic — there is an intermediate institutions' population size (around $N^I \approx 30$) that maximises the prevalence of state **HAF**, and hence, the social cost. This occurs because similarly to the increase of transition $\Lambda_{\mathbf{MNF\ HNF}}$ for decreasing N_I, also the transition $\Lambda_{\mathbf{HAF\ MAF}}$ increases. From **MAF**, the transition to **MNF** is likely, which would explain an increase in **MNF** at the cost of **HAF**. Thus, the prevalence of these two states — **HAF** and **MNF** — depends on a fine balance of several transitions. There is a point where the pathway **HAF** → **MAF** → **MNF** starts dominating, hence the decrease in **HAF** for low institutions' population sizes.

As for the manipulation-proof scenario (right), the decrease in state **MNI** (for decreasing institutions' population size) occurs only around $N^I \approx 20$. Therefore, even in a highly robust scenario, with

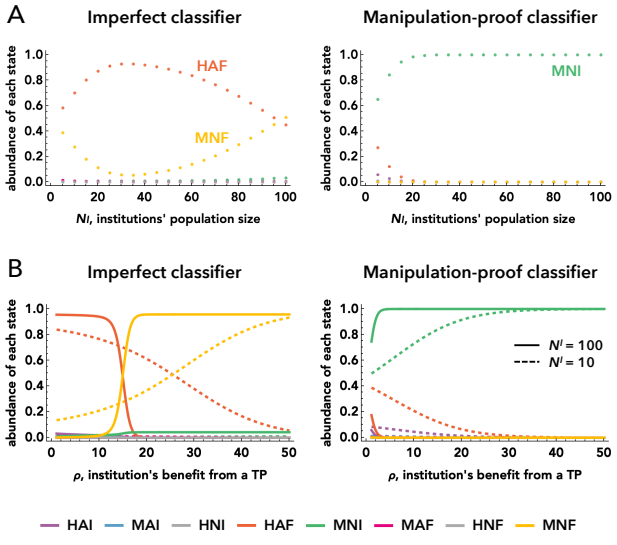

Figure 5: Abundance of each state for varying (A) population sizes N^I and (B) institution's benefit ρ from a true positive TP. In panel B, we also compare the case of symmetric population sizes ($N^I = 100$ in solid lines) and an asymmetric case ($N^I = 10$ in dashed lines). Fixed parameters: $\rho = 15$ **(panel A)**, $\lambda = 50, b = 50, c_F = 1, c_I = 5, \beta = 0.02, p_G = 0.8, N = 300$.

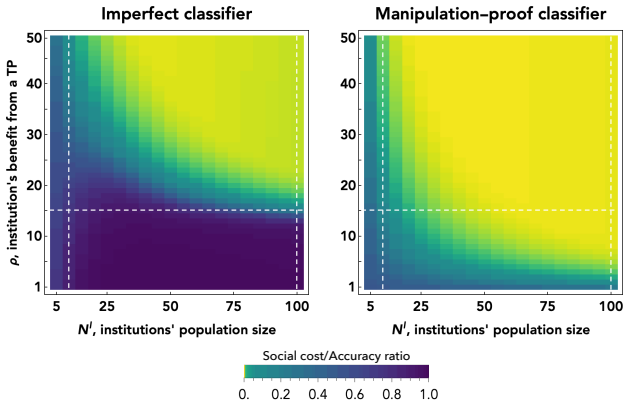

Figure 6: Social cost to algorithmic accuracy ratio for varying population sizes N^I and institution's benefit ρ from a true positive TP. We define accuracy as the sum of true positive and true negative rates. The white dashed lines indicate the parameters used in Figure 5. Fixed parameters: $\lambda = 50, b = 50, c_F = 1, c_I = 5, \beta = 0.02, p_G = 0.8, N = 300$.

manipulation-proof classifiers, a lending ecosystem with only a few institutions can significantly alter the dynamics to the detriment of users. In the presence of a small population of institutions, stochastic effects more easily drive the system away from the desirable state **MNI**. Once **High** institutions fixate (even if only due to chance), then the users rapidly transition to **Adapt** and **Fake** because those are their best-responses to harsh institutions.

In Figure 5B, we show the effect of the parameter ρ, that is, the benefit that an institution receives when it correctly accepts a user (e.g. the bank's gain from a successful loan repayment). The higher ρ, the more frequent the states **MNF** and **MNI** are, respectively, in the imperfect and robust scenarios. Therefore, a larger ρ tends to favour the users. Regarding the imperfect classifier regime, this is consistent with the observation in Figure 2 that state **MNF** becomes robust for $p_G > p_G^*$. This condition can be rewritten as $\rho > \lambda(1 - p_G)/p_G$. For the parameters in Figure 5B, the condition yields $\rho > 12.5$. Figure 5B (left) shows that when the population sizes are symmetric ($N^I = 100$ in solid lines), the state **MNF** starts replacing **HAF** close to $\rho = 12.5$. This state conversion is driven towards higher values of ρ for asymmetric population sizes ($N^I = 10$ in dashed lines). Furthermore, we see that the effect of population sizes depends on the parameter ρ.

