Algorithmic Transparency With Strategic Users

Qiaochu Wang Yan Huang Stefanus Jasin Param Vir Singh*

Abstract

Should firms that apply machine learning algorithms in their decision-making make their algorithms *transparent* to the users they affect? Despite the growing calls for algorithmic transparency, most firms have kept their algorithms *opaque*, citing potential gaming by users that may negatively affect the algorithm's predictive power. In this paper, we develop an analytical model to compare firm and user surplus with and without algorithmic transparency in the presence of strategic users, and present novel insights. We identify a broad set of conditions under which making the algorithm transparent actually benefits the firm. We show that, in some cases, even the predictive power of the algorithm can increase if the firm makes the algorithm transparent. By contrast, users may not always be better off under algorithmic transparency. These results hold even when the predictive power of the opaque algorithm comes largely from correlational features and the cost for users to improve them is minimal. We show that these insights are robust under several extensions of the main model. Overall, our results show that firms should not always view manipulation by users as bad. Rather, they should use algorithmic transparency as a lever to motivate users to invest in more desirable features.

Keywords*: Algorithmic Transparency, Game Theory, Machine Learning, Strategic Classification, Signaling Game*

*Qiaochu Wang, Yan Huang and Param Vir Singh are at Carnegie Mellon University. Stefanus Jasin is at University of Michigan.

1 Introduction

Machine learning algorithms are being increasingly applied to decision-making processes with far-reaching impacts extending to employment, access to credit, and education (Schellmann and Bellini, 2018, Fu et al., 2019). However, firms typically keep these algorithms as closely guarded secrets, on par with KFC or Coca-Cola's recipes. As a result, these algorithms stay opaque to the people whom they affect and lack clear explanations for the decisions they make.

Our study is motivated by the growing calls from different parts of society to require firms to make their algorithms transparent. According to American privacy law expert Marc Rotenberg: "At the intersection of law and technology – knowledge of the algorithm is a fundamental right, a human right."[1] The European Union's General Data Protection Regulation (GDPR) dictates that, whenever personal data is subject to automated decision making, people have "the right to obtain human intervention on the part of the controller" or the right to explanation.[2]

While making algorithms transparent is desirable, it can open a door to gaming, which potentially adversely affects the algorithms' predictive power. If strategic agents were to know the information about a classifier (i.e., how observed attributes affect the classification outcome), they may manipulate their attributes to receive a more desirable classification, hurting the predictive power of the algorithm. In financial and economic policy making, this problem is widely known as Goodhart's Law, which proclaims that "when a measure becomes a target, it ceases to be a good measure" (Goodhart, 1984). A similar notion is captured in the Lucas critique (Lucas et al., 1976). In fact, Fair Isaac Corporation keeps its exact credit scoring formula secret to make it more difficult for consumers to game the algorithm (Citron and Pasquale, 2014). Similarly, Google continues to modify its secret search ranking algorithm to keep businesses and people from gaming the system (Segal, 2011).

Motivated by the calls for algorithmic transparency and the threat of manipulation by agents to transparent algorithms, we investigate how algorithmic transparency may affect firms and agents. First, from the perspective of the firm (the decision-maker), we ask, is there any advantage in making its algorithm transparent even when there is the potential of manipulation by the agents? Second, we ask, would the agents be better off or worse off if firms make their algorithms transparent (i.e., if the agents receive more information about the factors affecting algorithmic decisions)? Third, we ask, how are the results affected by the predictive power of those features that are more susceptible to gaming by the agents? Finally, we ask, how does the market composition in terms of desirable and undesirable agents affect these results? In this paper, we develop and analyze a

[1] See "Algorithmic Transparency: End Secret Profiling," https://epic.org/algorithmic-transparency/

[2] See "Algorithmic transparency and the right to explanation: Transparency is only the first step," https://www.apc.org/en/blog/algorithmic-transparency-and-right-explanation-transparency-only-first-step

game-theoretic model to answer these questions.

We explicitly model the agents as strategic and the algorithm designer (the firm) is aware of the potential for manipulation. Hence, the firm can react to gaming by the agents. For example, consider a setting where the firm collects data, trains an algorithm that maps a set of observed features to a classification outcome, and publishes a decision rule. If agents desire to be positively classified, they would manipulate their values of the features to achieve it. However, the firm would be aware that the behavior of the agents has changed. It will then update the model and the decision rule. The agents' behavior would change once again. Over time, the firm will iterate to a *fixed point* – this decision rule would be the best response to the agents' strategic behavior.

More specifically, in this paper, we consider a *job hiring* scenario where a risk-neutral firm offers a fixed wage and wants to recruit only highly productive agents. There are two types of agents, High talent (H) and Low talent (L). The H type agents are more productive compared to the L type agents. While the type of each agent is fixed, it is not observed by the firm ex-ante. The firm, however, does have access to a number of observed features (i.e., the *observables*), which it uses to infer the agents' types (i.e., using historical data and an algorithm) and to determine a decision rule for hiring the agents.

We model two types of observables, *causal* and *correlational*. Typically, in machine learning models, the designer focuses primarily on model accuracy, not causality. However, any features that are captured by the machine learning model can still be classified as either *causal* or *correlational*. For simplicity, we consider only two features; one is causal and the other is correlational. There are several characteristics of these features that are important for our model. By definition, the causal feature impacts the productivity of the agent, whereas the correlational feature does not. The difference between 'causal' and 'correlational' features is similar to the difference between 'improvement' and 'gaming', two terms commonly used in the strategic classification literature (Kleinberg and Raghavan, 2019, Alon et al., 2020, Haghtalab et al., 2020). Putting an effort to change the value of a causal feature in the decision maker's favorable direction is called an 'improvement' because this is beneficial to the decision maker or the society. By contrast, putting effort on to change the value of a correlational feature does not benefit the firm and is a waste from the perspective of social welfare.

The agents can game (alter) both features by incurring a cost. As is typically assumed in most signaling game models (Spence, 1973), the H type agents have a cost advantage on the causal feature over the L type agents. As for the cost of improving the correlational feature, in the main model, we assume that this cost is independent of the agents' type and its value is marginally above zero.[3]

[3]For completeness, we also discuss the setting where the cost is significantly larger than zero in Appendix B.1.

The assumptions behind the cost structures of the causal and correlational features warrant some discussion. When an H type agent has a significant cost advantage over an L type agent on the causal feature, it will trivially lead to a separating equilibrium where only the H type agents get high values on the causal feature. In the case where the cost advantage of the H type agents on the causal feature is not too large, the firm would want to include the correlational feature in the machine learning model and in the decision rule. In this case, the correlational feature would provide an additional value in separating the two types. It is easy to see that the correlational feature is the one that is more susceptible to gaming. If the agents game it, this feature can lose its predictive power. This is precisely the reason that is typically purported for opposing algorithmic transparency. If the cost to alter the correlational feature is very high, or if the H type agents have an advantage on it, either gaming will not happen or gaming will be more favorable to the H type agents. In such a case, making the algorithm transparent would be either better or at least as good as keeping it opaque for the firm. Our interest is in investigating whether algorithmic transparency can be better for the firm as opposed to keeping the algorithm opaque even when the H type agents have no cost advantage on the correlational feature and the cost to manipulate the correlational feature is minimal, which makes the algorithm particularly highly susceptible to gaming.

We solve the game between the firm and the agents in two scenarios, an *opaque* algorithm scenario and a *transparent* algorithm scenario. In both scenarios, we use *perfect Bayesian equilibrium* (PBE) as the solution concept. In the opaque scenario, the agents move first. In this scenario, we assume that the agents are aware of the causal feature but have little knowledge of the correlational feature. Consequently, the agents can only improve the causal feature. The agents know that the firm uses a correlational feature, but they do not know what that feature is. However, it is common knowledge that the feature has predictive power that can help the firm separate the H type agents from the L type agents. This assumption is reasonable: If a feature has no predictive power in separating the two types of agents, the machine learning algorithm would simply discard it.

In the transparent scenario, the firm moves first and publishes its algorithm. The agents observe this algorithm and know what features are considered by the algorithm including their respective weights. In this scenario, the agents can game both the causal and the correlational features. They also know how the correlational feature correlates with their types before any gaming occurring. Finally, the firm decides who to hire based on the agents' ending feature values.

It is worth noting that although we assume the agents and the firm move sequentially in both the opaque and transparent scenarios, we do not model these processes as Stackelberg leadership models in our main analysis. The reasons are as follows. In a Stackelberg model, the first mover typically anticipates that any deviation from the current strategy will trigger the second mover to react accordingly. However, when the first mover consists of many uncoordinated agents, this

condition will be violated. Specifically, in the opaque scenario, given that the agents are non-cooperative, the unilateral deviation of a single agent will not change the follower's (i.e., the firm's) strategy. As for the transparent scenario, there is a single first mover (i.e., the firm), and it is possible for the firm to commit to the published algorithm. Consequently, a Stackelberg model is a valid model in this scenario. However, a Stackelberg model provides an advantage to the first mover. As a result, the firm may prefer the transparent algorithm over the opaque algorithm simply because of the first mover advantage that the transparent algorithm provides. To clearly explain the mechanisms of the firm's preference between algorithmic transparency and opacity and show that the firm could still prefer the transparent algorithm over the opaque algorithm even in absence of the first mover advantage, we will first focus our analysis in this paper on the case where the firm either does not have the commitment power or the firm only reveals which features it uses but not the full details of its hiring strategy (akin to 'partial transparency'). Later, in Section 5, we will discuss the case where the firm has commitment power and reveals the full details to the agents (akin to 'full transparency'). Overall, we will show that the key insights from the non-Stackelberg model will only be strengthened in the Stackelberg model.

In addition to analyzing the models discussed above, as briefly described in Section 6 and further elaborated in Appendix B, we also analyze three extensions by relaxing some of the assumptions made in the main model. These extensions include: (1) the case where the cost of improving the correlational feature is significantly larger than zero, (2) the case where the wage is endogenously chosen by the firm, and (3) the case where the agents have incorrect beliefs about the predictive power of the correlational feature. Our analysis for these three extensions shows that the results from the main analysis are robust to these alternative assumptions and modeling choices. Specifically, the main results of our paper will not change qualitatively in extension (3), and they will only be strengthened in extensions (1) and (2).

Key results and insights

Our first result in this paper challenges the conventional wisdom that making algorithms transparent will always hurt the firm economically. We identify a broad set of conditions under which making the algorithm transparent is actually beneficial for the firm. The key intuition behind this result is driven by how the H type and L type agents respond to algorithmic transparency. Since investment on the causal feature is costly and since the H type agents have an advantage on the correlational feature (we assume that an H type agent has a higher probability of being 'high' on the correlational feature, otherwise the firm will have no incentive to use this feature in the first place), the H type agents would invest in improving the causal feature only to the extent that it, along with the correlational feature, helps them separate themselves from the L type agents. When the algorithm is made transparent, the L type agents game the correlational feature. As a result, the L type agents

become similar to the H type agents on that feature, and the predictive power of the correlational feature decreases. Hence, the H type agents have to invest more in the causal feature to separate themselves from the L type agents. When the H type agents have a significant cost advantage over the L type agents on the causal feature, this leads to full separation, which benefits the firm. When the H type agents only have a marginal cost advantage over the L type agents on the causal feature, the L type agents will also invest significantly in the causal feature. In this case, both the H type and L type agents become more productive because of their higher investment in the causal feature. Although the firm cannot separate the two types of agents in this case, when the impact of the causal feature on productivity is above a certain threshold, the average productivity of the hired agents is significantly higher than in the opaque scenario. In other words, making the algorithm transparent allows the firm to motivate the agents to invest in improving features that are valuable to the firm.

Our second result in this paper is that the agents are not always better off under the transparent scenario. This might appear counter-intuitive at first: Since the agents would have access to more information under the transparent scenario, one would think that they should be better off under the transparent scenario. However, we show that there are conditions under which the agents are worse off in the transparent scenario. Interestingly, in most cases where the firm prefers the transparent scenario, the agents would prefer the opaque scenario, and vice versa. But we also identify a set of conditions where both the firm and the agents prefer the transparent scenario. The intuition for our second result is similar to that for the first result. The firm prefers the transparent scenario in situations where transparency motivates the agents to invest highly in the causal feature. When the cost of investment is high and the H type agents have a significant advantage, only the H type agents will invest in the causal feature. In this situation, although only the H type agents are hired, they are worse off due to the high cost that they incur. When the investment cost is moderate and the H type agents have a marginal cost advantage over the L type agents, both types of agents invest in the causal feature, and both are hired. The agents are better off because they have to incur only a moderate cost for being hired. Simultaneously, the firm is also better off since the average productivity of the hired agents is higher when the impact of the causal feature on productivity is above a certain threshold.

Our third result shows that it is possible for the firm to prefer algorithmic transparency when the correlational feature has high predictive power and prefers opaque algorithm when the correlational feature has low predictive power. This result also appears to be counter-intuitive since one would naturally expect that, as the predictive power of the correlational feature increases, the firm would be better off keeping the algorithm opaque. We find that this is not always the case. The key intuition behind this result is as follows: When the correlational feature is good at separating the

H type agents from the L type agents, the H type agents have little incentive to invest in the causal feature in the opaque scenario. However, under transparency, the H type agents lose this advantage and have to invest in the causal feature to separate themselves from the L type agents. As a result, the firm can hire more productive agents under transparency.

Our fourth result is that, when the fraction of the H type agents on the market is higher, the firm may have a stronger preference for the transparent algorithm under certain conditions. More specifically, the fraction of H type agents affects the firm's surplus significantly but does not have a large enough impact to alter the firm's decision for/against transparency when the cost for improving the causal feature is either too high or too low, or the H type agents have a large cost advantage over the L type agents. However, when the cost for improving the causal feature and the cost advantage of the H type agents are both moderate, the firm has a stronger preference for algorithmic transparency as the fraction of the H type agents on the market increases. In this cost range, both the H type and L type agents would improve the causal feature. While the firm is unable to separate the two types and hires both types, the number of the L type agents that are hired becomes smaller as the fraction of the H type agents on the market increases.

This paper makes several contributions. It is one of the first to provide an analytical model that systematically compares a firm's decision for algorithmic transparency *versus* opacity in the presence of strategic agents. We show that, counter to conventional wisdom, the firm can be better off under algorithmic transparency. Moreover, in most cases where the firm prefers algorithmic transparency, the agents will be worse off. Agents underinvest in the causal feature when the algorithm is opaque. Consequently, the firm depends heavily upon the correlational feature to separate different types of agents from one another. Our analysis and results show that the firm should not always worry about the potential loss of the predictive power of the correlational feature in its machine learning model under transparency. Rather, it can use algorithmic transparency as a lever to motivate the agents to invest more in the causal feature. The firm would typically be reluctant to adopt algorithmic transparency when their machine learning model derives large predictive power from the correlational feature. However, we show that the firm should recognize that investment in the causal feature by agents is endogenous. The H type agents are less likely to invest in the causal feature when the correlational feature is sufficient to separate them from the L type. This is the situation where the firm should be willing to sacrifice the predictive power of the correlational feature. We have demonstrated our results in the setting where (1) the firm is certain it will lose the predictive power of the correlational feature and (2) the firm does not have a first mover advantage under algorithmic transparency. Intuitively, one would think that algorithmic transparency would be bad for the firm when these two conditions hold. However, we show that, once we consider the endogenous investment in the causal feature, the firm would be

better off making the algorithm transparent.

Organization of the paper

The rest of the paper is organized as follows. We discuss how we build upon and contribute to the literature in Section 2. The details of our main model are presented in Section 3. In Section 4, we present the analysis of the main model. In Section 5, we discuss the model of Stackelberg competition. In Section 6, we describe three extensions of the main model. We conclude the paper in Section 7. Unless otherwise noted, all the proofs can be found in Appendix D.

2 Literature

Algorithmic transparency is a relatively new topic, but it is closely related to the literature on information asymmetry. Following the canonical job market signaling model developed by Spence (1973), a rich stream of research has focused on the interaction between a decision maker and strategic agents under asymmetric information. Some of this research is focused on the agents' side and study how agents strategically reveal their type under various market conditions (e.g., the stream of signaling game literature). Other research is focused on the decision makers' side, studying how they can design an optimal algorithm to extract agents' private information (e.g., the stream of strategic classification literature). Our work on algorithmic transparency is built upon and contributes to both these streams of literature.

In both the opaque and transparent scenarios, the interaction between the firm and the agents can be adapted into the signaling game framework, where individuals (senders) first send signals, and the firm (receiver) then makes hiring decisions based on the observed signals (Spence, 1973, Engers, 1987, Weiss, 1983, Daley and Green, 2014). Signaling game models often focus on specifying the equilibrium outcome under various market conditions. Receivers are assumed to be in a competitive environment and earns zero profit in the equilibrium. All the surplus is extracted by senders. The equilibrium concept typically used in signaling game models is the perfect Bayesian equilibrium (PBE), where three conditions are satisfied: senders use optimal strategies facing the wage offer, receivers give wage offers such that they will obtain zero profit, and the receivers' beliefs about the senders' type given the signal are consistent with the truth. Similar to the signaling games, we also specify the equilibrium outcome under various market conditions in both the transparent and opaque scenarios. However, in our model, the firm offers a fixed wage and focuses on designing the algorithm to increase its chances of hiring the most productive agents, which contrasts with the signaling games where all agents are hired and the firm's objective is to decide how much salary to offer to each agent.

On the firm side, our model setup bears more similarity to the strategic classification problems – offering a fixed wage, the firm decides whether to hire the agents based on their signal (Kleinberg

and Raghavan, 2019, Frankel and Kartik, 2019, Bonatti and Cisternas, 2019). In other words, the firm is trying to classify agents based on whether their expected productivity exceeds the wage or not. This classification setting makes it possible for us to analyze the economic impact of algorithmic transparency on the decision maker (firm) by comparing its equilibrium payoff in the opaque and transparent scenarios. Moreover, we do not assume the signal (education level for example) to always be pure money-burning. Instead, we allow the causal feature to positively impact productivity and specify conditions regarding this positive effect under which algorithmic transparency benefits the decision maker. In that aspect, our work is also closely related to the strategic classification literature that assumes the existence of both causal and non-causal features.

Signaling Games. The signaling game literature studies how agents strategically reveal their type to a principle in a situation of information asymmetry. Traditional signaling models typically assume that costly actions are the only channels through which agents can signal their type (Spence, 1973). In these models, standard assumptions such as the Spence-Mirrlees single-crossing condition ensure the existence of separating equilibria: equilibria that fully reveal agents' private information. While the machine learning models are trying to solve the same problem (i.e., trying to identify the type of agents under information asymmetry), they differ from decision makers in the classical signaling models in the following way. A machine learning model uses multiple features to learn an agent's type. Each feature is essentially an action taken by the agent that signals her type. Some of these features are costly to improve, while others are not.

Our paper is related to recent signaling game papers that have also considered multiple actions as channels through which an agent can signal her type (Engers, 1987, Frankel and Kartik, 2019, Daley and Green, 2014, Alós-Ferrer and Prat, 2012). In these papers, the agents are always aware of the actions that are used as signals by the decision maker. In contrast, in our model's opaque scenario, the agents know that a correlational feature is being used by the firm, but they do not know exactly what feature that is.

In our model, agents can use causal and/or correlational features to signal their type. The causal feature is similar to the costly signal typically captured in the traditional signaling game literature. A key difference is that we allow it to not only act as a signal of an agent's type but also have an impact on the agent's productivity, similar to Weiss (1983). For example, if agents of the same type have different levels of education, in our model, the firm would receive different payoffs from hiring them. The correlational feature that we model bears some similarities to the information in cheap talk games (Crawford and Sobel, 1982). This feature is almost costless to share, and it affects the eventual payoff of both the firm and the agents where their incentives are not perfectly aligned. In cheap talk games, the agent strategically manipulates this information, whereas, in our model, the agent does not know about this feature and cannot manipulate it in the

opaque scenario.

Similar to our paper, a few recent papers have modeled the tradeoffs that an agent faces in the presence of multiple signals. For example, Daley and Green (2014) modeled a scenario where a student can send a costly signal (e.g., joint degree completion) to the recruiter or rely on a type-correlated noisy signal (e.g., grades). They characterized the results based on the informativeness of the noisy signal. The noisy signal is similar to the correlated feature in our model, and its informativeness is also modeled similarly to how we capture the predictive power of the correlated feature. A key finding of the paper is that, when grades are informative, H type individuals are less eager to send costly signals because they can now rely on grades to signal their type, while L type individuals are more willing to send a costly signal to de-emphasize the 'grades' dimension. Consequently, a separating equilibrium on the costly signal dimension is harder to sustain. Our model also shares some similarities with this paper in that we also consider the possibility that the existence of an extra noisy signal will change individuals' decision on the more costly signal. However, there are key differences in our model's assumptions and results. Unlike the 'grades' dimension, whose value is impossible to manipulate, in our model, the firm can give individuals the opportunity to game the correlational feature by making the algorithm transparent. On the one hand, manipulation will make this dimension less informative. On the other hand, individuals' behavior on the more costly signaling dimension will also change and could lead to a separating equilibrium in many cases.

