Transparency, Detection and Imitation in Strategic Classification Supplementary Material

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

A Supplementary Material

Here we extend the analysis of the impacts of (transparent) feedback shared by the Institution on different metrics already introduced in the main text. In particular, we consdier:

- The Variation in the number of False Positives, namely ΔFP, with a generic False Positive FP being *"an Individual who is wrongly classified as likely to be successfull"* (e.g., unlikely to repay the loan despite being classified as "good creditor").
- The Variation in the number of True Positives, namely ΔTP. Here, a True Positive TP indicates *"an Individual that is correctly classified as likely to be successfull"* (e.g., likely to repay the loan and rightfully classified as "good creditor").
- The Variation in the number of True Positives over the Total Variation (False Positives and True Positives), given by the ratio $\Delta TP/(\Delta FP + \Delta TP)$. This is intended to measure the fraction of individuals that honestly improve their condition, relative to everyone that adapts (also present in Fig. 4C, in the main text).

In Fig.S1 we explore both the difference between the number of False Positives FP after and before the strategic adaptation, i.e. ΔFP, and the variation in the number of True Positives TP, i.e. ΔTP. We can observe that increasing ϕ not only reduces the number of False Positives FP but also increases the number of True Positives TP under a scenario of transparent feedback (e.g. low σ). This happens as individuals are aware of their context, their true distance from the decision threshold, and do not have incentives to faking (e.g. increase their observable feature) due to efficient detection. As a result, individuals improve their context and there is a higher number of individuals correctly labeled as Positive. For reference, in Fig. S1A we reproduce Fig. 4A discussed in our main text.

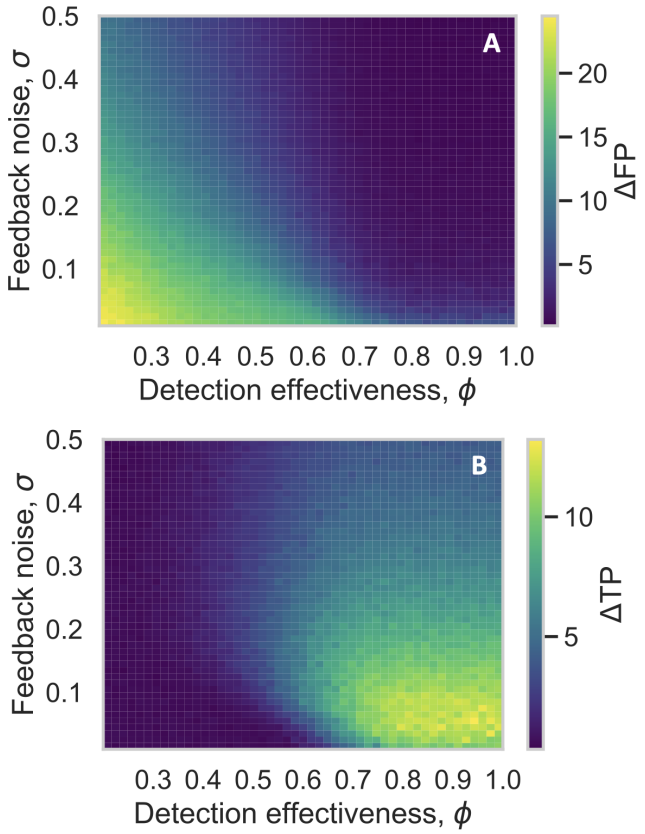

Figure S1: The role of detection effectiveness in mitigating the risks of transparent feedback. Panel **A** (top) reports the difference between the number of False Positives FP after and before the strategic adaptation, i.e. ΔFP. Panel **B** (bottom) reports the variation in the number of True Positives TP, i.e. ΔTP. Parameters used: $N = 100, b = 1.0, c_i = 3.0, \epsilon = 0, k = 0.5, \alpha = 0$. Results in this figure refer to the average over 200 runs starting from random initial conditions regarding individuals' placement in the feature space and noise in estimating the decision threshold by the Institution.