To study the interplay between population asymmetry and ρ, we vary both in Figure 6. Here, we plot the summary metric *social cost to algorithmic accuracy ratio* (cost-to-accuracy ratio for short), where we define accuracy as the sum of true positive and true negative rates. As such, it is desirable to maintain low values for this metric (greenish to yellow tones in Figure 6). We confirm that, in general, high ρ and institutions' population size N^I lead to a low cost-to-accuracy ratio. However, for the imperfect scenario (left), we also verify that the impact of population size depends on ρ. For low ρ, decreasing N^I decreases the cost-to-accuracy ratio. For intermediate ρ, there is a maximum of the cost-to-accuracy ratio for relatively low to moderate values of N^I. For high values of ρ, the cost-to-accuracy ratio has a minimum for intermediate values of N^I. This, again, reveals the nuanced non-monotonicity of the problem. Finally, while having robust classifiers (right) greatly alleviates the problem of strategic classification, also here caution needs to be taken, as regions where the cost-to-accuracy ratio is not optimal still persist.

4 CONCLUSION

Classification algorithms are increasingly used for decision-making in many domains affecting society. We take credit lending as our main motivating case study, but our model could be applied to other domains where strategic classificaiton is relevant, for instance hiring [10], recommender systems [23], collective action dilemmas [14], spam detection [16] or even predictive policing [3]. In such domains, individuals can adapt over time and be strategic about their disclosed features. This, in turn, demands algorithm retraining, creating complex feedback loops between users and institutions. We propose a game theoretic model to study the long-term dynamics of co-adaptation between users and institutions. Newly, we assume that the co-adapting populations are finite, to take into account (i) stochastic effects and (ii) asymmetries in population sizes, both naturally present in real populations, where usually the number of users is much larger than the set of institutions they interact with. The resulting non-trivial effects could not be captured with deterministic, infinite population models. Our key results highlight that when moderate institutions use algorithms susceptible to manipulation, users either game the system by misreporting personal information or incur excessive costs. We notice that unfavourable states for users can be amplified when the institutions' population is small. This negative impact on users may persist even under idealized scenarios. Although the proposed model already incorporates increasingly complex assumptions, we recognize that other aspects are oversimplified. For example, we consider both binary types of users, Good and Bad, and binary strategies. An extension of this work could take into account users that are more or less close to the decision boundaries and continuous strategies. In such a setting, we could obtain a better approximation of the decision threshold that institutions would converge to. Another future direction would be to consider diverse intensities of selection, either between users and institutions, or even between different groups of users. By using an approach that resonates with the general idea of human-AI co-evolution [28], our work sheds light on the trade-off between algorithmic performance and fairness in strategic classification. In particular, our results suggest that population structure in the form of size asymmetry plays a critical role in the interaction between users and institutions. Hence, it should be taken into account when assessing algorithmic accuracy and fairness in the long run.

Software Availability: The source code used to generate all the results can be found in this open source OSF repository [9].

ACKNOWLEDGMENTS

Flavia Barsotti: ING Research Project Lead. For the authors Flavia Barsotti and Fernando P. Santos, the names are reported in alphabetical order. 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

[1] Emrah Akyol, Cedric Langbort, and Tamer Basar. 2016. Price of Transparency in Strategic Machine Learning. (2016). arXiv:1610.08210

[2] Tibor Antal and István Scheuring. 2006. Fixation of strategies for an evolutionary game in finite populations. *Bulletin of Mathematical Biology* 68, 8 (2006), 1923–1944.

[3] Jane Bambauer and Tal Zarsky. 2018. The Algorithm Game. *Notre Dame Law Review* 94 (2018), 1–48.

[4] Flavia Barsotti, Ruya G. Kocer, and Fernando P. Santos. 2022. Can Algorithms be Explained Without Compromising Efficiency? The Benefits of Detection and Imitation in Strategic Classification. In *International Conference on Autonomous Agents and Multiagent Systems*. 1536–1538.

[5] Flavia Barsotti, Ruya G. Kocer, and Fernando P. Santos. 2022. Transparency, Detection and Imitation in Strategic Classification. In *Proceedings of the 31st IJCAI International Joint Conference on Artificial Intelligence*. 67–73.

[6] Yahav Bechavod, Chara Podimata, Zhiwei Steven Wu, and Juba Ziani. 2022. Information Discrepancy in Strategic Learning. *Proceedings of Machine Learning Research* 162 (2022), 1691–1715.

