# The Gatekeeper's Effect

Moran Koren.*

Department of Mathematics, Tel Aviv University

February 10, 2022

**Abstract**

Many selection processes contain a "gatekeeper". The gatekeeper's goal is to examine an applicant's suitability for a position before both parties incur substantial costs. Intuitively, a gatekeeper should reduce selection costs by sifting unlikely applicants. However, as we show, this is not always the case since the gatekeeper's introduction inadvertently interferes with the candidate's self-selection. We study the conditions under which a gatekeeper improves the system's efficiency and those under which it induces inefficiency. Additionally, we show that selection correctness can, at times, be improved by allowing for strategic gatekeeping.

**JEL: M55,M551,D82,D83,C72,C44**

*This research was supported by the Center of Mathematical Sciences and Applications of Harvard University and the Ministry of Science and Technology of Israel (Yitzhak Shamir Fellowship). A special thanks to Itai Ashlagi, Faidra Monachou, Scott Duke Kominers, Dudu Lagziel, Itai Arieli, and Rann Smorodinsky for early discussions on this topic. I would also like to thank the participants of the Sixth Marketplace Innovation Workshop (MIW2021) and the participants of Harvard University's CMSA's Big Data 2021 Conference for raising valuable questions. Most of the work was completed during my term as a Postdoctoral Research Fellow at Harvard University's Lab for Economic Design, me@mkoren.org

Consider the process of academic publication. It is costly to all involved parties. When deciding whether to publish an article, the journal's editor invests resources in locating suitable reviewers, which perform the task of assessing the proposed publication's quality. The author incurs the submission costs and the alternative waiting costs as she can not simultaneously submit her work to a different venue. The fit of the proposed article to a specific venue depends on immeasurable characteristics such as writing quality, scientific contribution, etc. Therefore, both author and editor can only estimate the probability in which the article will pass the reviewing process and the probability it will become significant post-publication. To reduce selection costs, the editor usually performs a brief review of the paper and chooses whether to initiate a desk rejection or continue with the rigorous peer-review process. The availability of a desk rejection has two contradicting effects on the efficiency of the selection process. On the one hand, desk rejection protects both the author and the editor against unnecessary costs. On the other hand, it may induce a less careful behavior by the authors, as now, their costs will materialize only in the event that the editor finds the paper worthy, hence, have better odds of passing the review process. In general, selection processes are often costly. This cost is often imposed both on those who select and on those being selected. In this work we examine the effect of an intermediate filtering on the efficiency of noisy selection processes.

As a running example, consider the case of a hiring process. When hiring a new employee, the firm invests time and effort in assessing an applicant's fit to a proposed position. The applicant also invests time and effort in attending the interviews and preparing for professional tests. An applicant's fit to a position depends on immeasurable characteristics, such as demeanor, diligence, and intelligence. This fit can, however, be estimated from the candidate observable traits (e.g., education, previous experience, etc.). To reduce the aforementioned costs, mainly for the firm, many selection processes contain a pre-filtering stage called *a gatekeeper.* The gatekeeper predicts the applicant's success probability based on her observable traits and attempts to sift the wheat from the chaff, allowing the test only to those applicants who are most likely to pass it. The role of the gatekeeper is often filled by an HR representative or an AI algorithm that skims through applicants' CVs before the professional staff reviews them.

Intuitively, the introduction of a gatekeeper will reduce the firm's selection costs as it prevents the candidates that are less likely to be accepted from taking the test. However, as we show, the gatekeeper incurs an indirect consequence. By pre-rejecting less likely candidates, the gatekeeper interferes with the candidate's self-selection. To see this indirect effect, consider a candidate who contemplates applying. Before applying to the position and incurring her costs (e.g., preparing for the test, travel arrangements etc.), each candidate evaluates her odds of eventually being selected. Naturally, she will choose to opt-out when, based on

her own assessment, the odds of her being selected are too low. When the selection process features a gatekeeper, the candidate knows that she will endure the costs only if the gatekeeper's information is in her favor. When this occurs, the marginal candidate would now prefer opting-in, thus lowering the average application quality. We ask whether this indirect effect has the potential to override the intended gatekeeper effect. In other words, can the correctness of a selection process decrease due to the introduction of a gatekeeper? and if so, under what conditions?

We present a game of incomplete information played between a candidate and a gatekeeper. The game comprises two states of nature. In one, the candidate and the position are highly compatible. In the other, the quality of the fit is low. The aforementioned state of nature is unknown, yet both agents receive a noisy signal over the identity of the realized state. The candidate decides whether to apply to the position. The gatekeeper decides whether to submit the candidate (provided she chose to apply) to a costly test. First, we assume that the gatekeeper is technocratic and naively follows its signal. Later, we allow for strategic behavior from both players. In this scenario, we find that: (1) the introduction of a gatekeeper interferes with the candidate self-selection process and lowers the average application quality; (2) we characterize the conditions under which the introduction of a gatekeeper decreases the overall correctness of the selection process; (3) we introduce a game between an applicant and a strategic gatekeeper and show that, in some cases, the gatekeeper can rectify the candidate self-selection by playing a mixed strategy, thus restoring the system's efficacy.

**Organization:** In Section 1 we describe the related work. Our model, analysis, and main results are presented in Section 2. We tackle strategic gatekeeping in Section 3. We demonstrate the implications of our results using a numeric example in Section 4. Section 5 contains a discussion on the applicability of our results to algorithmic bias and peer-review methodologies. Finally, we conclude in Section 6.