Strategic Classification. Strategic classification literature considers the problem of designing optimal classification algorithms when facing strategic users who may manipulate the input to the system at a cost (Hardt et al., 2016). Canonical strategic classification models deem that the user's manipulation always hurts the decision maker. Guided by this belief, a large stream of research on strategic classification is focused on developing algorithms that are robust to gaming (Meir et al., 2012, Cummings et al., 2015). Recently, several papers have argued that this gaming itself can be beneficial to the decision maker; thus, instead of focusing on manipulation-proof algorithms, these papers focus on designing algorithms that incentivize individuals to invest in desirable features (Kleinberg and Raghavan, 2019, Alon et al., 2020, Haghtalab et al., 2020). These papers are the ones we want to highlight since our paper also points out the difference between 'gaming' and 'improvement': gaming is bad for the decision maker because it deteriorates the information contained in the relevant features, but 'improvement' could be beneficial to the decision maker since it will causally impact the target variable.

Kleinberg and Raghavan (2019) studied the principle-agent problem where the agents' features (e.g., final exam score) can be improved in two ways: by investing effort in a desirable way (e.g., spending time on course material) or by investing effort in an undesirable way (e.g., cheating). The

effectiveness of each kind of effort on the feature is called the *effort profile*. The decision maker can observe the agents' performance on the features but cannot observe in which way the agents achieve their scores. Alon et al. (2020) examined a similar setting but extended Kleinberg and Raghavan (2019)'s model into a multi-agent scenario. Instead of assuming all individuals share the same *effort profile*, they focused on designing optimal algorithms that can work for different groups of individuals who may have different effort profiles. Our work is different from theirs with respect to one important aspect: while they assumed that a feature can be improved in either a causal or non-causal way, our model assumes that there are pure causal features and pure correlational features. Causal features (e.g., education level) can be improved only in a 'causal' way (such that the value of the target variable will also increase). Correlational features (e.g., whether an applicant wears glasses or not) can be improved only by gaming. If there was a 'causal' way to improve the correlational feature, then the firm's willingness to publish the correlational feature would be stronger and might come from the potential productivity-enhancing effect of individuals' improvement on the correlational feature. We show that even in the case where gaming on the pure correlational feature has no positive effect on productivity, the firm may still want to publish it. Furthermore, none of the papers above has studied how firms choose between opaque and transparent algorithms.

Another two papers we want to highlight are Frankel and Kartik (2019) and Bonatti and Cisternas (2019). These two economics papers showed that the decision maker could be ex-ante better off by committing to some ex-post sub-optimal strategies such as down-weighting some relevant features. Our paper is related to these papers in the sense that 'publishing the algorithms' could also be seen as a way to down-weight the correlational feature. However, there are critical differences between these papers and ours in terms of mechanisms. In their papers, ex-post sub-optimal behavior such as 'under-utilizing' some informative features might be preferred by the decision maker because individuals may have less incentive to manipulate these features if they anticipate that the features will be down-weighed. The decision maker loses some predictive accuracy due to under-utilizing those features, but the observed values for those features now better represent individuals' natural behavior instead of gaming behavior. Consequently, not fully exploiting the information contained in relevant features might be optimal ex-ante. Our paper, however, shows that even if such 'feature down-weighting' cannot reduce individuals' gaming, it may still benefit the decision maker. In our paper, the purpose of 'down-weighting' the correlational feature is not to reduce or eliminate gaming behavior, but rather to increase the competition intensity regarding the causal features on which the H type agents hold a cost advantage.

3 Model

In this section, we develop a parsimonious model that captures how the firm and agents act under opaque and transparent scenarios. We consider the hiring setting discussed above with two types of agents: high-talent (H type) agents and low-talent (L type) agents. For simplicity, we normalize the total number of agents to 1 and assume that a θ portion of them are of H type and the remaining $1-\theta$ portion are of L type.

Talent level is directly related to job performance and, ideally, the firm would like to hire only H type agents. However, the firm cannot directly observe an agent's type until she is hired and works at the firm for a while. Consequently, the firm can only use some observable agent features to differentiate the two types of agents. We classify these features into two types: causal features and correlational features. For simplicity, we assume that the firm only uses one causal feature (which is common knowledge to the firm and the agents, e.g., education level) and one correlational feature (which is unknown to the agents unless the firm decides to reveal it). Both features take on a discrete value of 0 (low) or 1 (high). Each agent can be characterized by one of four possible combinations (or states) in the two-dimensional feature space:

- State A (low causal, high correlational);
- State B (high causal, high correlational);
- State C (low causal, low correlational); and
- State D (high causal, low correlational).

The firm's hiring strategy can therefore be represented by four hiring probabilities for the four states. In the remainder of this paper, we will refer to these probabilities as P_A, P_B, P_C, and P_D, respectively.

We assume that it is costly for the agents to improve the common-knowledge causal feature, and that H type agents have a cost advantage on this feature. Specifically, we assume that C_H (the cost of improving the causal feature for the H type agents) is smaller than C_L (the cost of improving the causal feature for the L type agents). In contrast, the cost to improve the correlational feature is assumed to be the same for the H type and L type agents and is very small (i.e., marginally above zero). It is worth noting that, although the cost of improving any given correlational feature is small, there are many of them, and agents do not know which correlational feature will be used in the algorithm unless the firm decides to reveal it.

To model the situation where the firm has an incentive to include the correlational feature into its decision making, we further assume that a λ portion of the H type agents and a $1-\lambda$ portion of

the L type agents have value 1 on the firm's chosen correlational feature.[4] Moreover, $\lambda \in [0.5, 1]$, which indicates a positive correlation between an agent's value on the correlational feature and her type. The true value of λ is known to the firm. We assume that the agents also have a correct belief about λ in the main model. This assumption will be relaxed in Appendix B.3 where we allow the agents to have an 'incorrect belief' about λ.

The game between the firm and the agents is played as follows. In the first stage, the firm makes a decision on transparency (i.e., opaque or transparent), and this decision is known to all agents. If the firm chooses "opaque," the remainder of the game proceeds as follows: The agents first choose their strategies in terms of whether to improve their features, and then the firm makes its hiring decisions based on the observed agent features. If, on the other hand, the firm chooses "transparent," the remainder of the game proceeds as follows: The firm first discloses the algorithm to the agents; next, the agents choose their strategies; and, finally, the firm decides whom to hire based on the observed agent features. Recall from Section 1 that, in the main model, we focus our analysis on the case where the firm either does not have commitment power, or reveals only the features it uses but not its hiring probabilities.[5]

In order to focus on the more interesting and realistic cases, we make the following assumptions regarding the strategies of the agents:

1. *In the opaque scenario, the agents will only focus on whether to improve their causal feature.* This assumption is motivated by the fact that, in reality, while causal features are usually common knowledge between the firm and the agents (e.g., everyone knows that education level plays an important role in hiring decisions), correlational features are less so. In the opaque scenario, the firm does not reveal which correlational feature it will use in its algorithm to the agents. Consequently, the best that an agent can do is make a random guess. Since there are a large number of potential correlational features and the firm only uses one of them, from the individual agent's perspective, it is not profitable to improve any of the potential correlational features because the probability of hitting the right one is essentially zero. From a theoretical perspective, we note that this assumption is without loss of generality and is useful to simplify our analysis. It is not difficult to show that the main insights of the paper will not change qualitatively even if we assume a small probability $\epsilon > 0$ that the agents will hit the right correlational feature.

[4] While agents may not know which correlational feature the firm is using, they may attempt to game on as many correlational features as possible. In this scenario, λ $(1 - \lambda)$ can be interpreted as the probability of H (L) type agents hitting the feature that is actually used by the firm

[5] In Section 5, we will discuss the case where the firm has commitment power and reveals both the features it uses and its hiring strategy.

2. *In the transparent scenario, all agents will improve their correlational features.*[6] Once the firm reveals the correlational feature that it will use in its algorithm, the probability that an individual agent hits the right feature becomes 1. Since the cost of improving the correlational feature is minimal, as long as it increases an agent's probability of being hired, they will improve this feature. It is worth noting that assuming all agents will improve the correlational feature used in the algorithm does not cause a loss of generality. This is so because, under the scenario where all agents achieve a "high" state on the disclosed correlational feature, this feature completely loses its predictive power and, therefore, will drop out of the prediction algorithm. If it can be shown that the firm can still be better off by making its algorithm transparent in such an extreme case of "agent gaming," it sends a strong message that algorithmic transparency can indeed be economically beneficial. This assumption will be relaxed in Appendix B.1 where we allow gaming cost of the correlational feature to be significantly greater than zero. Our main result is strengthened after the relaxation of this assumption.

3.1 The Firm's Utility

As previously mentioned, the causal feature (for consistency, hereafter, we will use education level as an example of a causal feature) has a direct influence on the agents' performance, while the correlational feature does not. Thus, an agent's performance is determined by both their type ($T \in \{H, L\}$) and their education level ($Education \in \{0, 1\}$). Here, we follow Weiss (1983) and allow education to not only act as a signal of an agent's type but also contribute to the productivity of the agent. We use α and β to denote the marginal effects of type and education, respectively. For convenience, we normalize the performance of an uneducated L type agent to 0. The mathematical expression of an agent's performance is given by:

$$W(T, Education) = \alpha \times \mathbb{1}(T = H) + \beta \times \mathbb{1}(Education = 1). \tag{1}$$

In the opaque scenario, each agent is characterized by her type $T \in \{H, L\}$ and her state $S \in \{\text{A, B, C, D}\}$ at the end of the game. We can write an agent's performance as a function of her type and her final state as follows:

$$W_S^T = \alpha \times \mathbb{1}(T = H) + \beta \times (\mathbb{1}(S = B) + \mathbb{1}(S = D)). \tag{2}$$

In the transparent scenario, since all agents are "high" on the correlational feature, they only differ on the causal feature (i.e., education). This means that we can reduce the number of possible

[6] It's worth noting that although the correlational feature will be dropped if everyone improves it, the level of transparency actually increases. The reason is that agents know exactly the state they are in and thus are more certain about their chances of being hired.

states from four to two, i.e., state E (low education) and state F (high education). Using these state definitions, we can write an agent's performance as follows:

$$W_S^T = \alpha \times \mathbb{1}(T = H) + \beta \times \mathbb{1}(S = F). \tag{3}$$

Once the agents are hired, their performance will contribute to the firm's payoff, and the firm will pay them a fixed reward R (i.e., job compensation). In the main model, we assume that the reward R is the same for both transparent and opaque scenarios, and that its value is exogenously given. In Appendix B.2, we will endogenize R and allow the firm to potentially use different rewards for the transparent and opaque scenarios. We will show that our main insights still hold.

Let n_S^T denote the number of T type agents whose final states are S, and let $n_S = n_S^H + n_S^L$ denote the total number of agents whose final states are S. Furthermore, let $\gamma_S^e = n_S^H / n_S$. The firm's expected total payoff (for narrative convenience, we use 'payoff' to refer to 'expected payoff' hereafter) under hiring strategies (or probabilities) $P = (P_A, P_B, P_C, P_D)$, in the opaque scenario, or $P = (P_E, P_F)$, in the transparent scenario, can be mathematically expressed as

$$\begin{aligned} \Pi_{firm} &= \sum_S P_S \cdot \left[n_S^H (W_S^H - R) + n_S^L (W_S^L - R) \right] \\ &= \sum_S P_S \cdot n_S \cdot \left[\gamma_S^e (W_S^H - R) + (1 - \gamma_S^e)(W_S^L - R) \right]. \end{aligned} \tag{4}$$

3.2 The Agents' Utility

In the opaque scenario, the agents do not know which correlational feature will be used by the firm's algorithm; therefore, they will only focus their decisions on whether to improve the causal feature (i.e., education). Agents of the same type use the same strategy. Let u_T denote the expected utility (for narrative convenience, we use 'utility' to refer to 'expected utility' hereafter) of a T type agent. We have:

$$u_H = \begin{cases} \lambda P_B R + (1 - \lambda) P_D R - C_H & \text{if the } H \text{ type agent improves the causal feature,} \\ \lambda P_A R + (1 - \lambda) P_C R & \text{otherwise;} \end{cases} \tag{5}$$

$$u_L = \begin{cases} (1 - \lambda) P_B R + \lambda P_D R - C_L & \text{if the } L \text{ type agent improves the causal feature,} \\ (1 - \lambda) P_A R + \lambda P_C R & \text{otherwise.} \end{cases} \tag{6}$$

In the transparent scenario, all agents will have "high" values on the correlational feature, and their decisions are whether to improve the causal feature or not. The utility of a T type agent in the transparent scenario is:

$$u_H = \begin{cases} P_F R - C_H & \text{if the } H \text{ type agent improves the causal feature,} \\ P_E R & \text{otherwise;} \end{cases} \tag{7}$$

$$u_L = \begin{cases} P_F R - C_L & \text{if the } L \text{ type agent improves the causal feature,} \\ P_E R & \text{otherwise.} \end{cases} \tag{8}$$

3.3 Additional Parametric Assumptions

In Section 4, we will solve the game using backward induction. For each combination of (C_H, C_L), we will derive the payoff for the firm in both the opaque and transparent scenarios. We will then specify the range of values of the parameters under which the firm is better or worse off when choosing to be transparent instead of opaque. We will show our results in the C_H-C_L space.

We make the following three additional assumptions regarding the relationships among the parameters to allow us to focus on non-trivial and more interesting cases.

Assumption 1

$$0 < \beta < R < \alpha.$$

Assumption 1 says that the performance of an individual H type agent always exceeds the salary R regardless of her education level, whereas the performance of an individual L type agent is always smaller than the salary R. This condition ensures that the firm only wants to hire the H type agents.

Assumption 2

$$\frac{(\theta\lambda + (1-\theta)(1-\lambda))R}{\theta\lambda} < \alpha < \frac{R}{\theta}.$$

Recall that α denotes the performance advantage of H type agents over L type agents. Assumption 2 ensures that α falls in a certain range that guarantees that the firm will have an incentive to include the correlational feature in its algorithm even when all agents' education levels are 0. The derivations of the lower bound and the upper bound of α can be found in Appendix C.1. Intuitively, if α is too small, the firm will not hire anyone when all agents' education levels are 0. If, on the other hand, α is too large, the firm will hire everyone even when all agents' education levels are 0. In either case, the correlational feature is useless for the firm. We refer to the lower (upper) bound of α as $\underline{\alpha}$ ($\bar{\alpha}$) hereafter.

Assumption 3

$$R - \theta\alpha < \beta < R - \frac{\theta(1-\lambda)\alpha}{\theta(1-\lambda) + (1-\theta)\lambda}.$$

Assumption 3 says that the marginal effect of education on performance (β) falls in a certain range that guarantees the firm will have an incentive to include the correlational feature in the hiring algorithm. The derivations and interpretations of the lower and upper bound of β can be

found in Appendix C.1. Intuitively, if β is too small, the firm will not hire any agent even when everyone has a high level of education in the transparent scenario, which leads to trivial results. If, on the other hand, β is too large, the firm will hire everyone with a high level of education irrespective of their values on the correlational feature in the opaque scenario, which could not justify the firm's incentive to use the correlational feature in the first place. We refer to the lower (upper) bound of β as $\underline{\beta}(\bar{\beta})$ hereafter.[7]

A summary of notations can be found in Table 1 in Appendix A.

4 Analysis

Let γ_S^b denote the proportion that H type agents represent among all the agents in state S at the beginning of the game.[8] Per our discussions in Section 3, a θ portion of agents are of H type and the remaining $1-\theta$ portion are of L type. Moreover, a λ portion of the H type agents and a $1-\lambda$ portion of the L type agents have value 1 on the firm's chosen correlational feature. Therefore,

$$\begin{aligned}\gamma_A^b &= \frac{\lambda\theta}{\lambda\theta+(1-\lambda)(1-\theta)};\\ \gamma_C^b &= \frac{(1-\lambda)\theta}{(1-\lambda)\theta+\lambda(1-\theta)};\\ \gamma_B^b &= \gamma_D^b = 0.\end{aligned}$$

Recall, per our definition in Section 3.1, γ_S^e denotes the proportion that H type agents represent among all the agents in state S at the end of the game.

4.1 Opaque Scenario

Per our discussions in Section 3, in the opaque scenario, agents move first and will only decide on whether to improve the causal feature (i.e., education). Given the agents' strategies, based on Equation 4, the firm will be indifferent between hiring and not hiring agents with a final state S if $\gamma_S^e(W_S^H - R) + (1-\gamma_S^e)(W_S^L - R) = 0$, or equivalently,

$$\gamma_S^e = \frac{R - W_S^L}{W_S^H - W_S^L}. \tag{9}$$

By Equation 2, the above fraction equals $\frac{R}{\alpha}$ when $S \in \{\text{A,C}\}$ and equals $\frac{R-\beta}{\alpha}$ when $S \in \{\text{B,D}\}$. Let $\gamma_{th0} = \frac{R}{\alpha}$ and $\gamma_{th1} = \frac{R-\beta}{\alpha}$ (by Assumption 1, we have $0 < \gamma_{th1} < \gamma_{th0} < 1$). Quantities γ_{th0} and γ_{th1} are important in the analysis, especially in determining whether a certain outcome (i.e., a combination of the agents' strategies and the firm's strategy) can be sustained in the equilibrium.

[7] We provide discussion on the relaxation of Assumption 3 in Appendix C.2.

[8] If there are no agents in state S, we manually set $\gamma_S^b = 0$. The same applies to γ_S^e.

There is a total of nine possible classes of outcomes for agents' strategies. Five of them have the potential to be sustained in an equilibrium but the other four do not. The first five cases are: case 1 (neither H type nor L type agents improve education); case 2 (only H type agents improve education); case 3 (both H type and L type agents improve education); case 4 (H type agents improve education with some probability, and L type agents do not improve education); and case 5 (H type agents improve education, and L type agents improve education with some probability).

Aside from the above five cases, there are four other cases: case 6 (only L type agents improve education; case 7 (L type agents improve education with some probability but H type agents do not improve education); case 8 (L type agents improve education, but H type agents improve education only with some probability); and case 9 (both H type and L type agents improve education with some probability). It is not difficult to see that cases 6 through 9 cannot be sustained in an equilibrium. For cases 6 through 8, the L type agents have a higher value on education than the H type agents, and the firm will have an incentive to set higher hiring probabilities in states A and C than in states B and D. However, under this hiring strategy, the L type agents will have no incentive to improve education in the first place. As for case 9, the fact that both the H type and L type agents are using mixed strategies indicates that they are both indifferent between improving and not improving education. Since the cost of improving education for the L type agents is greater than that for the H type agents, to compensate for this higher cost, the L type agents must have a higher chance of being hired by the firm than the H type agents in the equilibrium. However, the firm has no incentive to use such a hiring strategy. We conclude that although there are nine possible classes of outcomes for the agents' strategies, only five of them (cases 1 through 5) can potentially be equilibrium outcomes.

Out of the five feasible cases, the actual equilibrium strategies of the agents and the firm depend on the values of (C_H, C_L). The following lemma summarizes the agents' equilibrium strategies for different values of (C_H, C_L) and the corresponding payoff for the firm. The proof can be found in Appendix D.1. Since we assume that the H type agents have a cost advantage to improve the causal feature (i.e., $C_H < C_L$), the region above the diagonal line in Figure 1 is infeasible.

Lemma 1 *The equilibrium outcome depends on the values of* (C_H, C_L)*, and this dependence is*

shown in Figure 1. The payoffs for the firm are given by

$$\Pi_{firm_{O1}} = \lambda\theta\alpha - (\lambda\theta + (1-\lambda)(1-\theta))R$$
$$\Pi_{firm_{O2}} = \theta(\alpha+\beta) - \theta R$$
$$\Pi_{firm_{O3}} = \lambda\theta(\alpha+\beta) + (1-\lambda)(1-\theta)\beta - (\lambda\theta + (1-\lambda)(1-\theta))R$$
$$\Pi_{firm_{O4}} = \theta(\alpha+\beta-R)\left(1 - \frac{R(1-\theta)(1-\lambda)}{(\alpha-R)\theta\lambda}\right)$$
$$\Pi_{firm_{O5}} = \frac{2\lambda-1}{\lambda}\theta(\alpha+\beta-R),$$

where $\Pi_{firm_{Oi}}$ *denotes the firm's total payoff in case* i*.*

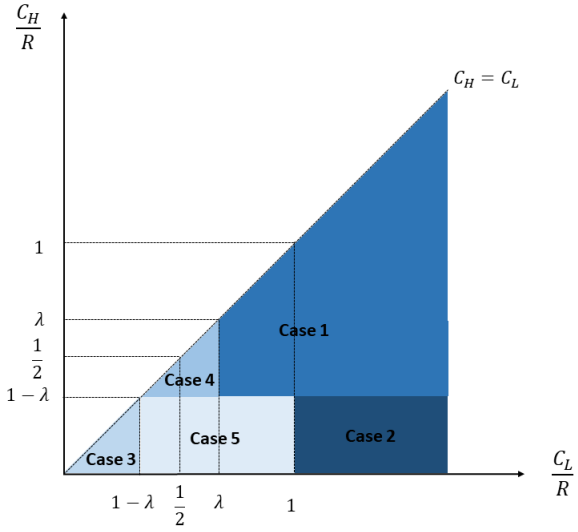


Figure 1: Equilibrium outcome in the opaque scenario

4.2 Transparent Scenario

In the transparent scenario, the firm moves first by announcing both the correlational feature that it uses and the probability of hiring for each state (i.e., P_E and P_F). Similar to the opaque scenario, there is a total of nine possible outcomes for the agents' strategies (we use the same numbering of the nine cases as in the opaque scenario). To determine whether a certain outcome can be sustained as an equilibrium, we use the fact that the firm is indifferent between hiring and not hiring agents with final state S iff $\gamma_S^e = \frac{R-W_S^L}{W_S^H-W_S^L}$. The fraction on the right-hand side of the equality equals γ_{th0} when $S = E$ and equals γ_{th1} when $S = F$.