As highlighted in the core of our paper, imitation has a net positive effect in reducing the number of False Positives after strategic adaptation. In Fig. S2 we can observe that this holds for different values of detection effectiveness obtained by varying parameter ϕ: $\phi = 0.5$ (Panel **A**), $\phi = 0.7$ (Panel **B**), $\phi = 0.9$ (Panel **C**). We can also observe that Imitation contributes to effective improvement: as the bottom panels show (Panel **D**, **E** and **F**), under high I and high α, the variation in True Positives (ΔTP) is higher than the variation in False Positives (ΔFP), as conveyed by

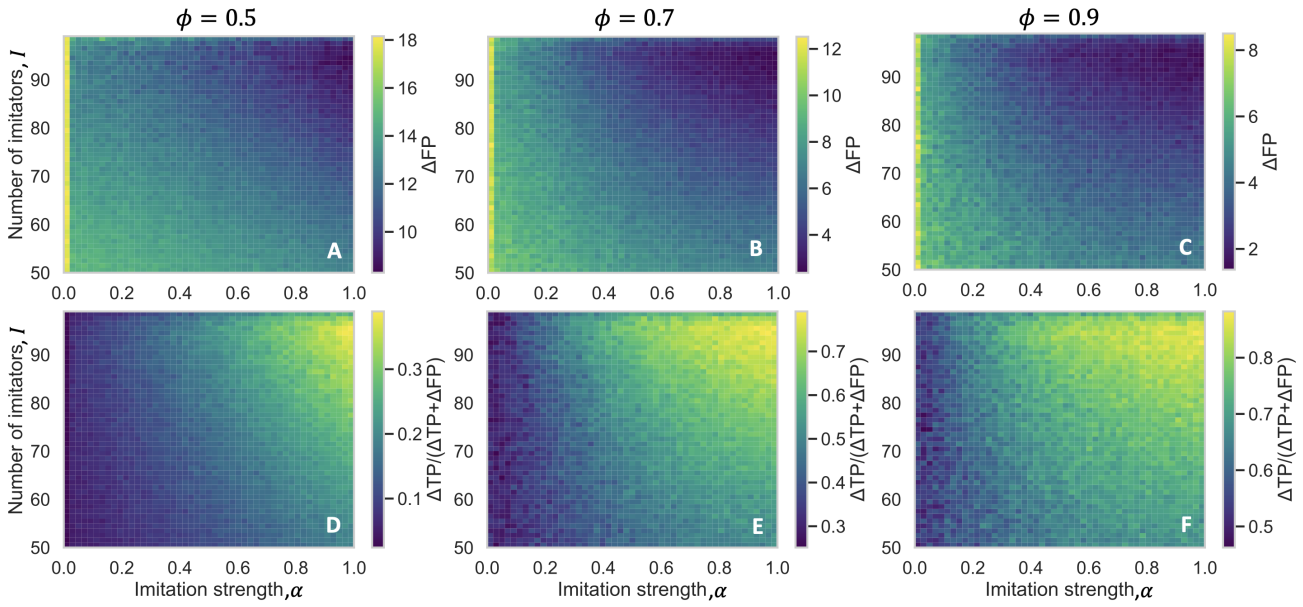

Figure S2: The role of imitation to mitigate the risks of transparent feedback. Increasing the number of imitators (I, vertical axis) and imitation strength (α, horizontal axis) reduces the number of False positives after strategic adaptation, i.e., ΔFP (top panels). This is observed for different values of the detection probability: $\phi = 0.5$ (Panel **A**), $\phi = 0.7$ (Panel **B**) and $\phi = 0.9$ (Panel **C**). Imitation also contributes to increase the relative variation in True Positives after strategic adaptation, compared with the variation in False positives, i.e., $\Delta TP/(\Delta TP + \Delta FP)$. This is observed for different values of the detection probability: $\phi = 0.5$ (Panel **D**), $\phi = 0.7$ (Panel **E**) and $\phi = 0.9$ (Panel **F**). Overall, increasing ϕ has a negative impact on ΔFP and positive impact on ΔTP. Results here are the average over 300 runs, each starting from random initial conditions.