# 1 Related work

The impact of rational agents' strategic behaviors in relation to economic efficiency in the presence of uncertainty has been studied extensively over the last half-century. In his seminal paper, Akerlof (1970) presented a market model where the value of a proposed good (a used car) is known only by one side of the market, while the other side is oblivious regarding the value. He showed that in such markets, the average quality of the proposed commodity decreases due to adverse selection (e.g., "lemons" are driving out good cars from the market). In follow-up work, Levin (2001) showed that adverse selection depends on the relative quality

of the information available to both market sides. That is, while holding the information structure of the buyer fixed, a more informed seller increases the severity of the so-called buyer's curse and vice versa. Our model diverges from this line of thinking in three ways. First, it contains no asymmetric information. In our model, both sides are oblivious to the underlying state, and the quality of player signals may vary. Second, much of the literature on adverse selection assumes an inherent conflict of interest. In our model, if agents hold perfect information, both players' incentives are fully aligned. The tension between the players is due to the uncertainty both face. Finally, our results are not driven by an exogenously determined price, although the applicant cost parameter plays a similar role.

Additionally, our results relate to a growing line of research on screening. Lagziel and Lehrer (2019) studied the problem of screening efficiency. That is, whether a stricter filtering step improves the quality of the resulting decision. The authors found a non-monotone connection between the two. For example, there are cases in which a higher screening threshold induces a decrease in the overall expected valuation. Additionally, they characterized the conditions on the distribution under which such non-monotonically emerges. In a follow-up paper, Lagziel and Lehrer (2021) have shown that a noisily informed decision-maker can improve the quality of his decision by the introduction of an additional binary noise. Our results are similar as we show that, in some cases, one can improve correctness even by adding a gatekeeper whose signal is extremely noisy and economically equivalent to a coin toss. In the other cases, the gatekeeper's signal quality determines the added value of an additional filtering step. However, we differ from the existing screening literature as the driving force of our model is agents' strategic choice while, in their model, the primary focus is on characterizing the conditions of the probability structure under which additional "naive" screening is beneficial.

Our study of strategic gatekeeping is also reminiscent of the work on Bayesian Persuasion originated by Kamenica and Gentzkow (2011). There, a fully informed sender gains additional rewards by employing a strategy by which he lies to the receiver at some positive probability (See Kamenica, 2019 for a recent review). In our case, the influence of the gatekeeper's dishonesty is to incentivize a more selective candidate's behavior. We show that the benefits of this strategy are guaranteed when the public information is sufficiently favorable.

Finally, we borrow our information structure from the literature on information cascades and herding (Banerjee, 1992; Bikhchandani et al., 1992; Smith and Sørensen, 2000; Arieli and Mueller-Frank, 2019). In this group of models, sequentially arriving agents learn from (or herd on) the actions of agents who arrived before them. In most herding models, the strategic interaction between agents is

limited as agents are assumed to exist for a single period and are unaffected by future events (see Bikhchandani et al., 2021 for a recent survey).[1] In our model, the strategic interaction between both agents is the focal point of our analysis.

# 2 The Model

The game comprises a candidate ($C$) and a gatekeeper ($K$). There are two possible states of nature that we denote by $\Omega = \{H, L\}$. In state $H$, the candidate is highly compatible with the position. In state $L$, there is a low level of compatibility between the candidate and the position.

In order to fill the position, the candidate must pass a costly test. In state $\omega \in \Omega = \{H, L\}$, the probability of a candidate passing the test is denoted by $\varphi(\omega)$. We assume that $\varphi(H) > \varphi(L) > 0$. We normalize $\varphi(H) = 1$ and denote $\varphi(L) \equiv \varphi$.[2] If she chooses to apply, and is allowed to take the test, the candidate incurs a cost $\gamma \in (\varphi, 1)$. This cost materializes in the event that the test takes place, even if the candidate does not pass it.

The candidate can decide whether she wishes to apply or not, $a_C \in \{0, 1\}$. If she chooses not to apply, she receives a utility of zero. If she applies ($a_C = 1$), and passes the test, she gains positive utility (even when $\omega = L$). We normalize the utility of a high fitting candidate to 1 and assume that a candidate with a low fit gains a utility of $\alpha > 0$, such that $\alpha\varphi < \gamma$. The gatekeeper can decide whether to submit the applicant to the test $a_K = \{0, 1\}$. It receives a utility of 1 whenever a highly fitting candidate passes the test and endures a utility of $-d$ whenever a low fitting candidate passes the test, where $d > 0$.[3]

The realized state is unknown. Both players assign a prior probability of $\mu$ to the event $Pr(\omega = H)$. We will call $\mu$ the *public belief*. In addition, each player $i \in \{C, K\}$ receives a noisy signal $s_i \in S_i$. We assume that the gatekeeper's signal is binary, i.e., $S_K = \{h, l\}$.[4] The candidate draws her signal from an open interval over $S_C \subset \mathbb{R}$. The quality of the gatekeeper's signal $Pr(s_k = \omega|\omega) = q_K > \frac{1}{2}$ is known (hereafter, we suppress the subscript $K$). The candidate's signal is randomly drawn from a state-dependent distribution $F_\omega$ with PDFs $f_\omega$. The distributions $F_L, F_H$ are mutually absolutely continuous, i.e., no signal fully

[1] Recently, papers that either highlight the strategic game between the agents or consider the overall effect have begun to emerge. See for example Ban and Koren, 2020b; Arieli et al., 2019; Halac et al., 2020; Smith et al., 2021.

[2] Note that this assumption is without loss of generality. Our only limitation is that the above inequalities be strict.

[3] Note that the assumption on gatekeeper's utility is without loss as any cost structure can be fitted by choice of an appropriate $d$.