Consistent with the opaque scenario, cases 6 through 9 cannot be sustained in an equilibrium. Moreover, according to Assumption 3, if everyone is at state F, the firm will hire all agents. In case

5, some L type agents are at state E, which gives the firm even more incentive to hire agents in state F. However, again, to be an equilibrium, the mixed strategy outcome in case 5 requires the firm to be indifferent between hiring and not hiring agents from state F. Therefore, neither case 4 nor case 5 can be sustained in an equilibrium in the transparent scenario. Altogether, this leaves us with only the first three cases as possible equilibrium outcomes. The following lemma summarizes the agents' equilibrium strategies for different values of (C_H, C_L), as well as the corresponding payoff for the firm. The proof can be found in Appendix D.2.

Lemma 2 *The equilibrium outcome depends on the values of* (C_H, C_L)*, and this dependence is shown in Figure 2. The payoffs for the firm are given by*

$$\begin{aligned}\Pi_{firm_{T1}} &= 0\\ \Pi_{firm_{T2}} &= \theta(\alpha+\beta-R)\\ \Pi_{firm_{T3}} &= \theta(\alpha+\beta)+(1-\theta)\beta-R,\end{aligned}$$

where $\Pi_{firm_{Ti}}$ *denotes the firm's total payoff in case* i*.*

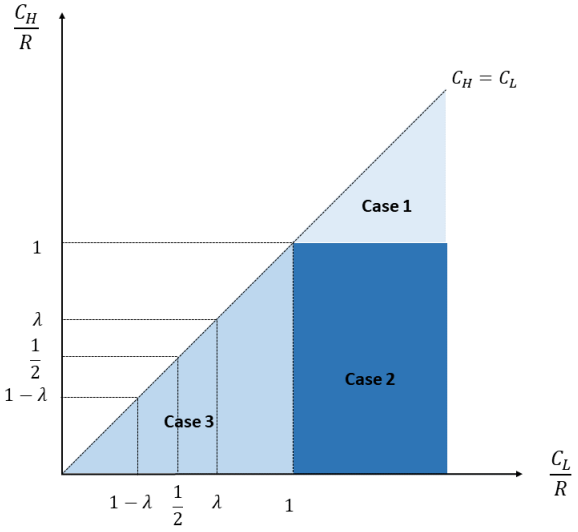


Figure 2: Equilibrium outcome in the transparent scenario

4.3 The Firm's Decision on Algorithmic Transparency

The firm can make a decision on algorithmic transparency by comparing the payoffs in the transparent and opaque scenarios. In this subsection, we will show how the firm is *not* always worse off making its algorithm transparent instead of opaque. We divide the blue region in Figures 1 and 2 into seven smaller regions: $N1$, $N2$, and $N3$ and $C1$, $C2$, $C3$, and $C4$ (see Figure 3).

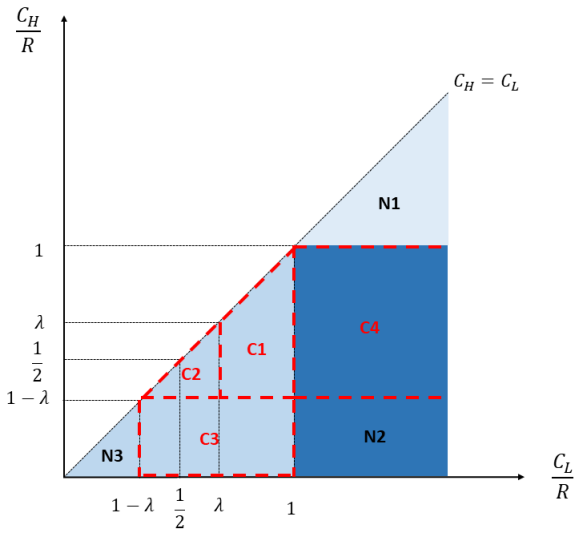


Figure 3: Comparison of agents' equilibrium behavior in the transparent and opaque scenarios

We first consider what happens in regions $N1$ through $N3$. Note that, in these regions, agents play the same equilibrium strategies on the causal feature in the opaque and transparent scenarios. For example, region $N1$ corresponds to case 1 in both the opaque and transparent scenarios, where neither type of agents improve their education. We now discuss the payoff comparison in these regions:

- In region $N1$, agents play the strategies in case 1 in both the opaque and transparent scenarios.[9] Mathematically, $\Pi_{firm_{O1}} > \Pi_{firm_{T1}} = 0$. Therefore, the firm will always prefer to be opaque in this region.

- In region $N2$, agents play the strategies in case 2 in both the opaque and transparent scenarios. Since $\Pi_{firm_{O2}} = \Pi_{firm_{T2}}$, in this region, the firm is indifferent between being opaque or being transparent.

- In region $N3$, agents play the strategies in case 3 in both the opaque and transparent scenarios. We can rewrite $\Pi_{firm_{T3}}$ as follows:

$$\begin{aligned}\Pi_{firm_{T3}} &= \theta(\alpha+\beta)(\lambda+(1-\lambda)) + (1-\theta)\beta(\lambda+(1-\lambda)) \\ &\quad -R(\theta\lambda+(1-\theta)(1-\lambda)+(1-\theta)\lambda+(1-\lambda)\theta).\end{aligned}$$

[9]It is worth mentioning that although agents' strategies on the causal feature are the same in the opaque and transparent scenarios, the firm's payoff is different because of the existence of the correlational feature.

Since $\Pi_{firm_{O3}} = \lambda\theta(\alpha+\beta) + (1-\lambda)(1-\theta)\beta - (\lambda\theta + (1-\lambda)(1-\theta))R$, we have:

$$\begin{aligned}\Pi_{firm_{T3}} - \Pi_{firm_{O3}} &= (1-\lambda)\theta(\alpha+\beta-R) - \lambda(1-\theta)(R-\beta) \\ &= \theta(1-\lambda)\alpha - (\theta(1-\lambda) + (1-\theta)\lambda)(R-\beta) < 0.\end{aligned}$$

where the inequality follows Assumption 3. This means that the firm will always prefer to be opaque in this region.

We conclude that, in regions $N1$ through $N3$, being transparent is never strictly better than being opaque. This is quite intuitive since, in these regions, agents play the same strategies on the causal feature in both the opaque and transparent scenarios. Hence, the firm can only be worse off revealing its algorithm due to the loss in the predictive power of the correlational feature. Specifically, in regions $N1$ and $N3$, when the firm chooses to be opaque, the predictive power of the algorithm only comes from the correlational feature because the H type and L type agents play the same strategy on the causal feature. This suggests that the firm will incur a significant loss due to the reduction in the algorithm's prediction accuracy when the algorithm is made transparent.

We now discuss the payoff comparison in regions $C1$ through $C4$. In what follows, we first provide a summary of the payoff comparison result in each region, and then discuss the intuition.

- In region $C1$, the agents' strategies and the firm's payoff change from opaque scenario 1 to transparent scenario 3. $\Pi_{firm_{T3}} > \Pi_{firm_{O1}}$ iff

$$\beta > \beta_1 = \lambda\theta(\alpha - R) - (1-\lambda)(1-\theta)R + R - \theta\alpha. \tag{10}$$

  Thus, in this region, the firm will prefer to be transparent when $\beta > \beta_1$.

- In region $C2$, the agents' strategies and the firm's payoff change from opaque scenario 4 to transparent scenario 3. $\Pi_{firm_{T3}} > \Pi_{firm_{O4}}$ iff

$$\beta > \beta_2 = R - \frac{\alpha R(1-\lambda)}{(\alpha - R)\lambda + R(1-\lambda)}. \tag{11}$$

  Thus, in this region, the firm will prefer to be transparent when $\beta > \beta_2$.

- In region $C3$, the agents' strategies and the firm's payoff change from opaque scenario 5 to transparent scenario 3. $\Pi_{firm_{T3}} > \Pi_{firm_{O5}}$ iff

$$\beta > \beta_3 = \frac{\lambda\theta\alpha - \theta\alpha - 2\lambda\theta R + \theta R + \lambda R}{\lambda - 2\theta\lambda + \theta}. \tag{12}$$

  In this region, the firm will prefer to be transparent when $\beta > \beta_3$. It is interesting to note, however, that β_3 equals $\bar{\beta}$ defined in Assumption 3 (see below for an explanation of why this is the case). Since the value of β cannot exceed $\bar{\beta}$ (according to Assumption 3), this means that the firm will never prefer to be transparent in this region.

- In region $C4$, the agents' strategies and the firm's payoff change from opaque scenario 1 to transparent scenario 2. $\Pi_{firm_{T2}} > \Pi_{firm_{O1}}$ regardless of β. Thus, in this region, the firm will always prefer to be transparent.

According to Assumption 2, the value of α falls in the following interval:

$$\frac{(\theta\lambda + (1-\theta)(1-\lambda))R}{\theta\lambda} < \alpha < \frac{R}{\theta}.$$

Within the above interval, β_1 and β_3 are decreasing in α, and β_2 is increasing in α. Moreover, $\beta_1 = \beta_2$ when $\alpha = \underline{\alpha}$ and $\beta_2 = \beta_3$ when $\alpha = \bar{\alpha}$ ($\underline{\alpha}$ and $\bar{\alpha}$ are defined in Assumption 2). Thus, we have $\beta_1 < \beta_2 < \beta_3$. To understand why we have increasing thresholds for β as we move from regions $C1$ to $C3$ and why there is no threshold for β in region $C4$, we must look at how the firm's decision to be transparent changes the agents' strategies in different regions. To facilitate our discussions, we first define the concept of "degree of separation." Suppose that there are n_{H0} H type agents and n_{L0} L type agents who do not improve education and n_{H1} H type agent and n_{L1} L type agents who improve education. We define the degree of separation (Dos) between the H type and L type agents as follows:

$$Dos = 1 - \frac{\min(n_{H0}, n_{L0}) + \min(n_{H1}, n_{L1})}{n_{H0} + n_{L0} + n_{H1} + n_{L1}}.$$

Note that, if either all agents improve education or no one improves education, then Dos reaches its minimum value: $Dos_{min} = \max(\theta, 1-\theta)$. If all H type agents improve education and no L type agents improve education, then Dos reaches its maximum value: $Dos_{max} = 1$. If either H type or L type agents use a mixed strategy, the value of Dos is somewhere in between. The key observation here is that, the higher the value of Dos, the easier it is for the firm to differentiate H type agents from L type agents using the causal feature.

We now discuss how the firm's decision to be transparent changes agents' strategies in regions $C1$ through $C4$. Note that transparency intensifies agents' competition on the causal feature and that this intensified competition has the following two effects.

1. *The degree of separation between H type and L type agents on their causal feature changes.* As an illustration, consider region $C4$. In this region, when the firm switches from being opaque to being transparent, agents' strategies also switch from opaque scenario 1 to transparent scenario 2. Under opaque scenario 1, neither the H type nor the L type agents improve education, whereas under transparent scenario 2, only the H type agents improve education. This means that the H type agents are now more separated from the L type agents on the causal feature (i.e., there is a higher degree of separation). Similarly, it can also be verified that, in regions $C2$ and $C3$, the two types of agents become less separated and, in region $C1$, there is no change in the degree of separation.

2. *Agents' average value on the causal feature becomes higher and their work performance increases (according to Equation 2).* To see this, consider region $C1$. In this region, when the firm switches from being opaque to being transparent, agents' strategies also switch from opaque scenario 1 to transparent scenario 3. Although the change in agents' strategies does not affect the degree of separation, since both types of agents improve education, the average level of education for both agent types increases. Similarly, in region $C2$, the average level of education and, thus, the performance level of both types of agents increase. In region $C3$, only the average performance of the L type agents increases, whereas in region $C4$, only the average performance of the H type agents increases. We can see that, in regions $C1$ through $C4$, the agents' overall average performance always increases when the firm switches from being opaque to being transparent.

Since the firm always loses useful information from the correlational feature that helps it differentiate between the two types of agents when switching from the opaque to the transparent algorithm, the firm's decision on algorithmic transparency will depend on whether the above two effects (i.e., the change in the degree of separation and the increase in the agents' average performance) can offset the negative effect of information lost on the correlational feature. In region $C1$, even though the degree of separation does not change, both the H type and L type agents improve education, and the firm can benefit from the increase in the average agents' performance. Whether this benefit offsets the negative effect of the information loss on the correlational feature depends on the value of β. If β is large enough (i.e., $\beta > \beta_1$), then being transparent is preferred over being opaque. In region $C2$, the degree of separation decreases as the agents' distribution on the causal feature changes from partial separation to pooling. However, the average performance of the H type and L type agents increases, which suggests that, if β is large enough, the firm can still be better off making the algorithm transparent. The condition on β in this case is stricter than in region $C1$ (i.e., $\beta > \beta_2 > \beta_1$). This is so because the marginal effect of education must now be large enough to offset not only the previously mentioned negative effect of the loss of information on the correlational feature, but also the worse degree of separation on the causal feature.

In region $C3$, the firm's decision to be transparent affects fewer agents compared to in region $C2$. Switching from the opaque to the transparent algorithm incentivizes all L type agents and some H type agents to improve education in region $C2$, but it only incentivizes some L type agents to improve education in region $C3$. Since fewer agents are affected by the firm's switching from the opaque to the transparent algorithm in region $C3$ compared with region $C2$, and since the increase in the agents' average performance is proportional to the number of agents being affected, a larger β is needed in region 3 to achieve the same level of average performance found in region $C2$. This is why $\beta_3 > \beta_2$.

To see why $\beta_3 = \bar{\beta}$ (as defined in Assumption 3), note that in region $C3$ all H type agents have already improved education in the opaque scenario, and, only some L type agents will switch from not improving to improving education when the algorithm is made transparent. According to Assumption 1, the individual productivity of the L type agents cannot exceed R, which means that the firm will not benefit from hiring any L type agents even if their education levels are high. Consequently, making the algorithm transparent has a negative effect on the firm's payoff since the degree of separation on the causal feature decreases. This negative effect is partially offset by some L type agents' increased investment on the causal feature. The net effect is negative and its magnitude is decreasing in β. When β reaches $\bar{\beta}$, the net effect becomes 0. To see why, consider a range of β values that are very close to $\bar{\beta}$, according to Equation D.4 in Appendix D.1, nearly all L type agents have improved education in the opaque scenario; thus, the degree of separation is already near zero. Therefore, making the algorithm transparent has little negative effect on the degree of separation. Additionally, the positive effect of the increased investment on the causal feature is also close to zero. When $\beta = \bar{\beta}$, both effects are zero.

In region $C4$, the degree of separation increases when the firm switches from being opaque to being transparent. In fact, the agents' distribution on the causal feature changes from pooling to perfect separation. In other words, the firm now can perfectly separate the H type agents from the L type agents based on the causal feature alone without needing the correlational feature. This effect in itself is sufficient to offset the negative effect of information lost on the correlational feature. This is the reason why, in this region, the firm prefers to be transparent regardless of the value of β.

The following theorem summarizes our findings about the firm's decision on transparency:

Theorem 1 *In regions $N1$ through $N3$, being transparent is never strictly better than being opaque. In regions $C1$ through $C4$, depending on the value of β, the firm may prefer being transparent to being opaque. Specifically, in region $C1$, the firm will prefer to be transparent if $\beta > \beta_1$; in region $C2$, the firm will prefer to be transparent if $\beta > \beta_2$; in region $C3$, the firm will never prefer to be transparent; and, in region $C4$, the firm will prefer to be transparent regardless of the value of β.*

4.4 The Effect of the Predictive Power of the Correlational Feature (λ) and the Fraction of High-Talent Agents (θ)

We have shown that the firm will be strictly better off making the algorithm transparent if the (C_H, C_L) pair lands either in region $C4$, or in regions $C1$ or $C2$ with some additional conditions on β. In region $C4$, the transparent algorithm is preferred regardless of β, λ, and θ since removing the correlational feature changes the agents' behavior on the causal feature from pooling to perfect separation, and perfect separation is the case where the firm receives the maximum profit. In regions

$C1$ and $C2$, the main force that makes the transparent algorithm preferable is the agents' increased average level on the causal feature. As long as β exceeds the threshold $\beta_1(\beta_2)$, the increased level on the causal feature will have a large enough positive effect on work performance to make the firm better off. We now examine how the thresholds on β changes when λ or θ changes.

Taking the derivate of the expressions of β_1 and β_2 in Equations 10 and 11 with respect to λ yields the following:

$$\frac{\partial \beta_1}{\partial \lambda} = -2\theta R + \theta\alpha + R. \tag{13}$$

$$\frac{\partial \beta_2}{\partial \lambda} = \frac{\alpha R(\alpha - R)}{(2\lambda R - R - \alpha\lambda)^2}. \tag{14}$$

Both of these derivatives are greater than 0 in the parameter ranges that we consider (i.e., those given by Assumptions 1, 2, and 3). This means that, within each region, as λ becomes larger, a higher value of β is needed to make the transparent algorithm preferable. This is because a larger λ implies that more information is contained in the correlational feature, so a higher causal effect is needed to offset the loss of information from the correlational feature under the transparent algorithm. However, the effect of λ on algorithmic transparency is not this straightforward since, apart from the equilibrium payoff, λ can also determine the kind of strategy combination that can be sustained as an equilibrium given a (C_H,C_L) pair (i.e., the regions' shapes in Figure 3 will change as λ changes). Consider the region just beneath the dividing line of regions $N2$ and $C4$: as λ increases, it changes from belonging to region $N2$ to belonging to $C4$. This means that the firm will prefer the opaque algorithm when faced with a small λ but prefer a transparent algorithm when faced with a large λ. Although it appears counter-intuitive, it can be explained as follows. When λ is small, making the algorithm transparent is not effective enough to change the agents' behavior on the causal feature. However, when λ is large, the agents' investment on the causal feature will increase drastically, and the firm is able to hire agents who are on average more productive under the transparent scenario than under the opaque scenario.

The following proposition summarizes our findings about how λ affects the firm's decision on transparency:

Proposition 1 *An increase in λ has the following effects on the firm's decision on transparency:*

1. *The area of regions $C1$, $C2$, and $C4$ increases, which means that the transparent algorithm is preferred under more (C_H,C_L) value pairs.*

2. *Within regions $C1$ and $C2$, the conditions on β to make the transparent algorithm preferred to the opaque algorithm become stricter (i.e. a larger β is needed).*

We now discuss the impact of θ on algorithmic transparency. Taking the derivative of the expressions of β_1 and β_2 in Equations 10 and 11 with respect to θ yields:

$$\frac{\partial \beta_1}{\partial \theta} = -2\lambda R + \lambda\alpha + R - \alpha. \quad (15)$$

$$\frac{\partial \beta_2}{\partial \theta} = 0. \quad (16)$$

It can be shown that $\frac{\partial \beta_1}{\partial \theta}$ is smaller than 0 in the parameter ranges that we consider. This means that, in region $C1$, as the proportion of the H type agents increases, the conditions on β to make algorithmic transparency more desirable become milder. This is because the degree of separation on the causal feature in region $C1$ does not change (from pooling at 0 to pooling at 1). The only negative effect of algorithmic transparency is the loss of the correlational feature. When θ is high, most agents are of H type, and losing the correlational feature is less harmful to the firm (because there are fewer L type agents who can be mistakenly hired when the correlational feature is lost). Thus, a smaller β is needed to offset this negative effect. In region $C2$, however, θ has no influence on the firm's decision on algorithmic transparency. This is because, in region $C2$, there are two negative effects of algorithmic transparency: the loss of the correlational feature and a smaller degree of separation on the causal feature. A higher θ will mitigate the first effect (as explained above) and amplify the second (more H type agents will be left out). Overall, θ does not affect the value of β needed for making the transparent algorithm preferable.