$\Delta TP/(\Delta TP + \Delta FP) > 0.5$.

A.1 More complex scenarios

The results presented in this study are derived from a binary classification algorithm. We made this choice both for a substantive reason and a methodological motivation. The substantive reason is that binary classification is the most important discrimination task for financial organizations, especially in the credit-decisions domain. In fact, conventional discrimination methods (e.g., logistic regression or support vector machines) perform quite well for credit decision models due to linear separability properties that most frequently apply to data clusters in the model space. The methodological motivation for using a binary classification model emanates from the idea of parsimony: it is preferable to build a formal model in the simplest possible form in order to make the dynamics that it would generate as tractable as possible. Binary classification fulfils this condition of simplicity quite well, while as just mentioned also having the advantage of being the most common approach used by financial organizations.

Overall, we think that by using a simple approach in this contribution we render the main findings easily accessible while retaining their validity. It is also important to mention that our formal framework allows us to increase the level of complexity that the classification algorithm could tackle (e.g., by increasing ϵ in Eq. (1) of the main text, leading to a linearly non-separable problem). In fact, we were able to show that the effects of feedback noise, detection and imitation that we report for $\epsilon = 0$ can be extended to more complex scenarios (such as $\epsilon = 0.1$, a setting where there is a probability that individuals with high x_1 are not able to repay a loan). Here, in Fig. S3 we can observe that the same results discussed in the main text hold when considering $\epsilon = 0.1$.

A.2 Varying adaptation costs

In the main text we assume that the adaptation costs are related with each other following:

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

This allows us to use a unique parameter to control how appealing is for individuals to fake and, as a result, how challenging is for the Institution to prevent gaming. By controlling k we can interpolate between extreme scenarios in terms of challenge degree to the Institution: i) Individuals are unlikely to improve (i.e. $k = 0$, assuming faking is cheap and detection cost is low) and ii) Individuals are likely to improve (i.e. $k = 1$, assuming faking is as expensive as improvement and detection costs are high). In Fig. S4 we fix $c_f = 0.5$ and study c_ι and c_d independently. We can observe that the number of ΔFP increases with c_ι and decreases with c_d: naturally, it becomes more tempting to game if the cost of improvement is higher and if the costs of detection are lower.

The discontinuity observed in Fig.S4 represents the transition from a setting where individuals' utility increases