[4] Most of our results carry through even when we assume a richer information set. We discuss this further in Section 6.

reveals the unknown state. Additionally, we make two simplifying assumptions. First we assume that $F_L, F_H$ exhibit the Monotone Likelihood Ratio Property (MLRP). MLRP is defined as follows,

**Definition 1.** $F_L, F_H$ exhibit the *Monotone Likelihood Ratio Property (MLRP)* if for every $s, \hat{s} \in S_C$ such that $s > \hat{s}$ the following condition is satisfied,

$$\frac{f_H(s)}{f_L(s)} \geq \frac{f_H(\hat{s})}{f_L(\hat{s})}.$$

Second, we assume the candidate signals are unbounded. The following definition of unbounded signals is borrowed from Smith and Sørensen (2000),

**Definition 2.** Let $F_H, F_L$ be two mutually absolutely continuous distributions with a common support $supp(F)$. We say that signals are *unbounded* if the convex hull $co(supp) = [0, 1]$.

The implications of assuming an unbounded signal structure are that for every $\mu \in (0, 1)$ and every $q \in [0.5, 1)$, the candidate will always take either action with a positive probability. In social learning models, this creates the *overturning principle* that facilitates learning in infinite populations. In our context, it is a means to an end for maintaining clarity. As once combined with MLRP, it guarantees that the candidate follows a unique threshold strategy where both actions occur with positive probability. We believe that releasing this assumption will add little insight and will not alter the qualitative nature of our result.[5]

Let $\sigma$ denote a strategy profile for both players. Note that $\sigma$, together with the game's information structure define a probability distribution $\mathbb{P}$ over $\Omega \times S_C \times S_K$. Therefore, given a strategy profile $\sigma$, the players expected utilities are well defined and can be calculated.

## 2.1 Analysis and Results

Our first question is: What effect does embedding a "naive" gatekeeper in a selection process have on its efficiency? In other words, can we always improve the quality of the selection process by augmenting it with a gatekeeper who automatically follows its signal and is deprived of any strategic considerations? Intuitively, one would expect that the answer will depend on the gatekeeper's signal quality. If the gatekeeper's signal is of sufficiently high quality, the resulting allocation will improve, but it may be reduced by including a low-quality gatekeeper. As we soon find out, this intuition, while not without merit, is not always correct.

A *Naive* strategy for the gatekeeper is defined by $\sigma_K(h) = 1, \sigma_K(l) = 0$. A *Naive Gatekeeper* is a gatekeeper whose strategy space is restricted to playing only the naive strategy. We will use subscript $NK$ to denote a "Naive Keeper."

[5] We discuss this modeling choice further in Section 6.

The candidate uses Bayes rule to update her posterior belief. Upon receiving a signal $s \in S_C$, her expected utility from applying will be,

$$U_C(a=1|s) = \\ \frac{\mu p(s)}{\mu p(s) + (1-\mu)(1-p(s))}\sigma_K^H(1-\gamma) + \frac{(1-\mu)(1-p(s))}{\mu p(s) + (1-\mu)(\alpha - p(s))}\sigma_K^L(\alpha\varphi - \gamma), \tag{1}$$

where $\sigma_K^\omega = Pr_\sigma(a_K = 1|\omega)$ is the probability of passing the gatekeeper when she plays strategy $\sigma$ and the realized state is $\omega$, and $p(s) = \frac{f_H(s)}{f_H(s)+f_L(s)}$. Note that we can map each signal to its respective $p(s)$. Thus, without loss of generality, we assume that the signal $s$ admits the posterior $p(s)$ and all signals who share posterior are grouped. We will call $p(s)$ the *subjective quality* of the candidate as it measures the expected level of fit based on her private information when all other factors are held constant. Our assumptions on the signal structure translate to $\{x : s \in S_C \text{ and } x = p(s)\} = [0,1]$. Let $x(\sigma_K) = \inf\{x : x = p(s) \text{ and } s \in S_C \text{ and } u_C(a_C = 1|s, \sigma_K) > 0\}$. That is, when the gatekeeper's strategy is $\sigma_K$, $x(\sigma_K)$ is the lowest signal for which the candidate's optimal action is to apply. As we assume strict MLRP, we know that for every $\sigma_K$, $x(\sigma_K)$ is uniquely defined. Furthermore, one can map every signal $s$ to the posterior it induces $p(s)$. Therefore, we interchange between these terms in our analysis. We will use the term *applicant* to describe a candidate who chooses to apply. Under our assumptions, applicants are candidates whose subjective quality is greater than $x(\sigma_K)$. We now present our first result, which refers to the naive gatekeeper.

**Theorem 1.** *When facing a naive gatekeeper, the average subjective quality of the applicant decreases with the gatekeeper's signal quality.*

*Proof.* Let $F_L, F_H, S$ be an information structure and let $q, \hat{q}$ be two signal qualities such that $q > \hat{q}$. Given a naive gatekeeper of quality $q$, and by equation (1), a candidate will apply whenever the following condition is satisfied,

$$\frac{p(s)}{1-p(s)} \geq \frac{\gamma - \alpha\varphi}{1-\gamma}\frac{1-\mu}{\mu}\frac{1-q}{q}. \tag{2}$$

The threshold $x(\sigma_{NK}(q))$ is the subjective quality for which equation (2) holds with equality, that is, the signal that yields the lowest posterior belief of a high fit, for which the candidate still chooses to apply. Note that under our assumptions of MLRP and the possible set of subjective qualities, this threshold is uniquely defined for every $q$. To complete the proof of the theorem, note that the function $\frac{y}{1-y}$ is decreasing. □

In Theorem 1 we prove that the gatekeeper has an indirect effect on the candidate's choice. In addition to sifting out less likely applicants, it also interferes with the candidate's self-selection. Candidates whose subjective quality was marginal,

and would otherwise be deterred by the testing costs, will now apply, thinking that the gatekeeper's decision will protect them against unnecessary costs (as the event in which these costs materialize and the candidate fails becomes less likely). The question remains, though, can this indirect effect dominate?