The following proposition summarizes our findings about how θ affects the firm's decision on transparency:

Proposition 2 *In region $C1$, a higher θ will increase the firm's incentive to make the algorithm transparent. In other regions, θ has no impact on the firm's decision on algorithmic transparency.*

4.5 Agents' Welfare

In Section 4.3, we have specified conditions under which the transparent algorithm will yield a strictly higher payoff to the firm than the opaque algorithm. We will next investigate the impact of algorithmic transparency on the agents' welfare.

The total payoff across all agents in the equilibrium is summarized in the following lemma.

Lemma 3 *For each equilibrium outcome shown in Figure 1 (for the opaque scenario) and Figure*

2 (for the transparent scenario), the corresponding total payoff for all agents is given by

$$\Pi_{agents_{O1}} = (\lambda\theta + (1-\lambda)(1-\theta))R$$
$$\Pi_{agents_{O2}} = \theta(R - C_H)$$
$$\Pi_{agents_{O3}} = (\lambda\theta + (1-\lambda)(1-\theta))R - C_H\theta - C_L(1-\theta)$$
$$\Pi_{agents_{O4}} = \frac{(\lambda\theta + (1-\lambda)(1-\theta))(R - C_H)}{\lambda}$$
$$\Pi_{agents_{O5}} = \frac{(2R\lambda - R - C_H\lambda - C_L\lambda + C_L)\theta}{\lambda}$$

$$\Pi_{agents_{T1}} = 0$$
$$\Pi_{agents_{T2}} = \theta(R - C_H)$$
$$\Pi_{agents_{T3}} = R - C_H\theta - C_L(1-\theta),$$

where $\Pi_{agents_{Oi}}$ *denotes the agents' total payoff in case* i *of the opaque scenario and* $\Pi_{agents_{Ti}}$ *denotes the agents' total payoff in case* i *of the transparent scenario.*

As previously discussed, Figure 3 shows how the agents' behavior changes on the causal feature when the algorithm is made transparent. First, we consider the three regions where the agents' behavior on the causal feature does not change (regions $N1$, $N2$, and $N3$). In directly comparing the agents' payoff in the equilibrium, we have the following observations:

- In region $N1$, the agents play the strategies in case 1 in both the opaque and transparent scenarios. $\Pi_{agents_{O1}} > \Pi_{agents_{T1}}$. Therefore, the agents will receive a higher total payoff under the opaque algorithm in this region.
- In region $N2$, the agents play the strategies in case 2 in both the opaque and transparent scenarios. Since $\Pi_{agents_{O2}} = \Pi_{agents_{T2}}$, in this region, the agents are indifferent to whether the algorithm is opaque or transparent.
- In region $N3$, the agents play the strategies in case 3 in both the opaque and transparent scenarios. $\Pi_{agents_{O3}} < \Pi_{agents_{T3}}$. Therefore, the agents will receive a higher payoff under the transparent algorithm in this region.

In regions $N1$, $N2$, and $N3$, the agents' behavior on the causal feature does not change. Since improving the correlational feature is assumed to be costless, the agents' total cost stays the same before and after the algorithm is made transparent. The only thing that varies is the benefit they can obtain under the firm's hiring strategy. In region $N1$, more agents will be hired under the opaque algorithm (a λ portion of the H type agents and a $1-\lambda$ portion of the L type agents are

hired under the opaque algorithm but no one will be hired under the transparent algorithm). In region $N3$, more agents will be hired under the transparent algorithm (a λ portion of the H type agents and a $1-\lambda$ portion of the L type agents are hired under the opaque algorithm and everyone will be hired under the transparent algorithm). In region $N2$, the same number of agents will be hired regardless of whether the algorithm is opaque or transparent (only the H type agents will be hired).

Next, we consider the four regions where agents' behavior on the causal feature changes after the algorithm is made transparent (regions $C1$, $C2$, $C3$, and $C4$). By directly comparing the agents' payoffs in the equilibrium, we obtain the following observations:

- In region $C1$, the agents' strategies and their total payoff change from opaque scenario 1 to transparent scenario 3. $\Pi_{agents_{T3}} \leq \Pi_{agents_{O1}}$ iff

$$C_H\theta + C_L(1-\theta) \geq (1-\lambda\theta-(1-\lambda)(1-\theta))R. \tag{17}$$

  The smallest possible value for the left-hand side (LHS) of the inequality is reached when a (C_H,C_L) pair lands at the lower left corner in region $C1$, or in other words, when $C_H = (1-\lambda)R$ and $C_L = \lambda R$. It can further be shown that this smallest value equals the right-hand side (RHS). Thus, Equation 17 is satisfied for any (C_H,C_L) pair in region $C1$. In this region, the transparent algorithm will give the agents a lower total payoff compared with the opaque algorithm.

- In region $C2$, the agents' strategies and their total payoff change from opaque scenario 4 to transparent scenario 3. $\Pi_{agents_{T3}} \geq \Pi_{agents_{O4}}$ iff

$$\frac{1-\lambda}{\lambda}C_H - C_L \geq (\frac{1}{\lambda}-2)R. \tag{18}$$

  The smallest possible value for the LHS of the inequality is reached when a (C_H,C_L) pair lands at the lower right corner in region $C2$, or in other words, when $C_H = (1-\lambda)R$ and $C_L = \lambda R$. It can further be shown that this smallest value equals the RHS. Thus, Equation 18 is satisfied for any (C_H,C_L) pair in region $C2$. In this region, the transparent algorithm will give agents a higher payoff compared with the opaque algorithm.

- In region $C3$, the agents' strategies and their total payoff change from opaque scenario 5 to transparent scenario 3. $\Pi_{agents_{T3}} \geq \Pi_{agents_{O5}}$ iff

$$(2\theta - \frac{\theta}{\lambda} - 1)C_L \geq (2\theta - \frac{\theta}{\lambda} - 1)R. \tag{19}$$

  Since $2\theta - \frac{\theta}{\lambda} - 1 \leq 0$, and $C_L \leq R$ in region $C3$, Equation 19 is satisfied for any (C_H,C_L) pair in region $C3$. In this region, the transparent algorithm will give agents a higher payoff compared with the opaque algorithm.

- In region $C4$, the agents' strategies and their total payoff change from opaque scenario 1 to transparent scenario 2. $\Pi_{agents_{T2}} \leq \Pi_{agents_{O1}}$ iff

$$\theta(R - C_H) \leq (\lambda\theta + (1-\lambda)(1-\theta))R. \tag{20}$$

The largest possible value for the LHS is reached when a (C_H,C_L) pair lands at the lower bound of region $C4$, or in other words, when $C_H = (1-\lambda)R$. It can be shown that this largest value is smaller than the RHS. Thus, Equation 20 is satisfied for any (C_H,C_L) pair in region $C4$. In this region, the transparent algorithm will give agents a lower payoff compared with the opaque algorithm.

The following theorem summarizes our findings regarding the agents' welfare under the opaque and transparent algorithms.

Theorem 2 *Whether the agents are better off under either the opaque or the transparent algorithm depends on the values of (C_H,C_L), and this dependence is shown in Figure 4. In regions $N3$, $C2$ and $C3$, the agents' welfare is higher under the transparent algorithm. In regions $N1$, $C1$ and $C4$, the agents' welfare is higher under the opaque algorithm. In region $N2$, the agents' welfare is not affected by algorithmic transparency.*

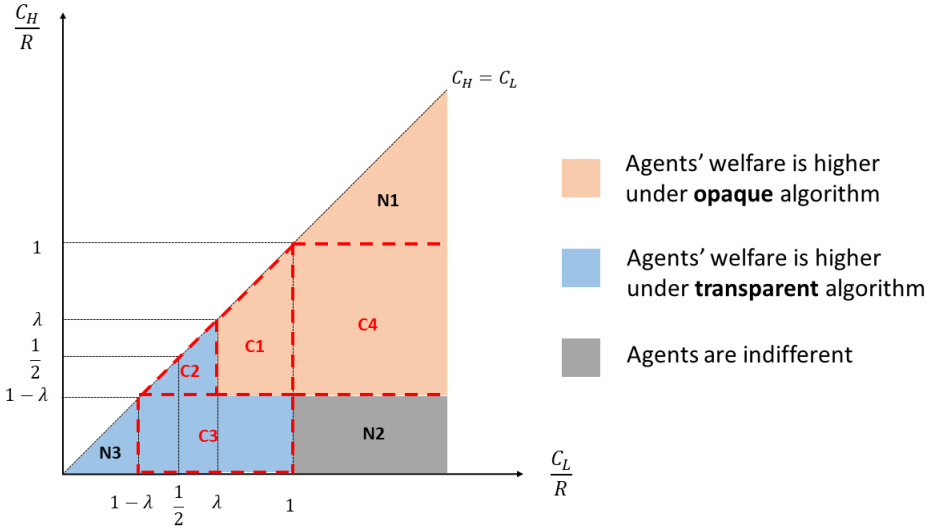


Figure 4: Comparison of agents' welfare in the transparent and opaque scenarios

The intuition behind Theorem 2 can be explained as follows. The pattern here is that the agents as a whole will prefer the opaque algorithm when C_H and C_L are large. They will prefer the transparent algorithm when C_H and C_L are small, and will be indifferent when C_L is large

but C_H is small. Making the algorithm transparent may force the agents to invest in the causal feature. Of course, they can also benefit from the increased investment in the causal feature (i.e., a higher chance of being hired). The cost of this investment increases with C_H and C_L, but the benefit does not vary with C_H and C_L. Consequently, the agents will be worse (better) off under the transparent algorithm if C_H and C_L are large (small).

Comparing Theorem 1 with Theorem 2, we find that the firm's and agents' interests conflict in some regions. For example, in regions $N3$ and $C3$, the transparent algorithm will give the agents a higher payoff, but the firm prefers the opaque algorithm. In region $C4$, the opaque algorithm will give the agents a higher payoff but the firm prefers the transparent algorithm. In regions $C1$ and $C2$, whether there is a conflict of interest between the firm and the agents depends on the value of β.

In region $C2$, if the condition of $\beta > \beta_2$ is satisfied, both the firm and the agents would prefer the transparent algorithm over the opaque algorithm. In other words, making the algorithm transparent will Pareto dominate keeping the algorithm opaque.

5 The Stackelberg Model

Our analysis in Section 4 has focused on the setting where the firm does not have commitment power. As noted in Section 1, this is useful to highlight the insight that the firm could still prefer the transparent algorithm to the opaque algorithm even in the absence of the 'first mover advantage'. In this section, we consider the case of 'full transparency' where the firm publishes all details of the hiring algorithm, including the features being used and the hiring strategy, and the firm has commitment power on the published algorithm. In this case, the Stackelberg model would be a more appropriate model when analyzing the transparent scenario. Our main objective in this section is to show that the key insights in the previous section will only be further strengthened when the firm switches from using partial transparency to using full transparency.

We start with discussing the firm's equilibrium payoffs. The exact payoffs for all cases are summarized in Lemma 4.

Lemma 4 *The equilibrium outcome depends on the values of (C_H, C_L), and this dependence is shown in Figure 5. The corresponding total payoffs for the firm are given by*

$$\begin{aligned}
\Pi_{firm_{FT1}} &= 0 \\
\Pi_{firm_{FT2}} &= \theta(\alpha + \beta - R) \\
\Pi_{firm_{FT3}} &= \theta(\alpha + \beta) + (1 - \theta)\beta - R \\
\Pi_{firm_{FTS}} &= \frac{C_L \theta(\alpha + \beta - R)}{R},
\end{aligned}$$

where $\Pi_{firm_{FTi}}$ *denotes the firm's total payoff in case* i, $i \in \{1, 2, 3, S\}$.

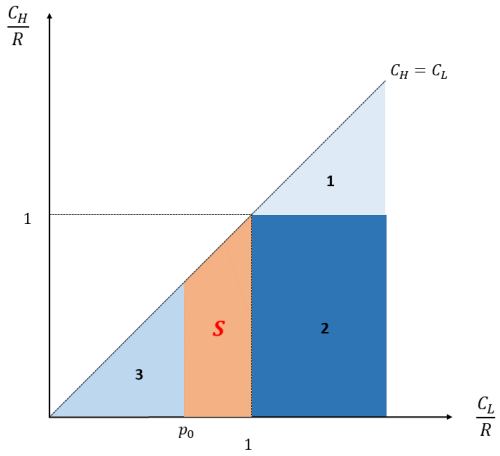


Figure 5: Equilibrium outcome in the transparent scenario in the Stackelberg model

Compared with the payoffs discussed in Lemma 2, we can see that the firm always gets a *weakly* higher payoff in the full transparency scenario than in the partial transparency scenario. This is so because, in the full transparency scenario, the firm can always commit to the equilibrium hiring strategy in the partial transparency scenario and achieve the same payoff as in the partial transparency scenario, which sets a lower bound on the firm's payoff. Given that the firm's payoff is weakly higher in the full transparency scenario than in the partial transparency scenario, naturally, the conditions under which full transparency is preferred over opacity are less strict than the conditions under which partial transparency is preferred over opacity. In particular, it can be shown that there are conditions under which the firm prefers full transparency, but not partial transparency, over opacity. These conditions are summarized in Proposition 3.

Proposition 3 *The firm prefers full transparency but not partial transparency over opacity if and only if any of the following conditions is satisfied:*

- *Value pair* $(\frac{C_L}{R}, \frac{C_H}{R})$ *falls in region* $C1$ *and* $\beta \leq \beta_1$*;*
- *Value pair* $(\frac{C_L}{R}, \frac{C_H}{R})$ *falls in region* $C2$ *and* $\beta \leq \beta_2$ *and* $C_L > C_L^{C2}$*;*
- *Value pair* $(\frac{C_L}{R}, \frac{C_H}{R})$ *falls in region* $C3$ *and* $C_L > C_L^{C3}$*;*
- *Value pair* $(\frac{C_L}{R}, \frac{C_H}{R})$ *falls in region* $N3$ *and* $C_L > C_L^{N3}$*;*

where $C_L^{C2} = R - \frac{R^2(1-\theta)(1-\lambda)}{(\alpha-R)\theta\lambda}$, $C_L^{C3} = \frac{(2\lambda-1)R\theta}{\lambda}$, $C_L^{N3} = R\lambda\theta + \frac{R(1-\lambda)(1-\theta)(\beta-R)}{\theta(\alpha+\beta-R)}$.

In Section 4.4, we discussed how the values of θ and λ affect the firm's preference on algorithmic transparency under partial transparency. Most of the results in Section 4.4 will not change qualitatively when full transparency is considered. The detailed discussions and analysis can be found in Appendix C.3. In what follows, we briefly discuss the key findings:

1. In general, when λ is large, the conditions under which the transparent algorithm is preferred over the opaque algorithm become stricter. However, there are also (C_H, C_L) pairs under which the firm switches from never preferring transparency to possibly preferring transparency (depending on the values of the other parameters) when λ gets larger. This finding is consistent with that under the partial transparency model.

2. In general, when θ increases, there will be more (C_H, C_L) pairs under which the firm will prefer the transparent algorithm over the opaque algorithm. This finding is unique under full transparency since, under partial transparency, θ does not affect the firm's decision on algorithmic transparency in most regions. This difference is due to the fact that in some regions, the firm can get a higher payoff under full transparency than under partial transparency, and the payoff difference is increasing in θ. In other words, when θ is large, the firm's payoff in the full transparency scenario will increase by a larger amount compared to the partial transparency scenario, which gives the firm more incentive to make the algorithm transparent.

Next, we discuss how agents' welfare is affected by full transparency. Lemma 5 summarizes the agents' equilibrium payoff in the full transparency scenario.

Lemma 5 *In the full transparency scenario, for each equilibrium outcome shown in Figure 5 the corresponding total payoff for agents is given by*

$$\begin{aligned}
\Pi_{agents_{FT1}} &= 0 \\
\Pi_{agents_{FT2}} &= \theta(R - C_H) \\
\Pi_{agents_{FT3}} &= R - C_H\theta - C_L(1-\theta) \\
\Pi_{agents_{FTS}} &= \theta(C_L - C_H),
\end{aligned}$$

where $\Pi_{agents_{FTi}}$ denotes the agents' total payoff in case i of the full transparency scenario.

Comparing the agents' payoff in the full transparency scenario with that in the partial transparency scenario, we can see that agents' welfare becomes strictly lower in region S and stays the same in all other regions. The finding is interesting as it shows that, as the level of transparency in the algorithm increases (i.e., from partial to full transparency), the firm may become better

off while the agents may become worse off. It strengthens our conclusion in Section 4.5 that, in most cases when the firm prefers transparency, the agents will be worse off under the transparent algorithm. The results presented in this section are consistent with recent research in both economics and computer science, which shows that compared to the case where the decision maker does not have commitment power, under the Stackelberg competition where the decision maker has commitment power, the decision maker can receive a higher payoff but at a cost of decreasing agents' welfare (Frankel and Kartik, 2019, Milli et al., 2019).

6 Extensions

In this section, we analyze several model extensions that relax some of the assumptions in our main model described in Section 3. Due to space limitation, we focus our discussion on the impact of algorithmic transparency on the firm, and briefly report the key findings here. The detailed discussions and proofs can be found in Appendix B. Overall, the results show that relaxing those assumptions does not alter the main results and insights of the study.

First, we relax the assumption that the cost of improving the correlational feature is close to zero. We consider the case where the cost of improving the correlation feature is substantial: c_h for the H type agents and c_l for the L type agents where $c_l \geq c_h > 0$. We find that under this condition, the firm's equilibrium payoff in the transparent scenario weakly increases while the firm's equilibrium payoff in the opaque scenario is not affected. Thus the main result of the paper, under certain conditions, algorithmic transparency benefits the firm, is strengthened. Second, we address the fixed wage assumption by allowing the firm to strategically choose the wage. After solving this extended model, we find that the firm in the transparent scenario is able to set a lower wage than in the opaque scenario without needing to worry about worsening the degree of separation between the two types of agents on the causal feature. Since the agents' productivity (affected by α and β) is assumed to be unaffected by wage, the firm will benefit more from endogenizing the wage in the transparent scenario than in the opaque scenario; thus, the firm's preference for the transparent algorithm will be strengthened. Lastly, we show that even when the agents severely underestimate or overestimate the prediction power of the correlational feature, λ, the regions in which the firm prefers the transparent algorithm will not vanish.

7 Conclusion

7.1 Summary of Results

In this paper, we studied how firm and agent welfare is affected by algorithmic transparency. We allowed the agents to be strategic such that they can invest in their causal and correlational features to increase their chances of being hired in response to the firm's algorithm. We also investigated how the predictive power of the correlational feature, the market composition in terms of the fraction of H type agents, and the impact of the causal feature on agent productivity affect the firm's decision to make their algorithm transparent or opaque.

As a first result, we identified a broad set of conditions under which the firm would be better off with algorithmic transparency than opacity. Our second result is that the agents may not always be better off under algorithmic transparency. Our third result is that, even when the correlational feature has high predictive power in the opaque scenario, the firm could still be better off making the algorithm transparent. Our final result is that, when the fraction of H type agents on the market is high, the firm would be better off by making its algorithm transparent. We also provided several extensions to our main model. After relaxing several assumptions and considering several model alternatives, we found that the main insights of the paper do not change.

7.2 Implications for Managers

Our paper shows that managers using machine learning models for decision making could be better off by making their algorithms transparent. Algorithmic transparency does not always mean a loss of predictive power. In some cases, it can in fact lead to greater predictive power. In other cases, while it may reduce predictive power, it can still make managers better off by improving the desirability of the whole market. Our results are particularly promising for managers, as they are now facing growing calls to make their algorithms transparent.

We identified a set of conditions where managers should prefer algorithmic transparency. There are three factors that managers should consider: (a) access to a good set of causal features; (b) the predictive power of the correlational features; (c) the market composition in terms of the fraction of H type (desirable) agents.

We provide some guidance on what makes a good causal feature here. The causal feature serves two purposes: (a) a signaling purpose – H type agents have a cost advantage on this feature, and thus, it can help separate H type agents from L type agents; and (b) a human capital purpose - the feature itself contributes to the productivity of the agents. To serve the signaling purpose well, the identified causal feature should be neither too costly nor too cheap to improve. If it is too costly, no one will improve it. By contrast, if it is too cheap, everyone will improve it. Furthermore, in

terms of the human capital purpose, the higher the feature's impact on productivity, the better it is. Even when this causal feature is unable to completely separate the H type agents from the L type agents, if it is moderately costly and contributes to productivity, the firm could still be better off. Typically algorithm designers are not focused on causality or identifying causal features. Our results indicate that they should. The recent stream of research in computer science that examines causal inference in machine learning models bodes well for them in this regard.