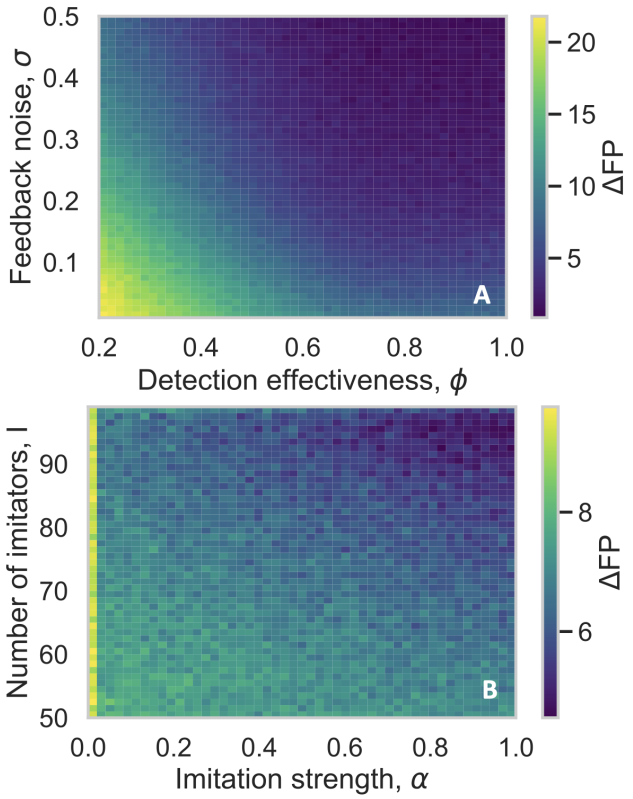

Figure S3: Same scenario explored in Fig. 4A and Fig. 4B of the main document, yet here with $\epsilon = 0.1$. The same conclusions discussed in the main text are obtained here. Parameters used: $N = 100$, $b = 1.0$, $c_\iota = 3.0$, $\epsilon = 0.1$, $k = 0.5$, $\alpha = 0$ (Panel **A**) and $\sigma = 0.001$, $\phi = 0.7$ (Panels **B** and **C**). Average over 400 runs.

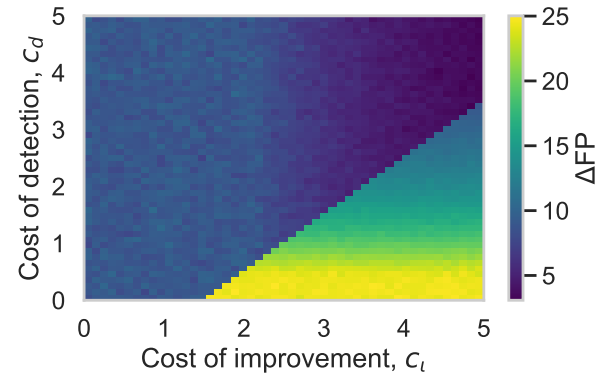

Figure S4: Exploration of the effect of varying independently the cost of improvement (c_ι) and the cost incurred by an individual that fakes and is detected (c_d). Parameters used: $c_f = 0.5$, $N = 100$, $b = 1.0$, $\epsilon = 0.0$, $\alpha = 0$, $\sigma = 0.01$, $\phi = 1.0$. Average over 400 runs.

monotonously with their real feature, $x_1(t+1)$, to a setting where it decreases with $x_1(t+1)$. Recall that expected utility is given by

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

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

$$\Delta_1(i,t+1) := (x_1(i,t+1) - x_1(i,t)), \quad (3)$$

and

$$f(i) := x_2(i,t) - x_1(i,t) \quad (4)$$

representing how much individuals fake at time t.

For now we assume that detection varies linearly with $f(i,t+1)$ i.e., $\phi = 1.0$. We also simplify notation by dropping individuals identities i, focusing on the feature space region where individuals are classified as positive, i.e., $\hat{\Theta}_i(S_i(x_2(i,t+1)), \hat{\theta}_i) = 1$. Let us further simplify notation by using $u(t+1) \equiv u$, $x(t+1) \equiv x$ and $x(t) \equiv x'$ Utility, as a function of improvement, thus reads as

$$u(x_1) = (1 - x_2 + x_1)b - (x_1 - x_1')c_\iota - (x_2 - x_1)c_f - (x_2 - x_1)c_d. \quad (5)$$

We can now study how expected utility at time $t+1$, $u(x_1)$, grows with increasing x_1, that is, real improvement:

$$\frac{du}{dx_1} = b - c_\iota + c_f + c_d, \quad (6)$$

As a result, for linear detection we have that $\frac{du}{dx_1} > 0$ if $c_d > c_\iota - b - c_f$. This means that, if individuals benefit from adapting to be classified as positive (i.e., $\hat{\Theta}_i(S_i(x_2(i,t+1)), \hat{\theta}_i) = 1$ — which naturally depends on the distance $x_1(t+1) - x_1(t)$), then utility grows with improvement and individuals have an incentive to improve rather than providing fake information. This condition captures the discontinuity observed in Fig. S4, where $b = 1$ and $c_f = 0.5$.