To discuss "efficiency," we first define it. Our notion of efficiency will be the probability of reaching the correct decision. We augment the term correctness of Arieli et al. (2018). The correctness of the game, denoted by $\theta$, is defined as,

$$\theta = Pr(H)Pr(\psi|H) + Pr(L)(1 - Pr(\psi|L)), \tag{3}$$

where $\psi$ is the event in which the candidate passes the test (i.e., the candidate applies, is approved by the gatekeeper, and passes the test). In what follows, one can think of a system designer who wishes to maximize the selection process's correctness.[6]

In our second result, we discuss the influence of introducing a gatekeeper on the correctness of the selection process. In Theorem 2 we compare the correctness of a selection process where no gatekeeper is present to the case of a naive gatekeeper with signal quality $q$. Let $\hat{x}$ denote the candidate's threshold type when no gatekeeper is present.

**Theorem 2.** *For any information structure $F_H, F_L, S, \mu$,*

1. *there exist $\bar{q}$ such that for all $q > \bar{q}$, introducing a naive gatekeeper of quality $q$ improves correctness.*
2. *if $\frac{\mu(1-F_H(\hat{x}))}{(1-\mu)(1-F_L(\hat{x}))} > \varphi$, there exists $\underline{q}$ such that whenever $q < \underline{q}$ the introduction of a naive gatekeeper lowers the correctness of the selection process.*
3. *Let $\bar{q} > 0.5$. if $\frac{\mu(1-F_H(\hat{x}))}{(1-\mu)(1-F_L(\hat{x}))} < \varphi$, the introduction of a naive gatekeeper of quality $q \in (0.5, \bar{q})$ improves the correctness of the selection process.*

The first and second parts of Theorem 2 state that a high-quality gatekeeper is always beneficial, while a low-quality gatekeeper can be detrimental. The third part is somewhat surprising, stating that correctness can improve even with the introduction of a gatekeeper of arbitrarily low quality. For example, take an arbitrarily small $\varepsilon$. A gatekeeper with signal quality $\frac{1}{2}+\varepsilon$ is economically equivalent to a coin toss. Despite this equivalence, part three tells us that some selection processes benefit from the gatekeeper's introduction.

The intuition behind this result is as follows. Consider an alternative model in which the candidate sees the gatekeeper's signal before making her decision.

[6]While we assume that the designer's utility is affected by false positives and false negatives in the same magnitude, we believe that releasing this assumption will have little effect over the qualitative nature of our results. We release this assumption when we discuss the strategic gatekeeper.

Thus, she can utilize it in her choice. If the candidate's signal is low, her decision to opt out will not be affected by the additional information. If the candidate's private signal is intermediate, her best response is to follow the gatekeeper signal. Thus, the correctness will equal that of the original model. However, when the candidate's signal is sufficiently strong, the candidate's optimal choice is to disregard the new information and apply even if the gatekeeper's signal is negative. For this this group of signals, the correctness in the alternative model will be higher than the correctness when the gatekeeper is present. In fact, the introduction of a gatekeeper decreases the correctness when the candidate's signal is sufficiently strong. The higher the gatekeeper's signal quality is, the smaller the set of candidate signals that are negatively effected. An additional effect comes into play when the gatekeeper's signal is sufficiently noisy. In this case, the decrease in overall correctness due to those candidate types who receive strong signals is countered by the increase of correctness due to the increase in selectivity of candidates of intermediate types.

## Proof of Theorem 2.

First we calculate $\hat{x}$. By equation (1) we find that $\hat{x} = \frac{(\gamma-\alpha\varphi)(1-\mu)}{(\gamma-\alpha\varphi)(1-\mu)+(1-\gamma)\mu}$. By the definition of correctness, when no gatekeeper is introduced, the correctness will be

$$\begin{aligned}\hat{\theta} = \mu(1 - F_H(\hat{x})) + (1-\mu)(F_L(\hat{x}) + (1 - F_L(\hat{x}))(1-\varphi)) = \\ \mu(1 - F_H(\hat{x})) + (1-\mu)(1 - (1 - F_L(\hat{x}))\varphi).\end{aligned} \tag{4}$$

When introducing a naive gatekeeper with quality q, the process correctness is,

$$\theta(q) = \mu q(1 - F_H(x(q))) + (1-\mu)(1 - (1 - F_L(x(q)))(1-q)\varphi)$$

where

$$x(q) = \frac{(\gamma - \alpha\varphi)(1-\mu)(1-q)}{(\gamma-\alpha\varphi)(1-\mu)(1-q) + (1-\gamma)\mu q}. \tag{5}$$

We calculate the difference between both expressions and find:

$$\begin{aligned}&\theta(q) - \hat{\theta} = \\ &\mu(q(1 - F_H(x(q))) - (1 - F_H(\hat{x})) + (1-\mu)\varphi(1 - F_L(\hat{x}) - (1-q)(1 - F_L(x(q)))) \\ =&\mu(F_H(\hat{x}) - qF_H(x(q)) - (1-q)) + (1-\mu)\varphi(q + (1-q)F_L(x(q)) - F_L(\hat{x})).\end{aligned} \tag{6}$$

An introduction of a naive gatekeeper is beneficial whenever the above difference is positive. After rearranging, we get the following proposition:

**Proposition 1.** *A naive gatekeeper with quality q yields correctness that is higher than the benchmark case if and only if,*

$$\frac{\mu(1 - F_H(\hat{x})) + (1-\mu)\varphi(F_L(\hat{x}) - F_L(x(q)))}{\mu(1 - F_H(x(q))) + (1-\mu)\varphi(1 - F_L(x(q)))} \leq q.$$