The second factor that managers should consider is the predictive power of the correlated features. Intuitively, managers may think of keeping their algorithms opaque when the correlational features provide significant predictive power. Our results show that this thinking is incorrect. We show that firms are more likely to be better off by making their algorithms transparent when correlational features provide significant predictive power in the opaque counterpart. Though incorrectly, managers would be particularly concerned about making the algorithms transparent when the marginal effect of the causal feature in separating the H type agents from the L type agents is small in the presence of correlational features in an opaque algorithm. However, they should realize that the effect of the causal feature is suppressed largely due to the strategic behavior of the agents when the algorithm is opaque. All agents, but more importantly, the H type agents, underinvest in the causal feature when they know the correlational feature can separate them from the L type agents: the higher the predictive power of the correlational feature, the lower the incentive of the agents to invest in the causal feature. When the algorithm is made transparent, in the new equilibrium, the agents have a higher incentive to invest in the causal feature, making the firm better off.

The third factor managers should consider is the market composition in terms of the fraction of the H type agents. If the causal feature is unable to separate the H type agents from the L type agents, algorithmic transparency could still benefit the firm if the cost of improving the causal feature is moderate and the market is composed of more H type agents.

Overall, our results suggest that managers should not view manipulation by agents as bad. Rather, they should embrace it and use algorithmic transparency as a lever for motivating agents to invest in more desirable actions.

7.3 Implications for Public Policy

There are two key arguments typically put forth in support of algorithmic transparency. First, making algorithms transparent could highlight any hidden biases in algorithms and make them accountable (Citron and Pasquale, 2014). Second, the users who are affected by an algorithm's decision making have the right to see which factors affect decisions made about them. Our paper has implications related to this second argument.

Recent legislation like the General Data Protection Regulation has afforded individuals a right to explanation under which firms have to provide an explanation regarding how a decision has been made by their algorithms (Goodman and Flaxman, 2017). Our results show that such a regulation may not improve consumer welfare. When agents know which features in an algorithm affect important decisions about them, they can improve those features. Consequently, algorithmic transparency is generally viewed as helping the agents at the cost of the firm. However, our results show that the agents may not benefit from algorithmic transparency as expected. When algorithms are opaque, they derive their accuracy from both causal and correlational features. Therefore, the H type agents do not need to invest in the costly causal feature to separate themselves from the L type agents. In the transparent scenarios, in many cases, the agents have to invest in the costly causal feature to achieve similar or even less separation with no change in their wage.

7.4 Generalizability of the Results

While our model focuses on the job hiring scenario which is an example of a screening problem, our model and results are generalizable to other screening problems. Specifically, problems comprising of two types of economic agents who have asymmetric information and who are attempting to engage in a transaction. The "screener" (i.e. agent with less information, e.g. the firm in our case) attempts to gain further insight or knowledge into the private information of the other agents (i.e. hidden type of the agent in our case). ML/AI algorithms are proving useful in helping the firms "screen" the applicants better. Job hiring is only one example of a screening problem. Other example include, insurance markets (e.g. car insurance, health insurance) or credit lending markets.

7.5 Future Research Directions

Firms have typically kept their algorithms opaque to protect them from gaming by agents. In this study, we show that this strategy may not be the best, and the firm could be better off by making its algorithm transparent in the presence of strategic users. However, there are three other reasons as to why firms may still not want to make their algorithm transparent – interpretability, privacy and competition. With access to big data and large computational power, machine learning models have become complicated to the extent that they are rendered uninterpretable. While recent research has made advances in developing interpretable machine learning models, Bertsimas et al. (2019) show that model interpretability comes at a cost of accuracy. As a result, when considering the issue of algorithmic transparency, it may be interesting to consider the tradeoff between interpretability and accuracy. Similarly, when a firm makes its algorithm transparent, this can lead to privacy concerns. Others may be able to infer information about agents when they are selected by a transparent algorithm. In these cases, algorithmic transparency may impose a privacy

cost on agents. Future research can investigate algorithmic transparency in the presence of privacy concerns. Finally, it would be interesting to investigate whether and how algorithmic transparency may affect the intensity of competition among firms. We believe these are interesting avenues for future research.

References

T. Alon, M. R. C. Dobson, A. D. Procaccia, I. Talgam-Cohen, and J. Tucker-Foltz. Multiagent evaluation mechanisms. In *AAAI*, 2020.

C. Alós-Ferrer and J. Prat. Job market signaling and employer learning. *Journal of Economic Theory*, 147(5):1787 – 1817, 2012.

D. Bertsimas, A. Delarue, P. Jaillet, and S. Martin. The price of interpretability. *arXiv preprint arXiv:1907.03419*, 2019.

A. Bonatti and G. Cisternas. Consumer Scores and Price Discrimination. *The Review of Economic Studies*, 87(2):750–791, 2019.

I.-K. Cho and D. M. Kreps. Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221, 1987.

D. K. Citron and F. A. Pasquale. The scored society: Due process for automated predictions. *Washington Law Review*, 89, 2014.

V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, pages 1431–1451, 1982.

R. Cummings, S. Ioannidis, and K. Ligett. Truthful linear regression. In *Conference on Learning Theory*, pages 448–483, 2015.

B. Daley and B. Green. Market signaling with grades. *Journal of Economic Theory*, 151:114 – 145, 2014.

M. Engers. Signalling with many signals. *Econometrica*, 55(3):663–674, 1987.

A. Frankel and N. Kartik. Improving information from manipulable data. *arXiv preprint arXiv:1908.10330*, 2019.

R. Fu, Y. Huang, and P. V. Singh. Crowd, lending, machine, and bias. *Available at SSRN 3206027*, 2019.

C. A. Goodhart. Problems of monetary management: the uk experience. In *Monetary Theory and Practice*, pages 91–121. Springer, 1984.

B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

N. Haghtalab, N. Immorlica, B. Lucier, and J. Wang. Maximizing welfare with incentive-aware evaluation mechanisms. In *29th International Joint Conference on Artificial Intelligence*, 2020.

M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

J. Kleinberg and M. Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844, 2019.

R. E. Lucas et al. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46, 1976.

G. Mailath, M. Okuno-Fujiwara, and A. Postlewaite. Belief-based refinements in signalling games. *Journal of Economic Theory*, 60(2):241–276, 1993.

R. Meir, A. Procaccia, and J. Rosenschein. Algorithms for strategyproof classification. *Journal of Artificial Intelligence*, 186:123–156, 07 2012.

S. Milli, J. Miller, A. D. Dragan, and M. Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

H. Schellmann and J. Bellini. Artificial intelligence: The robots are now hiring. *The Wall Street Journal*, 20, 2018.

W. Schmidt and R. W. Buell. Experimental evidence of pooling outcomes under information asymmetry. *Management Science*, 63(5):1586–1605, 2017. doi: 10.1287/mnsc.2015.2407.

D. Segal. The dirty little secrets of search. *The New York Times*, 12(02), 2011.

M. Spence. Job market signaling. *Quarterly Journal of Economics*, 87:355–374, 1973.

A. Weiss. A sorting-cum-learning model of education. *Journal of Political Economy*, 91(3):420–442, 1983.

A Notation Summary

Notation	Meaning
α	The marginal effect of type on performance
β	The marginal effect of the causal feature on performance
λ	The portion of H type agents who are high on the correlational feature
θ	The portion of H type agents in the population
C_H	H type agents' cost of improving on the causal feature
C_L	L type agents' cost of improving on the causal feature
R	The job compensation paid by the firm once an agent is hired
State A	Low on the causal and high on the correlational feature in the opaque scenario
State B	High on the causal and high on the correlational feature in the opaque scenario
State C	Low on the causal and low on the correlational feature in the opaque scenario
State D	High on the causal and low on the correlational feature in the opaque scenario
State E	Low on the causal feature in the transparent scenario
State F	High on the causal feature in the transparent scenario

Table 1: Notation summary

B Discussions of Extensions of the Main Model

In this section, we analyze several extensions by relaxing some of the assumptions in our main model in Section 3. Due to space limitation, we will focus our discussions on the impact of algorithmic transparency on the firm, and the main goal is to show that relaxing some of the assumptions does not alter the main results and insights of the study. Specifically, in Appendix B.1, we relax the assumption that the cost of improving the correlational feature is zero; in Appendix B.2, we relax the assumption that the wage is fixed and exogenously given; in Appendix B.3, we consider the case where agents have incorrect belief of λ. Unless otherwise noted, all the proofs of the results in this section can be found in Appendix D.

B.1 Gaming is Costly

In the main model, we have assumed that the cost of improving the correlation feature (gaming) is minimal regardless of the true type of the agent (we refer to this as the 'costless' setting). We now consider the case where the cost of improving the correlational feature is c_h for H type agents and c_l for L type agents with $c_l \geq c_h > 0$ (we refer to this as the 'costly' setting). This captures

the situation where the correlational feature is not easy to game. For example, a Ph.D. program admission office might use the number of international conferences an applicant has attended in the past as a correlational feature, or a firm may use AI to analyze applicants' behavior (e.g., pace or pitch of speaking) in the interview and use these verbal clues as correlational features. All these features require a certain amount of effort to game.

In the following, we first discuss the case where $c_h = c_l = c$ (we call this the 'no-advantage' case) and then we discuss the case where $c_h < c_l$ (we call this the 'advantage' case). Note that the change in the cost of improving the correlational feature has no impact on our analysis for the opaque scenario since, in the opaque scenario, the agents cannot manipulate their values on the correlational feature. As for the transparent scenario, we will show that the firm will get a weakly higher payoff in the 'costly' setting than in the 'costless' setting.

Recall that, in the 'costless' setting, all agents will eventually have value 1 on the correlational feature under the transparent algorithm; thus, States A and C collapse into a single State E whereas States B and D collapse into a single State F according to the discussions in Section 3.2. In the 'costly' setting, some agents may decide not to improve their correlational feature even when the algorithm is transparent; therefore, agents' possible ending states are A, B, C, and D as in the opaque scenario.

The firm's equilibrium payoffs for the 'no-advantage' case are given below. Note that we divide the regions in the same way for Lemmas 6 and 7 below. However, the firm's payoff in each region is different depending on the relative size of c with respect to R.

Lemma 6 *In the 'costly' setting, when $c \geq R$, the equilibrium outcome depends on the values of (C_H, C_L), and this dependence is shown in Figure 6. The corresponding total payoffs for the firm are given by*

$$\Pi_{1a}^{ex1} = \lambda\theta(\alpha - R) - (1-\lambda)(1-\theta)R$$
$$\Pi_{2a}^{ex1} = \theta(\alpha + \beta - R)$$
$$\Pi_{3a}^{ex1} = \Pi_{3b}^{ex1} = \Pi_{3c}^{ex1} = \lambda\theta(\alpha + \beta - R) + (1-\lambda)(1-\theta)(\beta - R),$$

where Π_i^{ex1} denotes the firm's total payoff in region i. All payoffs are non-negative.

Lemma 7 *In the 'costly' setting, when $c < R$, the equilibrium outcome depends on the values of (C_H, C_L), and this dependence is shown in Figure 6. The corresponding total payoffs for the firm*

are given by

$$\begin{aligned}
\Pi_{1a}^{ex2} &= 0 \\
\Pi_{2a}^{ex2} &= \theta(\alpha + \beta - R) \\
\Pi_{3a}^{ex2} &= \theta(\alpha + \beta - R) + (1 - \theta)(\beta - R) \\
\Pi_{3b}^{ex2} &= \theta(\alpha + \beta - R) + (1 - \lambda)(1 - \theta)(\beta - R) \\
\Pi_{3c}^{ex2} &= \lambda\theta(\alpha + \beta - R) + (1 - \lambda)(1 - \theta)(\beta - R),
\end{aligned}$$

where Π_i^{ex2} *denotes the firm's total payoff in region* i*. All payoffs are non-negative.*

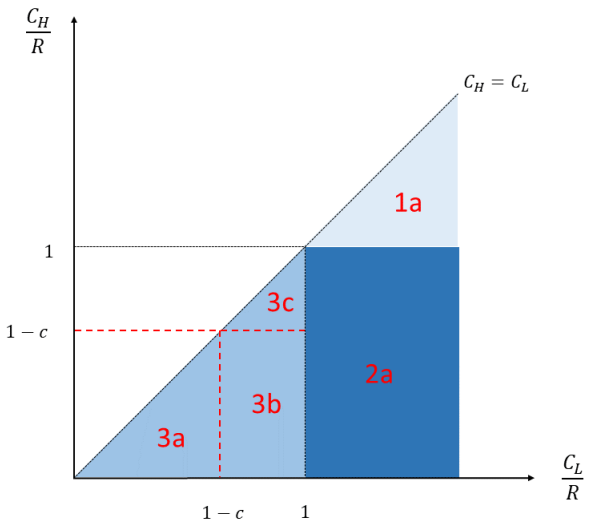


Figure 6: Equilibrium outcome in the 'costly' setting

By directly comparing the firm's total payoffs in Lemmas 6 and 7 with those in Lemma 2, it is not difficult to see that the firm is weakly better off in the 'costly' setting than in 'costless' setting. We formally summarize this observation in Proposition 4.

Proposition 4 *For the 'no-advantage' case, the firm gets weakly better payoffs in the 'costly' setting than in the 'costless' setting regardless of* C_H*,* C_L*, and* c*.*

In general, as c increases, the conditions under which algorithmic transparency is preferred become less strict (note that, for each region, the payoff in the 'costly' setting is at least as large as the payoff in the 'costless' setting). The intuition is as follows: If c is substantial, the firm might be able to exploit the information on the correlational feature without worrying about agents gaming the feature even when the algorithm is transparent. To be more specific, the firm can treat the

agents in states {A, B} and states {C, D} differently without worrying that the agents in states {C, D} will move to states {A, B}, reducing the degree of separation. Since the average productivity of agents in states {A, B} is higher than the average productivity of the whole population, the firm can always get a higher payoff in the 'costly' setting than in the 'costless' setting by hiring a larger number of agents in states {A, B} than in states {C, D}.

We next study the 'advantage' case, where the cost of improving on correlational feature for H type agents stays at $c_h = c$ while the cost for L type agents increases to $c_l > c$. Proposition 5 summarizes the main findings:

Proposition 5 *The firm gets weakly better payoffs in the 'advantage' setting than in the 'no advantage' setting regardless of C_H, C_L, c_h and c_l.*

The intuition of the above result is straightforward: Increasing c_l could discourage L type agents from gaming when the algorithm is transparent, and this is always beneficial to the firm since agents in states A and B are on average more productive than agents in states C and D, and the firm always hire more agents from the former states (i.e. $P_A \geq P_C, P_B \geq P_D$). Discouraging L type agents from gaming will further increase the difference in agents' quality (i.e., the proportion of H type agents) between states {A, B} and states {C, D} which, in turn, helps the firm to hire more H type agents while avoiding more L type agents.

To summarize, when the cost of improving the correlational feature is not zero, it can only strengthen the firm's incentive to choose a transparent algorithm for any value of C_H and C_L.

B.2 Endogenous Wage

In the main model, we have shown that, when the wage R is fixed, the firm will be better off in making the algorithm transparent under certain conditions. A natural question is whether this result is primarily driven by the fixed wage assumption, that is, whether the result is driven by the fact that agents' overall education and productivity increase in the transparent scenario and the firm can take full advantage of this increase without incurring any additional cost. If, in the opaque scenario, the firm can lower the wage while still keeping the agents' strategies unchanged, it can offset some of the firm's loss due to agents' lower education level. In such a case, the equilibrium profit in the opaque scenario will become higher because the cost of hiring gets lower. Consequently, the opaque algorithm may become more attractive than the transparent algorithm and our result might be altered. Our objective in this subsection is to show that the above argument is only partly true: If we allow the firm to adjust the wage, while it is true that the equilibrium payoff in the opaque scenario can be increased by lowering the wage, the firm can further lower the wage in the transparent scenario while making sure that it does not worsen the agents' degree of separation on

the causal feature. As a result, making the algorithm transparent would still be preferable under certain conditions.

In the main model, the firm is restricted to pay the same wage R regardless of the decision on algorithmic transparency. In this extension we allow the firm to use R_O in the opaque scenario and R_T in the transparent scenario, where R_O and R_T are smaller than R,[10] and we denote the optimal R_O and R_T as R_O* and R_T* respectively. We further assume that $R_O \geq \underline{R}$ where $\underline{R} = \max(\theta\alpha, \beta + \frac{\theta(1-\lambda)\alpha}{\theta(1-\lambda)+(1-\theta)\lambda})$, consistent with Assumptions 2 and 3. *The purpose of this extension is not to fully solve for the equilibrium in the endogenous wage setting, but to show that the major insights from our main model still hold even if the firm is allowed to strategically lower the wage R.*

In this discussion, we will only restrict our attention to the regions that have been previously identified to favor transparent scenario (i.e., regions $C1$, $C2$, and $C4$ defined in Section 4.3). Specifically, we assume the firm will prefer the transparent algorithm if the wage is fixed at R_0, where $(\frac{C_L}{R_0}, \frac{C_H}{R_0})$ lands in regions $C1$, $C2$, and $C4$ defined in Section 4.3. We then let the firm lower the wage from R_0 to R_O* and R_T* in the opaque and transparent scenarios, respectively. We denote the firm's payoff under R_O* (R_T*) where R_0 is in region Ci in the opaque (transparent) scenario as Π^O_{Ci} (Π^T_{Ci}), $i \in \{1, 2, 4\}$.

Proposition 6 $\Pi^O_{Ci} \leq \Pi^T_{Ci}$, $\forall i \in \{1, 2, 4\}$.

Proposition 6 tells us that, for the cases that we have previously identified to favor the transparent scenario, the transparent scenario will still be preferred when the firm is allowed to strategically lower the wage. The intuition behind this proposition is as follows: In the opaque scenario, H type agents hold an advantage on the correlational feature and this will impede them from improving the causal feature. Only when the wage is high will they be incentivized to improve the causal feature and separate themselves from L type agents. Such an advantage does not exist after the algorithm is made transparent so a lower wage is sufficient to separate different types of agents. Consequently, the firm is able to set a lower wage in the transparent scenario than in the opaque scenario without worrying of decreasing agents' *Dos* on the causal feature or decreasing the average value on the causal feature. Since agents' productivity (captured by α and β) is assumed to be unaffected by wage, the firm will benefit more from endogenizing the wage (due to lower a wage expense) in the transparent scenario than in the opaque scenario; thus, the firm's incentive to choose a transparent algorithm will be strengthened.

[10] Here we only consider the firm's decision on lowering R. The case where the firm strategically raises R is similar but less straightforward.

B.3 Agents Have Incorrect Belief of λ

In the main model we have assumed that agents have a correct belief of λ. We now relax this assumption and consider two possibilities: (1) Agents overestimate λ as λ_O and (2) Agents underestimate λ as λ_U. We assume that

$$0.5 < \lambda_U < \lambda < \lambda_O < 1.$$

Furthermore, similar to Assumptions 2 and 3, we assume

$$\frac{(\theta\lambda_U + (1-\theta)(1-\lambda_U))R}{\theta\lambda_U} < \alpha.$$

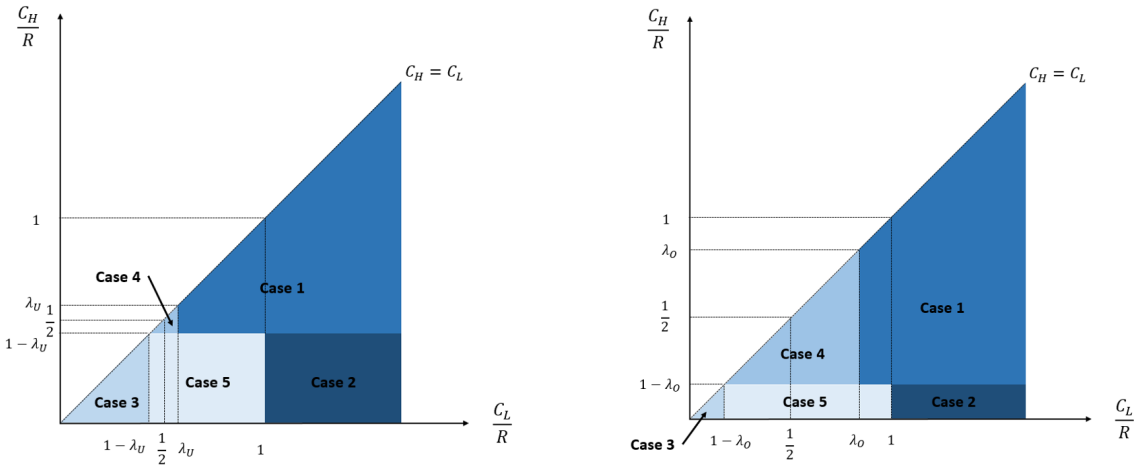
and

$$\frac{\theta(1-\lambda_U)\alpha}{\theta(1-\lambda_U) + (1-\theta)\lambda_U} < R - \beta.$$

The above conditions ensure that, in agents' (incorrect) belief, the firm will have the incentive to include the correlational feature in the algorithm.