B Simulation Pseudocode

Here we present a summary of the code used in our simulations. First, Algorithm 1 presents the main simulation cycle, where, at each run, individuals start from a random (truthful, $x_2 = x_1$) state and, after one step of classification, can adapt their features:

Algorithm 1 Main simulation cycle

for $i \in N$ (Inititialize all agents i in set N) **do**
  $x_1(i,0) \leftarrow x \sim U(0,1)$
  $x_2(i,0) \leftarrow x_1(i,0)$
  $y_i(0) \leftarrow \begin{cases} 1 & \text{with probability } \rho_i(x_1(i,0)) \text{ Eq. (1)} \\ 0, & \text{otherwise} \end{cases}$
end for
Institution trains classifier (clf) using $x_2(i,0)$, $y_i(0)$
Classify all individuals using clf and compute metrics $FP(0), TP(0), FN(0), TN(0)$
Assuming $\vec{x}(i,t) = [x_1(i,t), x_2(i,t)]$:
for $i \in N \setminus I$ (First-movers adapt) **do**
  $\vec{x}(i,1) \leftarrow \vec{x}(i,0)$ + ADAPT(i, $\vec{x}(i,0)$)
end for
for $i \in I$ (Imitators adapt) **do**
  $\vec{x}(i,1) \leftarrow \vec{x}(i,0)$ + IMITATE(i, $\vec{x}(i,0)$,α)
end for
Classify all individuals using clf and $x_2(i,1)$.
Compute true label based on $x_1(i,1)$ and Eq. (1) and compute $FP(1), TP(1), FN(1), TN(1)$
Compute $\Delta FP = FP(1) - FP(0)$ and $\Delta TP = TP(1) - TP(0)$

Adaptation is summarized in Algorithm 2 and Algorithm 2, where we represent, respectively, adaptation through utility maximization and adaptation biased through imitation.

Algorithm 2 Function used by individuals to update their own features after classification based on utility maximisation

function ADAPT(i, $\vec{x}(i,0)$)
  Input: individual id (i), and feature vector ($\vec{x}(i,0) = [x_1(i,0), x_2(i,0)]$)
  Output: Adaptation displacement vector of individual i, $\vec{x}(i,1) - \vec{x}(i,0)$
  ▷ Compute perceived threshold:
  $\hat{\theta}_i = \max(x_2(i,0)), N \sim (\theta, \sigma))$
  ▷ Chose $x_1(i,1), x_2(i,1)$ that maximize Eq. (9):
  $x_1(i,1), x_2(i,1) \leftarrow \arg\max_{x_1(i,1), x_2(i,1)} u(i, t+1)$
return $(x_1(i,1) - x_1(i,0), x_2(i,1) - x_2(i,0))$
end function

Algorithm 3 Function used by individuals to update their own features after classification based on utility maximisation and imitation

function IMITATE(i, $\vec{x}(i,0)$, α)
  Input: individual id (i), and feature vector ($\vec{x}(i,0) = [x_1(i,0), x_2(i,0)]$)
  Output: Adaptation displacement vector of individual i, $\vec{x}(i,1) - \vec{x}(i,0)$
  ▷ Chose $x_1(i,1), x_2(i,1)$ that maximize Eq. (9):
  $\vec{u}^*_m(i) \leftarrow$ ADAPT($i, \vec{x}(i,0)$)
  $\underline{\vec{u}}_P = [0,0]$
  for $j \in P$ (Imitate average displacement of other individuals in P) **do**
    $\underline{\vec{u}}_P = \underline{\vec{u}}_P +$ ADAPT$(j, \vec{x}(j,0))/|P|$
  end for
  return $(1-\alpha) \cdot \vec{u}^*_m(i) + \alpha \cdot \underline{\vec{u}}_P$
end function