To see part 1 of Theorem 2 we will show that for every $q$, the left-hand side of Proposition 1 is smaller than one. That is,

$$\mu(1-F_H(\hat{x}))+(1-\mu)\varphi(F_L(\hat{x})-F_L(x(q))) \leq \mu(1-F_H(x(q)))+(1-\mu)\varphi(1-F_L(x(q))).$$

Rearranging, we get,

$$F_H(x(q)) - F_H(\hat{x}) \leq \frac{1-\mu}{\mu}\varphi(1-F_L(\hat{x})).$$

By equation (5), for every $q \in [0.5, 1]$ we get that $\hat{x} = x(\frac{1}{2}) \geq x(q)$. Therefore, $F_H(x(q)) - F_H(\hat{x}) \leq 0$, thus concluding the proposition's proof. □

To prove the remaining parts of the theorem, we examine if

$$\frac{\mu(1-F_H(\hat{x})) + (1-\mu)\varphi(F_L(\hat{x}) - F_L(x(q)))}{\mu(1-F_H(x(q))) + (1-\mu)\varphi(1-F_L(x(q)))} < \frac{1}{2}$$

when $q \approxeq \frac{1}{2}$.

By Proposition 1, if this condition holds, than adding a gatekeeper of such low quality improves correctness. If it does not hold, adding the low-quality keeper hurts correctness.

Let $q = \frac{1}{2} + \varepsilon$. The condition in Proposition 1 can now be written as,

$$\frac{\mu(1-F_H(\hat{x})) + (1-\mu)\varphi(F_L(\hat{x}) - F_L(x(q)))}{\mu(1-F_H(x(q))) + (1-\mu)\varphi(1-F_L(x(q)))} < \frac{1}{2} + \varepsilon.$$

Recall that $\lim_{\varepsilon \to 0} F_\omega(x(q)) = F_\omega(\hat{x})$; hence, in the limit, for an arbitrarily small $\varepsilon$, correcntess improves only whenever,

$$\frac{\mu(1-F_H(\hat{x}))}{\mu(1-F_H(\hat{x})) + (1-\mu)\varphi(1-F_L(\hat{x}))} \leq \frac{1}{2}.$$

After rearranging, we get the above condition. □

# 3 Strategic Gatekeeping

In this section, we ask whether one can do better than the naive gatekeeper. In Theorem 2 we have shown that in some cases, the introduction of a gatekeeper causes a decrease in the efficiency of the selection process as it is captured by its correctness. Now we examine this question from the perspective of the gatekeeper. The gatekeeper's incentive is similar to correctness as both increase with the probability that a suitable candidate will pass the test. The gatekeeper's incentive structure differs from our notion of correctness because the gatekeeper endures negative utility only for low fitting candidates who pass the test. However, the correctness also decreased due to the opportunity cost, which discourages ex-post high fitting candidates from applying. We examine whether the gatekeeper could also benefit from the introduction of noise.

We will assume that the following condition on the gatekeeper's utility to avoid trivial cases,

$$\begin{aligned}&\mu q(1-F_H(x(\sigma_{NK}(q))))-d\varphi(1-\mu)(1-q)(1-F_L(x(\sigma_{NK}(q))))\\&>0>\\&\mu(1-q)(1-F_H(x(\sigma_{NK}(q))))-d\varphi(1-\mu)q(1-F_L(x(\sigma_{NK}(q)))).\end{aligned}\tag{7}$$

We model strategic gatekeeping by allowing the gatekeeper to mix between utilizing the naive strategy with probability $1-\sigma$, and with probability $\sigma$, allowing the applicant (i.e., the candidate who chooses to apply) to take the test without filtering.[7] In the following proposition, we show that this method indeed counters the adverse effects on the candidate's self-selection behavior from Theorem 1.

**Proposition 2.** *For every $q, \mu, F_L, F_H, S$ and a mixed strategy $\sigma$, the following inequality is satisfied,*

$$\frac{\partial x}{\partial \sigma} > 0,$$

*were $x(\sigma)$ is the type of the indifferent candidate when the probability of no gatekeeping is $\sigma$.*

*Proof.* Let $\sigma$ be the probability in which the candidate faces no gatekeeping. If she chooses to apply, her expected utility will be,

$$\frac{\mu p(s)(q+\sigma(1-q))(1-\gamma)-(1-\mu)(1-p(s))((1-q)+\sigma q)(\gamma-\varphi\alpha)}{\mu p(s)+(1-\mu)(1-p(s))}.$$

The candidate will be indifferent whenever the following equality holds,

$$x(\sigma)=\frac{(\gamma-\varphi\alpha)(1-\mu)((1-q)+\sigma q)}{(\gamma-\varphi\alpha)(1-\mu)((1-q)+\sigma q)+(1-\gamma)\mu(q+\sigma(1-q))}.$$

The derivative of $x(\sigma)$ is thus,

$$\frac{\partial x}{\partial \sigma}=\frac{(\gamma-\varphi\alpha)(1-\mu)(1-\gamma)\mu(2q-1)}{((\gamma-\varphi\alpha)(1-\mu)((1-q)+\sigma q)+(1-\gamma)\mu(q+\sigma(1-q)))^2}.\tag{8}$$

Recall that the expression above is positive for every $q \in (0.5, 1]$. □

Proposition 2 shows that strategic gatekeeping can counter the candidate's self-selection. This, however, comes at a cost as the process dismisses the gatekeeper's signal with positive probability. Next, we examine if the gatekeeper can gain additional expected utility from deploying such a strategy.

In Theorem 3 we show that there are cases in which even the gatekeeper benefits from the introduction of noise. In fact, when the prior is sufficiently high, we prove that the gatekeeper always benefits from such strategic behavior.