Note that the agents' belief about λ is only relevant in the opaque scenario. In the transparent scenario, the agents' equilibrium strategies are the same as discussed in Section 4.2. In the opaque scenario, for both possibilities (1) and (2), it is the agents' belief of λ, not the true λ, that will determine their strategies on the causal feature. However, the true λ will determine the firm's payoff given the agents' strategies. There are still five cases as defined in Section 4.1. However, as shown in Figure 7, the boundaries of these cases are now determined by λ_O or λ_U. Within each case, the firm's payoff function stays the same as in Section 4.1. The comparison between the opaque and transparent scenarios can be performed in a similar way as in Section 4: We can divide the area below the 45-degree line into seven regions, $N1$ to $N3$ and $C1$ to $C4$, as shown in Figure 8.

Similar to our results in Section 4, in regions $N1$ to $N3$, the firm will prefer the opaque algorithm. In region $C4$, the firm will prefer the transparent algorithm. In region $C1$ to $C3$, whether the firm prefers the transparent or opaque algorithm depends on the value of β. Generally speaking, when agents overestimate λ, under a wider range of conditions, the firm will prefer the transparent algorithm. This is reasonable: Overestimating the predictive power of the unknown correlational feature further prevents agents from competing on the causal feature in the opaque scenario. As previously discussed, making the algorithm transparent intensifies agents' competition on the causal feature, which benefits the firm. This benefit is larger when agents overestimate λ. Our result also shows that even if agents underestimate λ, there are still cases in which the transparent algorithm is preferred by the firm.



(a) Equilibrium outcome under belief λ_U

(b) Equilibrium outcome under belief λ_O

Figure 7: Equilibrium outcome when agents have an incorrect belief of λ

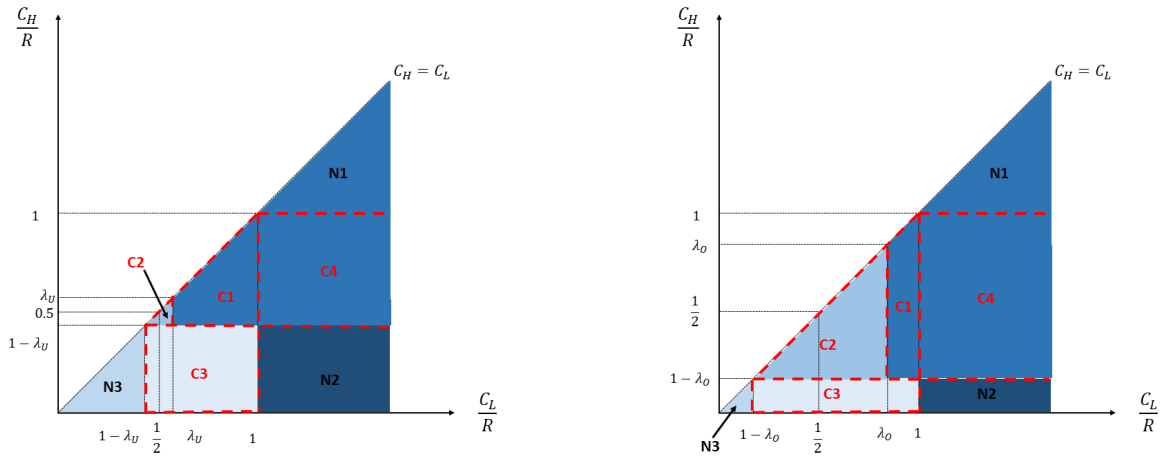
C Omitted Discussions

C.1 Derivation of the Lower and Upper Bound of α and β

In this paper we assume α to be in a certain range to eliminate uninteresting scenarios:

$$\frac{(\theta\lambda + (1-\theta)(1-\lambda))R}{\theta\lambda} < \alpha < \frac{R}{\theta}.$$

In the opaque scenario, when there is no agent who improves the causal feature, we want the firm to hire some agents based on the information in the correlational feature instead of not hiring anyone. (Not hiring anyone in this case is uninteresting because it will trivially drive everyone to improve the causal feature.) Thus we want α to be large enough to incentivize the firm to hire agents who have value 1 on the correlational feature. In the transparent scenario, when there is no agent who improves the causal feature, all the agents are mixed together in the feature space: they all have the same values on both the causal and correlational features. The firm will either hire everyone or not hire anyone, depending one whether the average productivity of all the agents exceeds the salary or not. We want α to be small enough that the firm will not hire anyone in this case. (Hiring everyone in this case is uninteresting because no one will have the incentive to improve on the causal feature regardless of the cost of improving.) Specifically, rewrite the left inequality as $\theta\lambda\alpha + (1-\theta)(1-\lambda) \times 0 > (\theta\lambda + (1-\theta)(1-\lambda))R$. In the initial distribution of the opaque scenario (i.e., where everyone has a value of 0 on the causal feature), there are $\theta\lambda$ H type agents and $(1-\theta)(1-\lambda)$ L type agents who have value 0 on the causal feature and value 1 on the correlational feature. The inequality means that their total productivity (left hand side) should



(a) Agents underestimate λ as λ_U

(b) Agents overestimate λ as λ_O

Figure 8: The comparison between transparent and opaque scenario when agents have incorrect belief on λ

be larger than the total salary paid to them (right hand side). In other words, the firm has an incentive to hire all of these agents. If this is not the case, then the firm will not hire any agents with value 0 on the causal feature even if they have value 1 on the correlational feature, which will trivially incentivize the agents to improve the causal features. The right inequality means that in the transparent scenario where the correlational feature is gamed, if everyone has value 0 on the causal feature, the firm will not hire anyone.

In this paper we also assume β to be in a certain range to eliminate uninteresting scenarios:

$$R - \theta\alpha < \beta < R - \frac{\theta(1-\lambda)\alpha}{\theta(1-\lambda) + (1-\theta)\lambda}.$$

β is the marginal effect of education on the agent's productivity.

Rewrite the left inequality as $\theta(\alpha+\beta)+(1-\theta)\beta > (\theta+(1-\theta))R$. In the transparent scenario, when everyone games on the correlational feature and everyone improves the causal feature, there are θ H type agents and $1-\theta$ L type agents who have value 1 on both features. The inequality means that their total productivity (left-hand side) is larger than the total salary paid to them (right-hand side). In other words, the firm will have an incentive to hire all of them. If this is not the case, then in the transparent scenario, no one will improve on the causal feature and the firm will end up hiring nobody. As for the right part inequality, rewrite it as $\theta(1-\lambda)(\alpha+\beta)+(1-\theta)\lambda\beta < (\theta(1-\lambda)+(1-\theta)\lambda)R$. In the opaque scenario, when no one games on the correlational feature but everyone improves the causal feature, there are $\theta(1-\lambda)$ H type agents and $(1-\theta)\lambda$ L type agents who have value 1 on the causal feature but value 0 on the correlational feature. This inequality

means that these agents' total productivity (left-hand side) is smaller than the total wage paid to them (right-hand side). In other words, the firm will have no incentive to hire anyone of them. If this is not the case, then in the opaque scenario, improving the causal feature will guarantee an agent to be hired regardless of her value on the correlational feature, which will again, lead to an uninteresting equilibrium.

C.2 Discussion on the case where β does not satisfy Assumption 3

In Assumption 3, we restrict β in a certain range to avoid uninteresting cases. We now discuss how the relaxation of this assumption will affect our result. Specifically, we consider two cases (1) the case where $\beta \geq R - \frac{\theta(1-\lambda)\alpha}{\theta(1-\lambda)+(1-\theta)\lambda}$ (i.e., β exceeds the upper bound ($\bar{\beta}$) defined in Assumption 3) and (2) the case where $\beta \leq R - \theta\alpha$ (i.e., β is below the lower bound ($\underline{\beta}$) defined in Assumption 3), this includes the case $\beta = 0$.

We first consider a case where $\beta \geq R - \frac{\theta(1-\lambda)\alpha}{\theta(1-\lambda)+(1-\theta)\lambda}$. The dependence of equilibrium outcomes on parameters in the transparent scenario is the same as shown in the original model. However, the dependence of equilibrium outcomes on parameters in the opaque scenario has changed, as shown in Figure 9. The meaning of different cases are defined in Section 4.1.

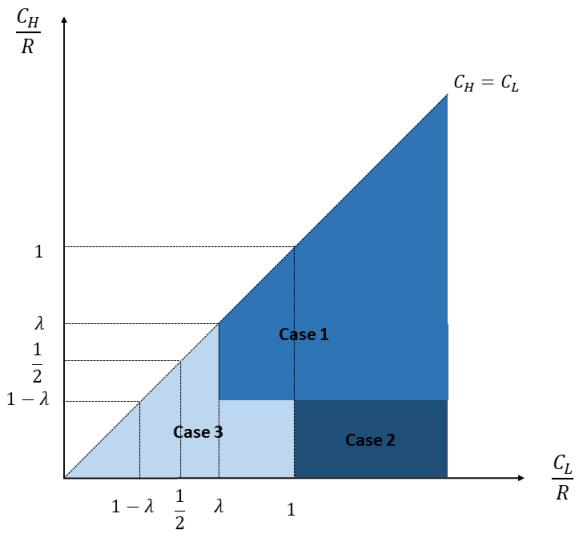


Figure 9: Equilibrium outcome in the opaque scenario when $\beta \geq R - \frac{\theta(1-\lambda)\alpha}{\theta(1-\lambda)+(1-\theta)\lambda}$

Compared with the opaque scenario in our original model, relaxing this assumption will eliminate the two partial separating equilibria. As a result, the firm will be indifferent between transparent and opaque algorithm in the region denoted as case 3 in Figure 9 (i.e., the case where all agents get education and the firm hires everyone). In other regions, the preference of transparency is exactly the same as our original model. This is a degenerate case of our original model.

We now consider a case where $\beta \leq R - \theta\alpha$. The dependence of equilibrium outcomes on parameters in the opaque scenario is the same as shown in the original model. However, the dependence of equilibrium outcomes on parameters in the transparent scenario has changed, as shown in Figure 10.

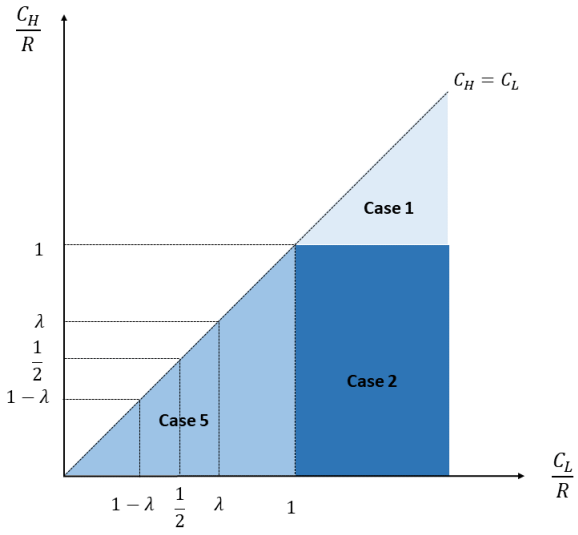


Figure 10: Equilibrium outcome in the transparent scenario when $\beta \leq R - \theta\alpha$

Compared with the transparent scenario in our original model, the previous pure strategy equilibrium case 3 will be replaced by a partial separating equilibrium case 5. As a result, the firm will get weakly higher payoff in the region denoted as case 5 in Figure 10, since in case 5 some L type agents do not get education and thus are separated apart from H type agents. The firm's incentive to choose transparent algorithm weakly increases in this region. In other regions, the preference of transparency is exactly the same as our original model.

C.3 Discussion on the Effect of θ and λ on Decision on Algorithmic Transparency Under the Stackelberg Model

The effect of θ and λ on the firm's choice on opacity and full transparency is summarized in Proposition 7.

Proposition 7 *An increase in λ has the following effects on the firm's decision on transparency:*

1. *The area of regions $C1$, $C2$, and $C4$ increases, which means that the transparent algorithm is preferred under more (C_H,C_L) value pairs.*

2. *Within regions $C2$, the conditions on β to make the transparent algorithm preferred to the opaque algorithm become stricter (i.e. a larger β is needed).*

3. *Within regions $C2$, $C3$, and $N3$ the conditions on C_L to make the transparent algorithm preferred to the opaque algorithm become stricter (i.e. a larger C_L is needed).*

An increase in θ has the following effects on the firm's decision on transparency:

1. *Within region $C2$, the conditions on β to make the transparent algorithm preferred to the opaque algorithm is not affected.*

2. *Within regions $C2$, $C3$, and $N3$ the conditions on C_L to make the transparent algorithm preferred to the opaque algorithm become stricter (i.e. a larger C_L is needed).*

Proposition 7 can be proved by checking

$$\frac{\partial C_L^{C2}}{\partial \lambda} > 0, \frac{\partial C_L^{C2}}{\partial \theta} > 0$$
$$\frac{\partial C_L^{C3}}{\partial \lambda} > 0, \frac{\partial C_L^{C3}}{\partial \theta} > 0$$
$$\frac{\partial C_L^{N3}}{\partial \lambda} > 0, \frac{\partial C_L^{N3}}{\partial \theta} > 0,$$

In general, as β and λ increase, the transparent algorithm becomes more attractive compared with the opaque algorithm. This is because when β and C_L increase, the firm's equilibrium payoff in the full transparency scenario increases sharply, while the equilibrium payoff in the opaque scenario is less affected by these two parameters.

D Mathematical Appendix (Proofs of All Results)

D.1 Proof of Lemma 1

In the proof, we will proceed as follows: First, we identify different regions in the C_H-C_L space in which a certain class of strategies can be sustained as an equilibrium; then we analyze the corresponding payoffs for the firm and agents.

In the opaque scenario, agents move first and then the firm moves after observing individuals' actions. Even though the players in this game move sequentially, we can show that the PBE of this game coincides with the Bayesian Nash Equilibrium (BNE) of the corresponding simultaneously move game. The reasons are as follows: The strategies of all the agents will influence the firm's choice of algorithms. However, the influence of a single agent's action on the firm's strategy can be neglected. Once an equilibrium is reached, an agent (first mover) assesses the profitability of a deviation by assuming that the firm's (second mover) strategy will not change accordingly. In

other words, all the PBEs that can be sustained in this scenario are actually the BNEs of the corresponding simultaneous move game.

Opaque scenario 1. We first look at the case where neither H type nor L agents improve education. In this case, $\gamma_S^e = \gamma_S^b$. The firm's best response is: $P_A = P_B = P_D = 1, P_C = 0$.[11] For this strategy combination to be a PBE, the following conditions need to be satisfied:

$$\begin{aligned} C_H &\geq (1-\lambda)R && (H \text{ type agents will not deviate}) \\ C_L &\geq \lambda R && (L \text{ type agents will not deviate}) \\ \gamma_A^e &\geq \gamma_{th0} && (\text{The firm will not deviate on } P_A) \\ \gamma_C^e &\leq \gamma_{th0} && (\text{The firm will not deviate on } P_C). \end{aligned}$$

(The first two conditions ensures that the H type and L type agents are better off (in terms of utility) not improving education, and the last two conditions ensure that the firm is better off (in terms of total payoffs) not deviating from its current P_A and P_C.) The first two conditions specify the regions in C_H-C_L space that this combination of strategies can be sustained as an equilibrium (Figure 11 shows this region). The last two conditions are the direct consequences of Assumption 2:

$$\frac{\lambda\theta}{\lambda\theta + (1-\lambda)(1-\theta)} \geq \frac{R}{\alpha} = \gamma_{th0} \qquad \frac{(1-\lambda)\theta}{(1-\lambda)\theta + \lambda(1-\theta)} \leq \theta \leq \frac{R}{\alpha} = \gamma_{th0}.$$

The total payoffs for the firm and each type of agents are given by:

$$\begin{aligned} \Pi_{firm_{O1}} &= \lambda\theta\alpha - (\lambda\theta + (1-\lambda)(1-\theta))R \\ \Pi_{H_{O1}} &= \theta\lambda R \\ \Pi_{L_{O1}} &= (1-\theta)(1-\lambda)R. \end{aligned}$$

where we use $\Pi_{H_{O1}}$ and $\Pi_{L_{O1}}$ to denote the total payoff of H type and L type agents, respectively. (In the remaining of the proof, we will use $\Pi_{H_{Oi}}$ and $\Pi_{L_O i}$ to denote the total payoff of H type and L type agents under case i, respectively.)

Opaque scenario 2. In this case, only H type agents improve education but L type agents do not, and we have: $\gamma_B^e = \gamma_D^e = 1$ and $\gamma_A^e = \gamma_C^e = 0$. The firm's best response is $P_A = P_C = 0$, $P_B = P_D = 1$. For this strategy combination to be a PBE, the following conditions on the

[11] Any $P_B \in [0,1]$ and $P_D \in [0,1]$ can be sustained since they are off the equilibrium path. Here we pick the maximum value of P_B and P_D to make sure the equilibrium could survive the intuitive criterion.

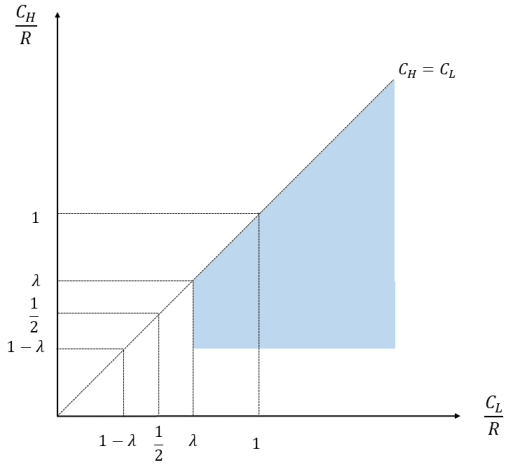


Figure 11: Opaque scenario 1

parameters need to be satisfied:

$$
\begin{aligned}
&C_H \leq R && (H \text{ type agents will not deviate})\\
&C_L \geq R && (L \text{ type agents will not deviate})\\
&\gamma_B^e \geq \gamma_{th1}, \gamma_D^e \geq \gamma_{th1} && (\text{The firm will not deviate on } P_B \text{ and } P_D)\\
&\gamma_A^e \leq \gamma_{th0}, \gamma_C^e \leq \gamma_{th0} && (\text{The firm will not deviate on } P_A \text{ and } P_C).
\end{aligned}
$$

The first two conditions specify the regions (see Figure 12) where this combination of strategies can be sustained as an equilibrium. The last two constraints are trivially satisfied. The total payoffs for the firm and each type of agents are given by:

$$
\begin{aligned}
&\Pi_{firm_{O2}} = \theta(\alpha + \beta) - \theta R\\
&\Pi_{H_{O2}} = \theta(R - C_H)\\
&\Pi_{L_{O2}} = 0.
\end{aligned}
$$

Opaque scenario 3. In this case, both H type and L type agents improve education. The values of γ^e's are given below:

$$
\begin{aligned}
&\gamma_A^e = \gamma_C^e = 0\\
&\gamma_B^e = \frac{\lambda\theta}{\lambda\theta + (1-\lambda)(1-\theta)}\\
&\gamma_D^e = \frac{(1-\lambda)\theta}{(1-\lambda)\theta + \lambda(1-\theta)}.
\end{aligned}
$$

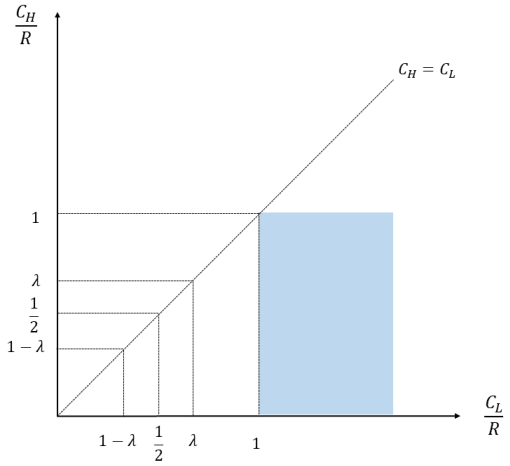


Figure 12: Opaque scenario 2

The firm's best response is to set $P_A = P_C = P_D = 0, P_B = 1$. For this strategy combination to be a PBE, we need the following conditions:

$$\begin{aligned}
&C_H \leq \lambda R && (H \text{ type agents will not deviate})\\
&C_L \leq (1-\lambda)R && (H \text{ type agents will not deviate})\\
&\gamma_B^e \geq \gamma_{th1} && (\text{The firm will not deviate on } P_B)\\
&\gamma_D^e \leq \gamma_{th1} && (\text{The firm will not deviate on } P_D).
\end{aligned}$$

The first two conditions specify the regions (see Figure 13) where this combination of strategies can be sustained as an equilibrium. The last two conditions are direct consequences of Assumption 1 and 2. The total payoffs for the firm and each type of agents are given by:

$$\begin{aligned}
&\Pi_{firm_{O3}} = \lambda\theta(\alpha+\beta) + (1-\lambda)(1-\theta)\beta - (\lambda\theta + (1-\lambda)(1-\theta))R\\
&\Pi_{H_{O3}} = \theta\lambda R - \theta C_H\\
&\Pi_{L_{O3}} = (1-\theta)(1-\lambda)R - (1-\theta)C_L.
\end{aligned}$$