[7] Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. 2020. Manipulation-Proof Machine Learning. *arXiv preprint* arXiv:2004.03865 (2020).

[8] Marta C. Couto, Flavia Barsotti, and Fernando P. Santos. 2025. Collective dynamics of strategic classification. (2025). arXiv:2508.09340

[9] Marta C. Couto, Flavia Barsotti, and Fernando P. Santos. 2026. Code Repository. https://doi.org/10.17605/OSF.IO/J2UXS.

[10] Lee Cohen, Connie Hong, Jack Hsieh, and Judy Hanwen Shen. 2025. Two Tickets are Better than One: Fair and Accurate Hiring Under Strategic LLM Manipulations. In *Forty-second International Conference on Machine Learning*.

[11] Sarah Dean, Evan Dong, Meena Jagadeesan, and Liu Leqi. 2025. Accounting for AI and Users Shaping One Another: The Role of Mathematical Models. *Transactions on Machine Learning Research* (2025).

[12] Hidde Fokkema, Damien Garreau, and Tim van Erven. 2024. The risks of recourse in binary classification. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 550–558.

[13] 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.

[14] António Góis, Mehrnaz Mofakhami, Fernando P Santos, Simon Lacoste-Julien, and Gauthier Gidel. 2025. Performative Prediction on Games and Mechanism Design. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1855–1863.

[15] Nika Haghtalab, Thodoris Lykouris, Sloan Nietert, and Alexander Wei. 2022. Learning in Stackelberg Games with Non-myopic Agents. In *Proceedings of the 23rd ACM Conference on Economics and Computation* (Boulder, CO, USA) *(EC '22)*. Association for Computing Machinery, New York, NY, USA, 917–918.

[16] 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.

[17] Josef Hofbauer and Karl Sigmund. 1998. *Evolutionary Games and Population Dynamics*. Cambridge Univ. Press, Cambridge, UK.

[18] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The Disparate Effects of Strategic Manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 259–268.

[19] Lorens A. Imhof, Drew Fudenberg, and Martin A. Nowak. 2005. Evolutionary cycles of cooperation and defection. *Proceedings of the National Academy of Sciences* 102, 31 (2005), 10797–10800.

[20] Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. 2021. Alternative Microfoundations for Strategic Classification. *Proceedings of Machine Learning Research* 139 (2021), 4687–4697.

[21] Samuel Karlin and Howard M. A. Taylor. 1975. *A First Course in Stochastic Processes* (2 ed.). Academic Press, London.

[22] Michael Kearns and Aaron Roth. 2020. Games People Play (With Algorithms). In *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (first ed.). Oxford University Press, New York, NY, 94–136.

[23] Taeyoung Kim and Il Im. 2025. Understanding users' AI manipulation intention: An empirical investigation of the antecedents in the context of AI recommendation algorithms. *Information & Management* 62, 1 (2025), 104061.

[24] 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.

[25] Carl D. Meyer. 2000. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[26] Smitha Milli, Anca D. Dragan, John Miller, and Moritz Hardt. 2019. The social cost of strategic classification. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (2019), 230–239.

[27] Martin A. Nowak. 2006. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, Cambridge, MA.

[28] Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, et al. 2025. Human-AI coevolution. *Artificial Intelligence* 339 (2025), 104244.

[29] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 119)*. 7599–7609.

[30] Manish Raghavan. 2023. *The Societal Impacts of Algorithmic Decision-Making*. Association for Computing Machinery. 364 pages.

[31] Eden Saig and Nir Rosenfeld. 2025. Evolutionary Prediction Games. (2025). arXiv:2503.03401

[32] Hen Shao, Shuo Xie, and Kunhe Yang. 2025. Should Decision-Makers Reveal Classifiers in Online Strategic Classification? (2025). arXiv:2506.01936v1

[33] Christine Taylor, Drew Fudenberg, Akira Sasaki, and Martin A. Nowak. 2004. Evolutionary game dynamics in finite populations. *Bulletin of Mathematical Biology* 66, 6 (2004), 1621–1644.

[34] Arne Traulsen, Martin A. Nowak, and Jorge M. Pacheco. 2006. Stochastic dynamics of invasion and fixation. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 74, 1 (2006), 1–5.

[35] Arne Traulsen, Jorge M. Pacheco, and Martin A. Nowak. 2007. Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of Theoretical Biology* 246, 3 (2007), 522–529.

[36] Milena Tsvetkova, Taha Yasseri, Niccolo Pescetelli, and Tobias Werner. 2024. A New Sociology of Humans and Machines. *Nature Human Behaviour* 8 (2024), 1864–1876.

[37] Tijana Zrnic, Eric Mazumdar, S. Shankar Sastry, and Michael I. Jordan. 2021. Who Leads and Who Follows in Strategic Classification? *Advances in Neural Information Processing Systems* 19, NeurIPS (2021), 15257–15269.