[7]Note that this is equivalent of giving a gatekeeper with a low signal the ability to play a mixed strategy.

**Theorem 3.** *There exists $\bar{\mu}$ such that for all $\mu \geq \bar{\mu}$, in every equilibrium $\sigma > 0$.*

*Proof.* The gatekeeper's expected utility from any strategy $\sigma$ can be written as,

$$U_K(\sigma) = \mu(1 - F_H(x(\sigma)))(q + \sigma(1-q)) - d\varphi(1-\mu)(1 - F_L(x(\sigma)))((1-q) + q\sigma).$$

In the theorem we claim that $\sigma > 0$. That is, when the public belief is sufficiently high, playing the naive gatekeeper's strategy is never optimal. Assume to the contrary that $\mu$ is arbitrarily close to one, yet $\sigma = 0$ is an equilibrium. This entails that $\frac{\partial U_K}{\partial \sigma}|_{\sigma \to 0} \leq 0$.

We calculate the derivative of $U_K(\sigma)$ as follows,[8]

$$\begin{aligned}
&\frac{\partial U_K}{\partial \sigma} = \\
&\mu(1-q)(1 - F_H(x(\sigma))) - \mu(q + \sigma(1-q))f_H(x(\sigma))\frac{\partial x}{\partial \sigma} \\
&- d\varphi(1-\mu)q(1 - F_L(x(\sigma))) + d\varphi(1 - q + \sigma q)f_L(x(\sigma))\frac{\partial x}{\partial \sigma}
\end{aligned}$$

and thus $\frac{\partial U_K}{\partial \sigma}|_{\sigma \to 0} \leq 0$ whenever,

$$\begin{aligned}
\mu(1-q)(1 - F_H(x(q))) - d\varphi(1-\mu)q(1 - F_L(x(q))) \leq \\
(\mu q f_H(x(q)) - d\varphi(1-q)(1-\mu)f_L(x(q)))\frac{\partial x}{\partial \sigma}.
\end{aligned} \tag{9}$$

To see the contradiction, recall that $\mu$ is arbitrarily close to one. Therefore, the left hand side of (9) is arbitrarily close to $1 - q$, while by equation (8), the expression $\frac{\partial x}{\partial \sigma}$ is arbitrarily close to zero. □

At first glance, Theorem 3 may seem intuitive for anyone who is well versed in Bayesian analysis. When disregarding the candidate's self-selection, whenever the prior is above the gatekeeper's signal quality, it is in the gatekeeper's best interest to disregard its signal and follow the public belief. In our model, however, any increase in the initial prior is balanced by a less selective candidate's behavior. When proving Theorem 3, we account for this behavior as well. We show that then $\mu$ is sufficiently close to one, $\sigma = 0$ is never a local maximum. Thus, cannot be a global maximum. Note that the opposite is not necessarily true. That is when $\mu$ an equilibrium in which the gatekeeper follows its signal can emerge whenever $\mu$ is sufficiently lower than one. The occurrence of such equilibrium is determined by the choice of the information structure.

[8] We slightly abuse notation here, and refer to $x(q)$ as defined by equation (5) rather than calling it $x(\sigma = 0)$.

# 4 A Numerical Example

As an example, assume that the candidate information structure is as follows,[9]

$$F_H(x) = x^2; F_L(x) = 1 - (1-x)^2; x \in (0,1).$$

Assume that $\gamma = 0.4, \varphi = 0.6, \alpha = 0.5$, and $d = 1$. In figure 1 we calculate the condition from Proposition 1 and check whether adding a gatekeeper improves or degrades the decision's correctness.

**The Gatekeeper's effect on correctness.**

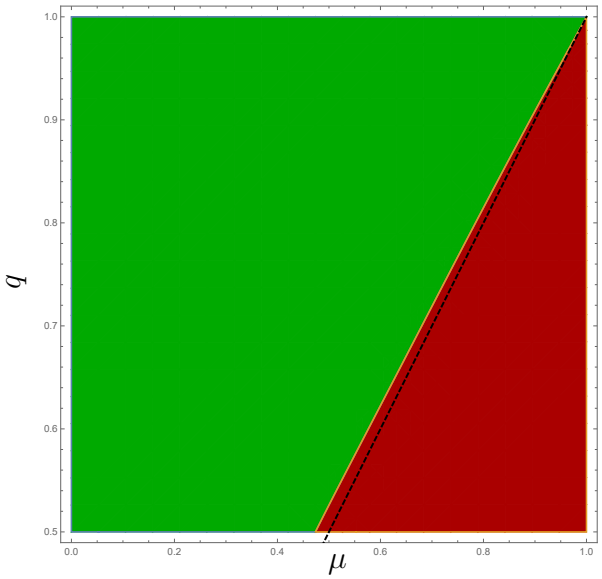


Figure 1: The public belief $\mu = Pr(\omega = H)$ is on the horizontal axis, the gatekeeper's signal quality is on the vertical axis. The dashed line is $\mu = q$. The rest of the parameters are $\gamma = 0.4, \varphi = 0.6, \alpha = 0.5$, and $d = 1$. By Proposition 1, the introduction of a gatekeeper improves correctness in the green regions and harms it in the red regions.

In figure 1 we can see the connection between the gatekeeper effect on correctness, the unconditional probability over the states of nature (i.e., the public belief), and the gatekeeper's signal quality. As expected, the gatekeeper positively affects correctness whenever her signal quality is sufficiently high for any value of public belief. One can see this by the green region at the top of Figure 1. The third part of Theorem 2 tells us that the gatekeeper's effect can, at times, be positive, even when its signal quality is extremely low. In Figure 1 we see that this occurs whenever the public belief is sufficiently low. Finally, we see that as the public belief increases, a positive gatekeeper's effect emerges only for increasingly higher signal qualities. This finding confirms our third result, which states that strategic gatekeeping may improve performance for sufficiently high public beliefs. We provide the inverse image in Figure 2, where we calculate the threshold above

[9]See Ban and Koren (2020a) for further details.

which playing the "naive gatekeeper" strategy is no longer an equilibrium. As one can see, it is increasing in the signal quality, yet is lower than $q$.