Opaque scenario 4. In this case, H type agents improve education with probability p_H and L type agents do not improve education. The values of γ^e's are given below:

$$\begin{aligned}
\gamma_A^e &= \frac{\theta(1-p_H)\lambda}{\theta(1-p_H)\lambda + (1-\theta)(1-\lambda)}\\
\gamma_C^e &= \frac{\theta(1-p_H)(1-\lambda)}{\theta(1-p_H)(1-\lambda) + (1-\theta)\lambda}\\
\gamma_B^e &= \gamma_D^e = 1.
\end{aligned}$$

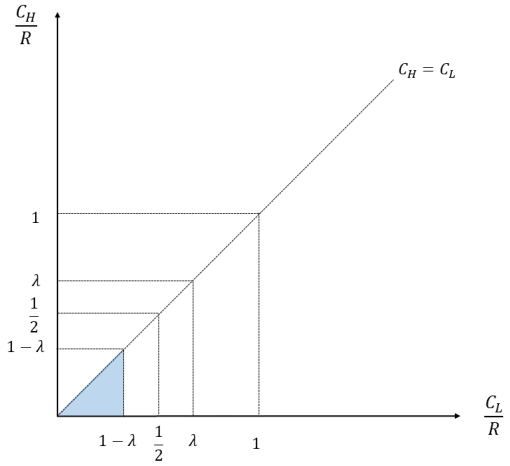


Figure 13: Opaque scenario 3

The firm's best response is to use $P_A = p_4, P_C = 0, P_B = P_D = 1$. For this strategy combination to be a PBE, the following conditions should hold:

$$\begin{aligned}
&C_H = (1-\lambda p_4)R && (H \text{ type agents are indifferent})\\
&C_L \geq (1-(1-\lambda)p_4)R && (L \text{ type agents will not deviate})\\
&\gamma_A^e = \gamma_{th0} && (\text{The firm is indifferent on } P_A)\\
&\gamma_C^e \leq \gamma_{th0} && (\text{The firm will not deviate on } P_C).
\end{aligned}$$

The last condition is satisfied following Assumption 2:

$$\gamma_C^e = \frac{\theta(1-p_H)(1-\lambda)}{\theta(1-p_H)(1-\lambda)+(1-\theta)\lambda} \leq \frac{(1-\lambda)\theta}{(1-\lambda)\theta+\lambda(1-\theta)} \leq \frac{R}{\alpha} = \gamma_{th0}.$$

The first condition could be used to express p as a function of the other parameters and, similarly, the third condition can be used to express p_H as a function of the other parameters. Specifically, we have:

$$p_4 = \frac{1}{\lambda}\left(1-\frac{C_H}{R}\right) \tag{D.1}$$

$$p_H = 1-\frac{R(1-\theta)(1-\lambda)}{(\alpha-R)\theta\lambda}. \tag{D.2}$$

Given the fact that p and p_H are values between 0 and 1, we can calculate the range for C_H and C_L in which this combination of strategies can be sustained as an equilibrium (see Figure 14):

$$\begin{aligned}
&(1-\lambda)R \leq C_H \leq R\\
&\frac{R-C_H}{R-C_L} \leq \frac{\lambda}{1-\lambda}.
\end{aligned}$$

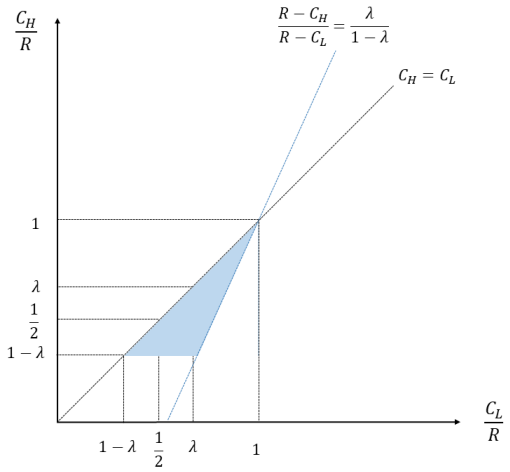


Figure 14: Opaque scenario 4

where the first inequality follows from the first condition and the second inequality follows from the second condition.

The total payoffs for the firm and each type of agents are given by:

$$\Pi_{firm_{O4}} = \theta p_H(\alpha+\beta) - \theta p_H R$$
$$\Pi_{H_{O4}} = \theta(R - C_H)$$
$$\Pi_{L_{O4}} = p_4(1-\theta)(1-\lambda)R.$$

Opaque scenario 5. In this case, H type agents improve education and L type agents improve education with probability p_L. We have:

$$\gamma_A^e = \gamma_C^e = 0$$
$$\gamma_B^e = \frac{\theta\lambda}{\theta\lambda + (1-\theta)p_L(1-\lambda)}$$
$$\gamma_D^e = \frac{\theta(1-\lambda)}{\theta(1-\lambda) + (1-\theta)p_L\lambda}.$$

The firm's best response is to use $P_A = P_C = 0, P_B = 1, P_D = p_5$. For this strategy combination to be a PBE, the following conditions should hold:

$$C_H \leq (\lambda + p_5(1-\lambda))R \qquad (H \text{ type agents will not deviate})$$
$$C_L = ((1-\lambda) + p_5\lambda)R \qquad (L \text{ type agents are indifferent})$$
$$\gamma_B^e \geq \gamma_{th1} \qquad (\text{The firm will not deviate on } P_B)$$
$$\gamma_D^e = \gamma_{th1} \qquad (\text{The firm will not deviate on } P_D).$$

The second condition can be used to express p as a function of the other parameters while the last condition can be used to express p_L as a function of the other parameters. Specifically,

$$p_5 = \frac{C_L - R}{\lambda R} + 1 \tag{D.3}$$

$$p_L = \frac{(\alpha + \beta - R)\theta(1-\lambda)}{(1-\theta)\lambda(R-\beta)}. \tag{D.4}$$

The third condition is satisfied following Equations 2 and 3:

$$\gamma_B^e = \frac{\theta\lambda}{\theta\lambda + (1-\theta)p_L(1-\lambda)} \geq \gamma_{th0}.$$

Given the fact that p and p_L are values between 0 and 1, we can calculate the range for C_H and C_L in which this combination of strategies can be sustained as an equilibrium:

$$(1-\lambda)R \leq C_L \leq R$$
$$\frac{R - C_H}{R - C_L} \geq \frac{1-\lambda}{\lambda}.$$

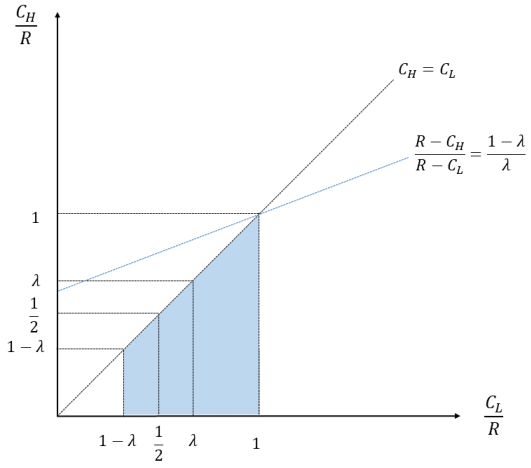


Figure 15: Opaque scenario 5

The total payoffs for the firm and each type of agents are given by:

$$\begin{aligned}
\Pi_{firm_{O5}} &= \theta\lambda(\alpha + \beta - R) + (1-\theta)p_L(1-\lambda)(\beta - R) \\
&= \frac{2\lambda - 1}{\lambda}\theta(\alpha + \beta - R) \\
\Pi_{H_{O5}} &= (\theta\lambda + \theta(1-\lambda)p_5)R - \theta C_H \\
\Pi_{L_{O5}} &= 0.
\end{aligned}$$

Dealing with multiple equilibria. Per our analysis of the above five cases, there are several regions where multiple equilibria exist. According to the dynamics of the game, in the opaque

scenario, the agents move first and the firm moves next. Thus, we first narrow down to the equilibria that survive the intuitive criterion (Cho and Kreps, 1987) and then select the equilibrium outcome which gives the largest total utilities for each agent type.[12] (In theory, finding such an equilibrium is not always possible; fortunately, it is possible in our case.)

- In the region where Case 4 and Case 5 overlap, Case 4 always gives higher payoffs to both H type and L type agents:

$$\Pi_{H_{O4}} = \theta(R - C_H) > (\theta\lambda + \theta(1-\lambda)p_5)R - \theta C_H = \Pi_{H_{O5}}$$
$$\Pi_{L_{O4}} = p_4(1-\theta)(1-\lambda)R > 0 = \Pi_{L_{O5}}.$$

  The inequalities hold since $p_4 = \frac{R-C_H}{\lambda R}$ and $p_5 = \frac{C_L - R}{\lambda R} + 1$ are values between 0 and 1.

- In the region where Case 4 and Case 1 overlap, Case 1 always gives higher payoffs to both H type and L type agents:

$$\Pi_{H_{O1}} = \theta\lambda R \geq \theta(R - C_H) = \Pi_{H_{O4}}$$
$$\Pi_{L_{O1}} = (1-\theta)(1-\lambda)R \geq p_4(1-\theta)(1-\lambda)R = \Pi_{L_{O4}}.$$

  The inequalities hold since $\frac{C_H}{R} \geq 1 - \lambda$ in the overlapped region and $p_4 = \frac{R-C_H}{\lambda R}$ is between 0 and 1.

- In the region where Case 1 and Case 2 overlap, Case 1 always gives higher payoffs to both H type and L type agents:

$$\Pi_{H_{O1}} = \theta\lambda R \geq \theta(R - C_H) = \Pi_{H_{O2}}$$
$$\Pi_{L_{O1}} = (1-\theta)(1-\lambda)R \geq 0 = \Pi_{L_{O2}}.$$

  The first inequality holds since $\frac{C_H}{R} \geq 1 - \lambda$ in the overlapped region.

- In the region where Case 1 and Case 5 overlap, Case 1 always gives higher payoffs to both H type and L type agents:

$$\Pi_{H_{O1}} \geq \Pi_{H_{O4}} \geq \Pi_{H_{O5}}$$
$$\Pi_{L_{O1}} \geq \Pi_{L_{O4}} \geq \Pi_{L_{O5}}.$$

  The first inequality holds since $\frac{C_H}{R} \geq 1 - \lambda$ in the overlapped region.

[12] We select the equilibrium that gives the largest total utility for each agent type, which shares the same spirit of the undefeated criterion introduced by Mailath et al. (1993). (See also (Schmidt and Buell, 2017))

D.2 Proof of Lemma 2

Similar to the proof of the opaque scenario, we proceed by analyzing the same five cases analyzed in the opaque scenario. We show that only the equilibrium outcomes corresponding to cases 1 to 3 are sustainable. Recall that in Lemma 2 we study the case where the firm does not have commitment power, thus the firm's action in the first stage (announcing the hiring probabilities) is not binding. Similar to what we have done in Lemma 1, we can show the PBEs of this game coincide with the BNEs of the corresponding simultaneous move game.

Transparent scenario 1. We first look at the case where neither H type nor L type agents improve education. In this case, $\gamma_E^e = \theta$ and $\gamma_F^e = 0$. The firm's best response is: $P_E = 0$, $P_F = 1$. For this strategy combination to be a PBE, the following conditions should be satisfied:

$$\begin{aligned} C_H &\geq R && (H \text{ type agents will not deviate}) \\ C_L &\geq R && (L \text{ type agents will not deviate}) \\ \gamma_E^e &\leq \gamma_{th0} && (\text{The firm will not deviate deviate on } P_E). \end{aligned}$$

The last condition is the direct consequence of Assumption 2:

$$\gamma_E^e = \theta < \frac{R-\beta}{\alpha} < \frac{R}{\alpha} = \gamma_{th0}.$$

The first two conditions specify the regions in C_H-C_L space that this combination of strategies can be sustained as an equilibrium (Figure 16 shows this region). The total payoffs for the firm and each type of agents are given by:

$$\begin{aligned} \Pi_{firm_{T1}} &= 0 \\ \Pi_{H_{T1}} &= 0 \\ \Pi_{L_{T1}} &= 0. \end{aligned}$$

where we use $\Pi_{H_{T1}}$ and $\Pi_{L_{T1}}$ to denote the total payoff of H type and L type agents, respectively. (In the remaining of the proof, we will use $\Pi_{H_{Ti}}$ and $\Pi_{L_{Ti}}$ to denote the total payoff of H type and L type agents under transparent case i, respectively.)

Transparent scenario 2. In this case, only H type agents improve education but L type agents do not, and we have: $\gamma_E^e = 0$ and $\gamma_F^e = 1$. The firm's best response is $P_E = 0$, $P_F = 1$. For this strategy combination to be a PBE, the following conditions on the parameters need to be satisfied:

$$\begin{aligned} C_H &\leq R && (H \text{ type agents will not deviate}) \\ C_L &\geq R && (L \text{ type agents will not deviate}) \\ \gamma_E^e &\leq \gamma_{th0} && (\text{The firm will not deviate on } P_E) \\ \gamma_F^e &\geq \gamma_{th1} && (\text{The firm will not deviate on } P_F). \end{aligned}$$

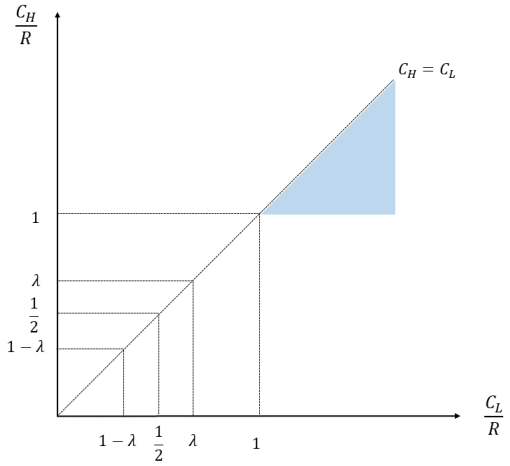


Figure 16: Transparent scenario 1

The last two conditions are trivially satisfied (by Assumption 1, we have $0 < \gamma_{th1} < \gamma_{th0} < 1$). The first two conditions specify the regions in C_H-C_L space that this combination of strategies can be sustained as an equilibrium (Figure 17 shows this region). The total payoffs for the firm and each type of agents are given by:

$$\Pi_{firm_{T2}} = \theta(\alpha + \beta - R)$$
$$\Pi_{H_{T2}} = \theta(R - C_H)$$
$$\Pi_{L_{T2}} = 0.$$

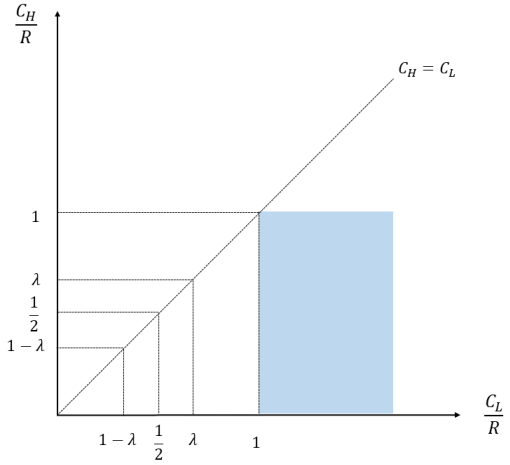


Figure 17: Transparent scenario 2

Transparent scenario 3. In this case, both H type and L type agents improve education, and

we have: $\gamma_E^e = 0$ and $\gamma_F^e = \theta$. The firm's best response is $P_E = 0$, $P_F = 1$. For this strategy combination to be a PBE, the following conditions on the parameters need to be satisfied:

$$\begin{aligned} C_H &\leq R && (H \text{ type agents will not deviate}) \\ C_L &\leq R && (L \text{ type agents will not deviate}) \\ \gamma_F^e &\geq \gamma_{th1} && (\text{The firm will not deviate on } P_F). \end{aligned}$$

The third condition is a direct consequence of Assumption 3. The first two conditions specify the regions in C_H-C_L space that this combination of strategies can be sustained as an equilibrium (Figure 18 shows this region). The total payoffs for the firm and each type of agents are given by:

$$\begin{aligned} \Pi_{firm_{T3}} &= \theta(\alpha + \beta) + (1 - \theta)\beta - R \\ \Pi_{H_{T3}} &= \theta(R - C_H) \\ \Pi_{L_{T3}} &= (1 - \theta)(R - C_L). \end{aligned}$$

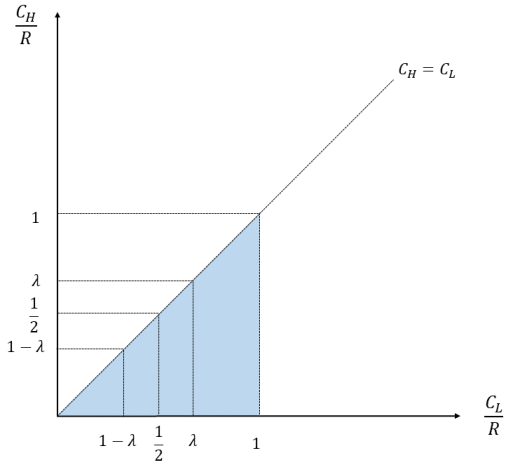


Figure 18: Transparent scenario 3

Transparent scenario 4. In this case, H type agents improve education with probability p_H and L type agents do not improve education. The values of γ^e's are given below:

$$\begin{aligned} \gamma_E^e &= \frac{(1 - p_H)\theta}{(1 - p_H)\theta + (1 - \theta)} \\ \gamma_F^e &= 1. \end{aligned}$$

The firm's best response is $P_E = p$, $P_F = 1$. For this strategy combination to be a PBE, the

following conditions on the parameters need to be satisfied:

$$\begin{aligned} C_H &= (1-p)R && (H \text{ type agents are indifferent}) \\ C_L &\geq (1-p)R && (L \text{ type agents will not deviate}) \\ \gamma_E^e &= \gamma_{th0} && (\text{The firm is indifferent on } P_E). \end{aligned}$$

Since $p_H \in [0,1]$, $0 < \gamma_E^e < \theta$. The last condition requires $0 < \frac{R}{\alpha} < \theta$, or equivalently $\alpha > \frac{R}{\theta}$. In the range of α that we are considering (i.e., $\frac{(\theta\lambda+(1-\theta)(1-\lambda))R}{\theta\lambda} < \alpha < \frac{R}{\theta}$, by Assumption 2), this combination of strategies cannot be sustained as an equilibrium.

Transparent scenario 5. In this case, H type agents improve education and L type agents improve education with probability p_L. The values of γ^e's are given below:

$$\begin{aligned} \gamma_E^e &= 0 \\ \gamma_F^e &= \frac{\theta}{\theta + (1-\theta)p_L}. \end{aligned}$$

The firm's best response is $P_E = 0$, $P_F = p$. For this strategy combination to be a PBE, the following conditions on the parameters need to be satisfied:

$$\begin{aligned} C_H &\leq pR && (H \text{ type agents will not deviate}) \\ C_L &= pR && (L \text{ type agents are indifferent}) \\ \gamma_F^e &= \gamma_{th1} && (\text{The firm is indifferent on } P_F). \end{aligned}$$

Note that, the second condition implies $p = \frac{C_L}{R}$. Given that $p \in [0,1]$, any value of $C_L \in [0,R]$ is valid. As for the last condition, $p_L \in [0,1]$ implies $\theta < \gamma_F^e < 1$. However, by Assumption 3, $\beta > R - \theta\alpha$, which implies $\gamma_{th1} = \frac{R-\beta}{\alpha} < \theta$. Thus, the last condition cannot be satisfied and, therefore, this combination of strategies cannot be sustained as an equilibrium.

D.3 Proof of Lemma 3

The proof of Lemma 3 is straightforward: $\Pi_{agents_{Oi}} = \Pi_{H_{Oi}} + \Pi_{L_{Oi}} \forall i \in \{1,2,3,4,5\}$, $\Pi_{agents_{Ti}} = \Pi_{H_{Ti}} + \Pi_{L_{Ti}} \forall i \in \{1,2,3\}$, where $\Pi_{H_{Oi}}$, $\Pi_{L_{Oi}}$, $\Pi_{H_{Ti}}$, $\Pi_{L_{Ti}}$ are defined in Appendix D.1 and D.2.

D.4 Proof of Lemma 4 and Lemma 5

We start from the equilibrium strategy in the partial transparency case and see whether there exists an ex-post sub-optimal strategy that the firm can commit to that can further increase the firm's payoff. In region 1 of Figure 5, any commitment on hiring strategy will not change agents' behavior on the causal feature, because in this region, neither type of agents will improve education due to the large values of C_H and C_L. Therefore, the firm will not commit to an ex-post sub-optimal strategy.

The equilibrium payoff in this region is equal to that in the partial transparency case $\Pi_{FT1} = 0$. In region 2, the firm has no incentive to commit to an ex-post sub-optimal strategy because the H type and L type agents are already perfectly separated on the causal feature, and the firm has achieved the maximum payoff. Thus the equilibrium payoff in this region is $\Pi_{FT2} = \theta(\alpha + \beta - R)$. However, in region 3, the firm may have an incentive to commit to a strategy that is ex-post sub-optimal. Given a (C_H, C_L) combination in this region, the firm could announce hiring probabilities $P_E = 0$ and $P_F = \frac{C_L}{R}$.[13] Observing this hiring strategy, H type agents improve education but L type agents choose not to. The firm then follows the pre-announced hiring strategy and hires $\frac{C_L}{R}$ portion of H type agents in State F, and this will result in a payoff of $\Pi_{Stackelberg} = \frac{C_L\theta(\alpha+\beta-R)}{R}$. This payoff is greater than the equilibrium payoff when the firm does not have the commitment power (i.e., $\Pi_{firm_{T3}} = \theta(\alpha + \beta) + (1 - \theta)\beta - R$ in Lemma 2) if and only if $\frac{C_L}{R} \geq p_0$ where $p_0 = \frac{\theta\alpha+\beta-R}{\theta(\alpha+\beta-R)}$. The region in which the condition $\frac{C_L}{R} \geq p_0$ holds is denoted as region S in Figure 5.

As for the agents welfare, it is easy to check that in region S, all H type agents improve education but only $\frac{C_L}{R}$ fraction of them will be hired, thus $\Pi_{agents_{FTS}} = \theta(C_L - C_H)$. In other regions, agents welfare stays the same as in the partial transparency case because neither agents' nor the firm's strategy changes.

D.5 Proof of Lemma 6 and Lemma 7

We start with the 'no-advantage case'. We first consider the case where $c \geq R$. In this case, no agents will improve the correlational feature. If $C_H > R$ and $C_L > R$, the firm sets $P_A = 1$, $P_B = P_C = P_D = 0$, no agents will move, and the equilibrium payoff to the firm is given by $\Pi_{1a}^{ex1} = \lambda\theta(\alpha - R) - (1 - \lambda)(1 - \theta)R$. If $C_H \leq R$ and $C_L > R$, the firm sets $P_A = P_C = 0$, $P_B = P_C = 1$, all H type agents improve the causal feature, and the equilibrium payoff to the firm is $\Pi_{2a}^{ex1} = \theta(\alpha + \beta - R)$. If $C_H \leq R$ and $C_L \leq R$, the firm sets $P_B = 1$, $P_A = P_C = P_D = 0$, all agents in state A move to state B, and the equilibrium payoff to the firm is $\Pi_{3a}^{ex1} = \Pi_{3b}^{ex1} = \Pi_{3c}^{ex1} = \lambda\theta(\alpha + \beta - R) + (1 - \lambda)(1 - \theta)(\beta - R)$.

We next consider the 'no-advantage case' with $c < R$. We identify an equilibrium for each of the 5 regions denoted in Figure 6. In region 1a, the firm sets $P_A = \frac{c}{R}$, $P_C = 0$, $P_B \in [0, 1]$, $P_D \in [0, 1]$, a fraction of agents in State C move to State A to make the firm indifferent between hiring and not hiring in State A. The equilibrium payoff to the firm is given by $\Pi_{1a}^{ex2} = 0$. In region 2a, the firm sets $P_A = P_C = 0$, $P_B = P_D = 1$, all H type agents improve the causal feature, and the equilibrium payoff to the firm is $\Pi_{2a}^{ex2} = \theta(\alpha + \beta - R)$. In region 3a, the firm sets $P_A = P_C = 0$, $P_B = P_D = 1$, all agents improve the causal feature, and the equilibrium payoff to

[13] Here we only consider the case where $\theta < \frac{R}{\alpha+\beta}$, the results in for the case where $\theta \geq \frac{R}{\alpha+\beta}$could be derived similarly.

the firm is $\Pi_{3a}^{ex2} = \theta(\alpha + \beta - R) + (1 - \theta)(\beta - R)$. In region 3b, the firm sets $P_A = P_C = P_D = 0$, $P_B = 1$, H type agents in State A and C move to State B, and the equilibrium payoff to the firm is $\Pi_{3b}^{ex2} = \theta(\alpha + \beta - R) + (1 - \lambda)(1 - \theta)(\beta - R)$. In region 3c, the firm sets $P_A = P_C = P_D = 0$, $P_B = 1$, H type agents in State A move to State B, and the equilibrium payoff to the firm is $\Pi_{3c}^{ex2} = \lambda\theta(\alpha + \beta - R) + (1 - \lambda)(1 - \theta)(\beta - R)$.

By comparing the equilibrium payoffs in the 'costly' and the 'costless' setting, we have $\Pi_{1a}^{ex1} > \Pi_{firmT1}$, $\Pi_{2a}^{ex1} = \Pi_{firmT2}$, $\Pi_{3a}^{ex1} = \Pi_{3b}^{ex1} = \Pi_{3c}^{ex1} > \Pi_{firmT3}$. Moreover, $\Pi_{1a}^{ex2} = \Pi_{firmT1}$, $\Pi_{2a}^{ex2} = \Pi_{firmT2}$, $\Pi_{3a}^{ex2} = \Pi_{firmT3}$, $\Pi_{3b}^{ex2} > \Pi_{firmT3}$, $\Pi_{3c}^{ex2} > \Pi_{firmT3}$. Thus, the firm will receive a weakly better payoff than in the 'costless' setting for any combination of C_H, C_L, and c.

D.6 Proof of Proposition 3

We first identify the conditions under which the firm will prefer the fully transparent algorithm over the opaque algorithm; this can be done by comparing the firm's equilibrium payoff in each scenario (shown in Lemma 1 and Lemma 4). The conditions are summarized in Theorem 3.

Theorem 3 *The firm will prefer the fully transparent algorithm over the opaque algorithm if and only if any of the following conditions is satisfied:*

- *Value pair* $(\frac{C_L}{R}, \frac{C_H}{R})$ *falls in region* $C1$.
- *Value pair* $(\frac{C_L}{R}, \frac{C_H}{R})$ *falls in region* $C4$.
- *Value pair* $(\frac{C_L}{R}, \frac{C_H}{R})$ *falls in region* $C2$ *and either* $\beta > \beta_2$ *or* $C_L > C_L^{C2}$.
- *Value pair* $(\frac{C_L}{R}, \frac{C_H}{R})$ *falls in region* $C3$ *and* $C_L > C_L^{C3}$.
- *Value pair* $(\frac{C_L}{R}, \frac{C_H}{R})$ *falls in region* $N3$ *and* $C_L > C_L^{N3}$.

where $C_L^{C2} = R - \frac{R^2(1-\theta)(1-\lambda)}{(\alpha-R)\theta\lambda}$, $C_L^{C3} = \frac{(2\lambda-1)R\theta}{\lambda}$, $C_L^{N3} = R\lambda\theta + \frac{R(1-\lambda)(1-\theta)(\beta-R)}{\theta(\alpha+\beta-R)}$.

Now we prove Theorem 3. Recall that, in Appendix D.4, we have shown that the firm's equilibrium payoffs are the same in the partial transparency scenario and the full transparency scenario if the value pair $(\frac{C_L}{R}, \frac{C_H}{R})$ falls in region $N1$, $N2$, and $C4$. Thus, in these regions, the firm's preference of transparency will not change when we switch from partial transparency to full transparency. If the value pair instead falls in regions $C1$, $C2$, $C3$, and $N3$, the firm may get a strictly higher payoff when we consider full transparency instead of partial transparency since these regions might overlap with region S. Consequently, in these regions, the firm's decision on algorithmic transparency might be altered if we study full transparency instead of partial transparency. First, notice that the left boundary of region S is $\frac{C_L}{R} = p_0 = \frac{\theta\alpha+\beta-R}{\theta(\alpha+\beta-R)}$. It is not difficult to check that

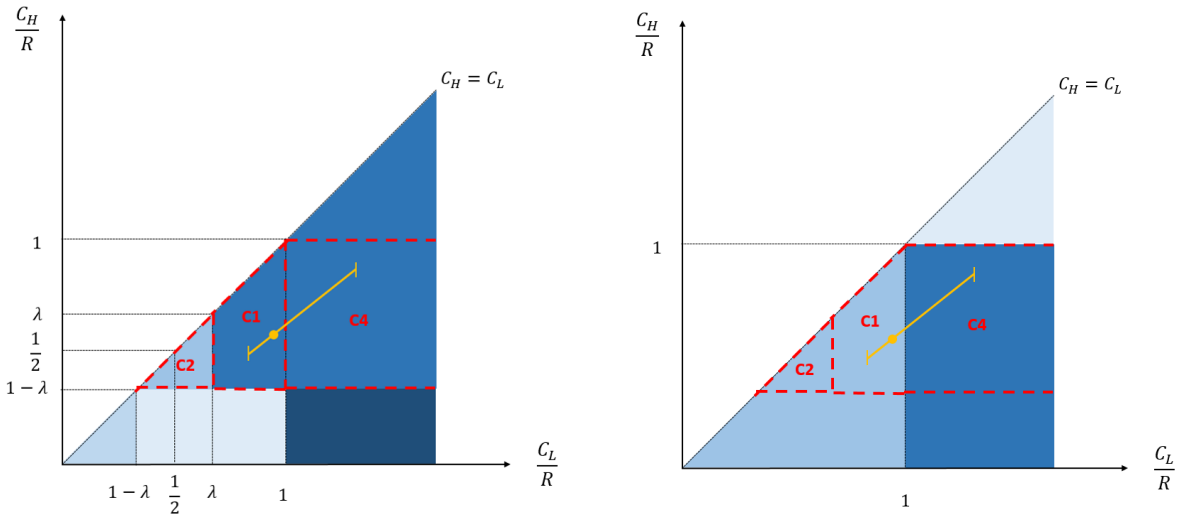
$p_0 \in [0, 2 - \frac{1}{\lambda}]$ and region $C1$ will be included in region S since the left boundary of region $C1$ is $\frac{C_L}{R} = \lambda$ and $\lambda \geq 2 - \frac{1}{\lambda}$. Within region $C1$, the firm's payoff is constant in the opaque scenario ($\Pi_O^{C1} = \lambda\theta\alpha - (\lambda\theta + (1-\lambda)(1-\theta))R$) but increases in C_L in the full transparency scenario. The minimum payoff for the firm in the full transparency scenario in region C_1 is reached at the left boundary and its value is $\min \Pi_{FT}^{C1} = \lambda\theta(\alpha + \beta - R)$. Since $\min \Pi_{FT}^{C1} - \Pi_O^{C1} = \lambda\theta\beta + (1-\lambda)(1-\theta)R > 0$, the firm will always prefer full transparency over opacity in region $C1$. For region $C2$, $\beta > \beta_2$ is a sufficient condition for the firm to prefer partial transparency over opacity and, thus, is also a sufficient condition for the firm to prefer full transparency over opacity since the firm's payoff is weakly higher in the full transparency scenario than in the partial transparency scenario. Another sufficient condition on $C_L > C_L^{C2}$ could be found by letting $\Pi_{FT}^{C2} = \Pi_O^{C2}$. Adding together, ($\beta > \beta_2$ or $C_L > C_L^{C2}$) constitutes a necessary and sufficient condition for the firm to prefer transparency over opacity in region $C2$. The necessary and sufficient conditions for the firm to prefer transparency in region $C3$ and $N3$ can be found similarly and the details are omitted here.

D.7 Proof of Proposition 5

Our discussion on the 'advantage' case where $c_l > c_h$ is built on the previous case where $c_h = c_l = c$ shown in Lemma 6 and Lemma 7. We show that an increase in c_l can weakly increase the firm's payoff in the transparent scenario. We will show this case by case in Figure 6. In region 1a, increasing c_l will not impact agents' behavior on the causal feature. It can only affect L type agents' incentive to game the correlational feature, and this is always beneficial to the firm since the two types of agents will become more separated on the correlational feature compared with the 'no advantage' case. In regions 2a, since L type agents will not improve on the correlational feature even in the 'no-advantage' case, increasing c_l will not affect the equilibrium outcome and the payoff for the firm. In region 3a, before c_l reaches R, it would have no impact on the equilibrium outcome. If c_l increases to a value greater than R, the firm will be better off compared with the 'no advantage' case since L type agents in State C will not be willing to move to State B and thus H type and L type agents can be better separated apart. In regions 3b and 3c, increasing c_l will have no impact on the equilibrium outcome since L type agents in State C will not improve the causal feature even in the 'no advantage' case. To summarize, the firm gets weakly better payoffs in the 'advantage' setting than in the 'no advantage' setting.

D.8 Proof of Proposition 6

To make our discussion concrete, we assume the firm starts from a fixed wage R_0 and can strategically lower the wage to R_O and R_T in the opaque and transparent scenario, respectively. On the one hand, the firm has an incentive to decrease R to reduce the cost of hiring; on the other



(a) Endogenous R in the opaque scenario

(b) Endogenous R in the transparent scenario

Figure 19: Illustration on how the firm chooses optimal wage

hand, the degree of separation among the agents on the causal feature may become worse if R is too small. The firm sets R by considering the trade-off between these two effects. We assume that R_O and R_T satisfy the conditions in Assumptions 1-3. Specifically, $R_O > \underline{R}$ and $R_T > \underline{R}$, where $\underline{R} = \max(\theta\alpha, \beta + \frac{\theta(1-\lambda)\alpha}{\theta(1-\lambda)+(1-\theta)\lambda})$[14]. This assumption helps us avoid several uninteresting scenarios such as the case where the firm sets R_O or R_T at a very low level such that it does not care about the true type of an applicant. Below, we analyze the firm's optimal choice of R_O and R_T and compare the equilibrium payoff in the opaque and transparent scenarios.

We start with the opaque scenario. Recall that, in Section 4.1, we divided the C_H-C_L space into different regions according to agents' strategies on the causal feature. Now, since R_O can be decreased by the firm, the (x, y) coordinates of any point $(\frac{C_L}{R_O}, \frac{C_H}{R_O})$ in Figure 19a can be increased proportionally. For example, the yellow point in Figure 19a will be moving along the yellow line, and the two ends of the line are defined as $(\frac{C_L}{R_0}, \frac{C_H}{R_0})$ and $(\frac{C_L}{\underline{R}}, \frac{C_H}{\underline{R}})$. As discussed before, we restrict our attention to the case where the firm prefers the transparent scenario under the fixed wage setting (R_0); in other words, we only consider the case where the left end of the line $(\frac{C_L}{R_0}, \frac{C_H}{R_0})$ lands in region $C1$, $C2$, and $C4$. We first study the case where the left end of the line lies in region $C1$ or $C4$. In this case, reducing R_O will not change agents' behavior on the causal feature (i.e., the whole yellow line lies in region 'case 1' defined in Section 4.1). Consequently, the firm will set $R_O* = \underline{R}$, and using the payoff function we have derived in Section 4.1, we have that the firm's equilibrium payoff is $\Pi^O_{C1} = \Pi^O_{C4} = \lambda\theta\alpha - (\lambda\theta + (1-\lambda)(1-\theta))\underline{R}$. We next move to the case where $(\frac{C_L}{R_0}, \frac{C_H}{R_0})$ lands in region $C2$. In this case, if R_O is set lower than $\frac{C_L}{\lambda}$, the yellow point will cross

[14]This assumption can be relaxed and it will not affect the intuition provided by the analysis

the boundary of case 1 and case 4 so agents' strategies on the causal feature will be altered. The optimal point is either at the boundary of case 1 and case 4 or at the right end of the yellow line, depending on which of them gives the firm a higher payoff. That is,

$$R_O* = \begin{cases} \frac{C_L}{R_0} & \text{if } \underline{R} < \frac{C_L}{R_0} \text{ and } \Pi_{firm_{O4}}(\frac{C_L}{R_0}) \geq \Pi_{firm_{O1}}(\underline{R}) \\ \underline{R} & \text{if } \underline{R} \leq \frac{C_L}{R_0} \text{ or } \Pi_{firm_{O4}}(\frac{C_L}{R_0}) < \Pi_{firm_{O1}}(\underline{R}) \end{cases}$$

where $\Pi_{firm_{O1}}(R)$ and $\Pi_{firm_{O4}}(R)$ are the firm's payoff functions as defined in Section 4.1: $\Pi_{firm_{O1}}(R) = \lambda\theta\alpha - (\lambda\theta + (1-\lambda)(1-\theta))R$, $\Pi_{firm_{O4}}(R) = \theta(\alpha+\beta-R)\left(1 - \frac{R(1-\theta)(1-\lambda)}{(\alpha-R)\theta\lambda}\right)$. The firm's optimal payoff in region $C2$ can then be shown as

$$\Pi^O_{C2} = \begin{cases} \lambda\theta\alpha - (\lambda\theta + (1-\lambda)(1-\theta))R_O* & \text{if } R_O* \leq \frac{C_L}{R_0} \\ \theta(\alpha+\beta-R_O*)\left(1 - \frac{R_O*(1-\theta)(1-\lambda)}{(\alpha-R_O*)\theta\lambda}\right) & \text{otherwise} \end{cases}$$

We next move to the transparent scenario. Figure 19b illustrates how $R_T \in [\underline{R}, R_0]$ affects the agents' behavior on the causal feature in the transparent scenario. Similar to our discussions in the opaque scenario, we only consider the case where $(\frac{C_L}{R_0}, \frac{C_H}{R_0})$ falls in $C1$, $C2$, and $C4$. We first consider the case where $\underline{R} \geq C_H$. It is straightforward to see that, in this case, $R_T* = \underline{R}$ regardless of which region $(\frac{C_L}{R_0}, \frac{C_H}{R_0})$ is in. Next, we consider the case where $\underline{R} < C_H$: it can be shown that in this case $R_T* = C_H$ regardless of which region $(\frac{C_L}{R_0}, \frac{C_H}{R_0})$ is in. The intuition is as follows: If the yellow line crosses the border between case 1 and case 2, the optimal point will be exactly on the border because a further decrease in R_T will change the agents' strategies on the causal feature from perfect separation to polling, and the latter will result in a zero payoff. The firm's optimal payoff can then be shown as

$$\Pi^T_{C1} = \Pi^T_{C2} = \Pi^T_{C4} = \begin{cases} \theta(\alpha+\beta-R_T*) & \text{if } R_T* \leq C_L \\ \theta(\alpha+\beta) + (1-\theta)\beta - R_T* & \text{otherwise} \end{cases}$$

We next compare the optimal wage and the maximum profit in the opaque and the transparent scenarios. It is straightforward to see $R_T* \leq R_O*$ if $\underline{R} \geq C_H$. Comparing the optimal payoff functions, we have $\Pi^T_{Ci} \geq \Pi^O_{Ci}, \forall i \in 1, 2, 4$. There is an intuitive explanation behind this result: If we force the firm to choose R_O* (a sub-optimal choice) in the transparent scenario, the firm can still get a better payoff than choosing R_O* (the optimal choice) in the opaque scenario, which means that the firm will get an even better payoff when choosing R_T* in the transparent scenario. We next consider the case $\underline{R} < C_H$. In this case, when $R_T* = C_H$ and $R_O* = \underline{R}$, $R_T* \leq R_O*$ does not hold, and therefore, the above logic does not work. We need to directly compare the firm's payoff in the transparent and the opaque scenarios. Note that in this case $(\frac{C_L}{R_T*}, \frac{C_H}{R_T*})$ falls in

region $C4$, which belongs to transparent scenario 2; $(\frac{C_L}{R_O*}, \frac{C_H}{R_O*})$ falls in region $N1$, which belongs to opaque scenario 1. Using the firm's payoff function shown in Lemma 1 and Lemma 2, we get $\Pi^T_{Ci} = \theta(\alpha + \beta - R_T*)$ and $\Pi^O_{Ci} = \lambda\theta\alpha - (\lambda\theta + (1-\lambda)(1-\theta))\underline{R}$. Since $R_T* < \beta + \theta\alpha$ (according to Assumption 3) and $R_O* = \underline{R} \geq \theta\alpha$ (according to Assumption 2), the following inequalities must hold: $\Pi^T_{Ci} = \theta(\alpha+\beta-R_T*) > \theta(\alpha+\beta-(\beta+\theta\alpha)) = \alpha\theta(1-\theta)$, and $\Pi^O_{Ci} = \lambda\theta\alpha-(\lambda\theta+(1-\lambda)(1-\theta))\underline{R} < \lambda\theta\alpha - (\lambda\theta + (1-\lambda)(1-\theta))\theta\alpha = \alpha\theta(1-\theta)(2\lambda-1)$. Considering the fact that $2\lambda - 1 \leq 1$, we have $\Pi^T_{Ci} > \Pi^O_{Ci}, \forall i \in 1, 2, 4$.