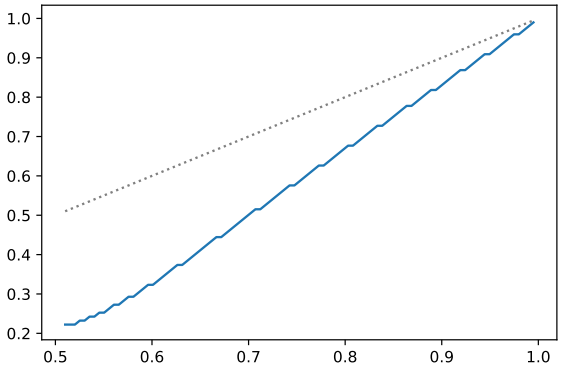


Figure 2: The public belief $\mu = Pr(\omega = H)$ is on the vertical axis. The gatekeeper's signal quality is on the horizontal axis. The rest of the parameters are $\gamma = 0.4, \varphi = 0.6, \alpha = 0.5$, and $d = 1$. The dashed line is the identity line. By Theorem 3, for every $q$, there exists a threshold $\bar{\mu}$ such that whenever the public history is above it, the Naive gatekeeper is no longer an equilibrium. The solid blue line describes this threshold.

As we can see in Figure 2, the threshold $\bar{\mu}$ is below the identity line. This is to be expected as when $\mu$ approximately equals $q$, the strength of the public belief cancels out a negative gatekeeper signal. This intuition has merit even when considering the strategic behavior of the candidate.

# 5 Discussion

Next we discuss the implications of our results in two important real world scenarios: (1) Biased Algorithms and (2) The role of reputation in the academic peer-review process.[10]

As AI based decision making tools gain popularity, there is a growing concern that their predictions, which are based on historic data, will fortify and magnify the under-representation of minority groups. This concern is due to the fact that learning is, by definition, based on existing observations. In many cases, these observations are skewed due to historic biases.

Consider the following extension to our model. Assume that, in addition to the model described in Section 2, the candidate can also be either "green" or "magenta." Due to historic reasons, the gatekeeper's signal quality when filtering green candidates is higher than its signal quality when facing magenta candidates.

[10]I would like to thank an anonymous reviewer for raising these issues.

According to Theorem 1, due to the difference in the gatekeeper's abilities, magenta candidates will be more cautious and exert a higher level of self-selection. As a result, the number of magenta applicants (i.e., candidates who applied) will be lower than their proportion in the population. Additionally, the average quality of magenta applicants will be higher than the average quality of green candidates. Note that if the algorithm is not biased, the difference in average applicant quality due to self-selection can, eventually, lead to a correction as the "magenta" color feature will be correlated with higher quality applicants, at least until the process converges and both groups are treated equally. This correction will not occur if the algorithm is designed to ignore this feature. Our findings suggest that it may be in the system's best interest to skip the gatekeeper when encountering a candidate from an under-represented group, provided her common prior is sufficiently high. This will improve both the system's correctness and its ability to better handle candidates from this group. The latter effect is out of our scope and is left for future work.[11]

A second example can be found in the peer-review process. This process differs significantly among academic disciplines. For example, in Economics, the process is single-blinded (i.e. reviewers know the identity of the authors but not vice versa). However, in Computer Sciences, the process is mostly double blinded (i.e. reviewers do not know the identity of the authors or vice versa). The role of the gatekeeper in this scenario is played by the editor (or the area chair in conferences, whenever a desk rejection is possible). In our model, the difference in methodologies is captured by the common prior. If the process is single-blinded, outstanding researchers will have high priors while novice researchers correspond to a balanced prior. If the process is double-blinded, there is no public information. Thus, the prior is balanced. Our results support the single-blind approach, as the gain from employing a gatekeeper varies with the reputation. Mainly, correctness decreases when the common prior is sufficiently high. Intuitively, in practice, one would expect an even more severe effect as there is a strong correlation between researchers with an outstanding reputation and those whose private signals are of high quality.

# 6 Conclusion

A gatekeeper is a common feature in many costly selection processes. Its goal is to reduce overall selection cost by sifting the wheat from the chaff before ad-

[11] In a recent example, Amazon have implemented an AI recruitment system. The system was scrapped as researchers found that it is biased against women (see `https://www.bbc.com/news/technology-45809919`). The bias occurred due to the scarcity of women C.Vs in the system's cohort.

ministering a costly exam. While not without merit, this intuition disregards the gatekeeper's indirect effects on strategic behavior by those who are being selected, i.e., the candidates. This work studies the implications of the gatekeeper's introduction on the resulting decision's quality, while also considering the candidate's behavior.

We introduce a game of incomplete information. A candidate must decide whether to participate in a costly selection process. A gatekeeper must decide whether she allows the candidate to take a costly exam. We find that the presence of a gatekeeper induces less self-selection on behalf of the candidate. Furthermore, the higher the quality of the gatekeeper signal, the less selective the candidate behavior becomes. We analyze the consequences of this indirect effect and find that (1) Provided that her signal quality is significantly high, the addition of a gatekeeper improves the probability of the process resulting in a correct decision. (2) Surprisingly, there are cases in which even the addition of a very low-quality gatekeeper improves correctness. This phenomenon occurs whenever the test is sufficiently refining.

We make two assumptions over the information structure. First, we describe the gatekeeper's private information as a binary signal structure. Second, we assume that the candidate signals are unbounded. We argue that these assumptions can be relaxed. As for the binary gatekeeper's signal, in a previous version (available upon request), we have assumed a richer gatekeeper's signal structure. Note that Theorem 1 depends on the ratio $\frac{Pr(H|a_K=1)}{Pr(L|a_K=1)}$. When the gatekeeper's signals are continuous, our analysis carries through. However, in this case, it will rely on comparing the gatekeeper's indifference signal generated by two separate information structures. Theorem 2 also carries through, although the exact formulation becomes cumbersome. As for Theorem 3, when the gatekeeper's signal is continuous, its strategy will follow a threshold structure. Thus, indifference will almost surely not occur. Finally, a binary classification seems to align with our motivation to examine the utilization of an AI classifier in selection processes.

When we have considered allowing the candidate signal structure to be bounded, two issues emerged. First, our results will carry through in full if the public belief is approximately balanced (i.e., $\mu \approx \frac{1}{2}$). Therefore, releasing the unbounded signals assumption will adversely affect the readability and approachability of our results. A second issue that emerges when signals are bounded is that now multiple equilibria may emerge.

Our goal was to present a model that is tractable on the one hand, yet is sufficiently general and make as few assumptions as possible about the candidate signal structure. Note that by assuming specific distributions, as we do in Section 4, one can calculate the exact premium generated from the gatekeeper's introduction. One can also use a more structured version of our model to analyze self-selection

in various settings. Our third result justifies adding noise into selection processes. This choice of strategy must be communicated to the candidates to induce a more selective candidate behavior. By employing such a strategy, a system designer can also gain critical insight into the selection process. For example, suppose an AI software of known quality plays the role of the gatekeeper. In this case, one can estimate the quality of the test by utilizing a strategic gatekeeper strategy. The designer can thus compare the performance of applicants that the gatekeeper flagged as highly fitting to those who were low fitting and have taken the test only due to the strategic play. When estimating the gatekeeper's quality, the public signal, and the test quality, the designer can approximate the effect of the candidate's self-selection.

Additionally, we argue that when the distributions are "well behaved" (that is, smooth and Lipschitz continuous), the optimal strategy will be one of the pure strategies, i.e., naive gatekeeper or no gatekeeper at all. This line of inquiry will require assuming additional structure. Therefore, We leave it to future work.

# References

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970. doi: 10.2307/1879431.

Itai Arieli and Manuel Mueller-Frank. Multidimensional Social Learning. *The Review of Economic Studies*, 86(3):913–940, may 2019. ISSN 0034-6527. doi: 10.1093/restud/rdy029. URL `https://academic.oup.com/restud/article/86/3/913/5034182`.

Itai Arieli, Moran Koren, and Rann Smorodinsky. The One-Shot Crowdfunding Game. In *EC '18: Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 213–214. ACM, 2018.

Itai Arieli, Moran Koren, and Rann Smorodinsky. The Implications of Pricing on Social Learning. In *ACM Conference on Economics and Computation (ACM-EC)*, Phoenix, Arizona, 2019.

Amir Ban and Moran Koren. A Practical Approach to Social Learning Analysis. *Working Paper*, 2020a. ISSN 00251895. URL `http://search.ebscohost.com/login.aspx?direct=true{&}db=buh{&}AN=6031936{&}site=ehost-live{&}scope=site`.

Amir Ban and Moran Koren. Sequential fundraising and social insurance. *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 45–46, 2020b.

Abhijit V. Banerjee. A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992. ISSN 00335533. doi: 10.2307/2118364. URL `http://www.jstor.org/stable/2118364`.

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. ISSN 0022-3808. doi: 10.1086/261849.

Sushil Bikhchandani, David Hirshleifer, Omer Tamuz, and Ivo Welch. Information cascades and social learning. (28887), June 2021. doi: 10.3386/w28887. URL `http://www.nber.org/papers/w28887`.

Marina Halac, Ilan Kremer, and Eyal Winter. Raising Capital from Heterogeneous Investors. *American Economic Review*, 110(3):889–921, mar 2020. ISSN 0002-8282. doi: 10.1257/aer.20190234.

Emir Kamenica. Bayesian Persuasion and Information Design. *https://doi-org.ezp-prod1.hul.harvard.edu/10.1146/annurev-economics-080218-025739*, 11:

249–272, aug 2019. doi: 10.1146/ANNUREV-ECONOMICS-080218-025739. URL https://www-annualreviews-org.ezp-prod1.hul.harvard.edu/doi/abs/10.1146/annurev-economics-080218-025739.

Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615, oct 2011. ISSN 0002-8282. doi: 10.1257/AER.101.6.2590.

David Lagziel and Ehud Lehrer. A Bias of Screening. *American Economic Review: Insights*, 1(3):343–56, sep 2019. doi: 10.1257/AERI.20180578.

David Lagziel and Ehud Lehrer. Screening Dominance: A Comparison of Noisy Signals. *American Economic Journal: Microeconomics (Forthcoming)*, 2021. ISSN 1945-7669. doi: 10.1257/MIC.20200284.

Jonathan Levin. Information and the Market for Lemons. *The RAND Journal of Economics*, 32(4):657, 2001. doi: 10.2307/2696386.

Lones Smith and Peter Sørensen. Pathological Outcomes of Observational Learning. *Econometrica*, 68(2):371–398, 2000.

Lones Smith, Peter Norman Sørensen, and Jianrong Tian. Informational Herding, Optimal Experimentation, and Contrarianism. *The Review of Economic Studies*, 88(5):2527–2554, sep 2021. ISSN 0034-6527. doi: 10.1093/RESTUD/RDAB001. URL https://academic.oup.com/restud/article/88/5/2527/6149316.