Watching the Watchers: The Credibility of Signaling Social Goodwill with Imperfect Monitoring*

Philippe Mahenc† Alexande Volle‡

March 20, 2019

Abstract

When consumers cannot verify corporate social goodwill, firms may be reluctant to uphold a pledge of social goodwill. We show how imperfect monitoring can mitigate this moral hazard problem.

We augment the standard model of price signaling by allowing consumers to use the results of independent monitoring as a complementary source of information. Before sending a price signal to consumers, firms pledge or not to invest in social goodwill. Monitoring corrects for consumers' arbitrary beliefs.

With no monitoring, firms do not abide by their pledges of social goodwill when they fail to send a credible signal *via* price,

With monitoring, there exist equilibria in which a firm invests in social goodwill and succeeds in signaling its choice *via* price. We conclude that independent monitoring, although imperfect, helps a firm fulfill its pledge of social goodwill by restoring the credibility of price signaling.

Keywords: credence good, monitoring, signaling.

JEL Code: D8, H4, L15, L31, Q5.

*This research received funding from the French Agence Nationale de la Recherche within the framework of the project "GREENGO – New Tools for Environmental Governance: the role of NGOs" (ANR-15-CE05-0008).

†CEE-M, Univ Montpellier, CNRS , INRA, Montpellier SupAgro, Montpellier, France, Avenue Raymond Dugrand, Site de Richter, C. S. 79606, 34960 Montpellier Cedex 2 - France - Email: philippe.mahenc@umontpellier.fr

‡CEE-M, Univ Montpellier, CNRS , INRA, Montpellier SupAgro, Montpellier, France, Avenue Raymond Dugrand, Site de Richter, C. S. 79606, 34960 Montpellier Cedex 2 - France - Email: alexandre.volle@umontpellier.fr

1 Introduction

As surprising as it may sound, firms often claim their pledges for social objectives, whether environmental or ethical. Besley and Ghatak (2007) find evidence of a broader social role for private enterprise in the development of corporate social responsibility. Baron (2010) recognizes that an increasing number of firms engage in the mitigation of externalities, the redistribution of wealth or the provision of public goods; the researcher categorizes all these pro-social activities under the generic term of "self-regulation". A firm may find it worthwhile to champion a social cause, as long as doing so increases the product value to consumers and the consumer surplus that the firm can extract in its self-interest. However, self-regulation can also be motivated by concerns that range from the deterrence of public regulation, as in Lyon and Maxwell (2004), to moral concerns, as in Baron (2010).

Firms that spend resources on self-regulation must in turn communicate information about their social goodwill; otherwise, they may receive no credit for doing so. Firms usually claim social goodwill by displaying their own social labels or relying on external certification, whether public or private. These claims are aimed at persuading consumers that the firm wishes to maintain a balance between its own interest and any social concerns that a catchword has brought to the fore.

There are many examples of ecolabels claiming compliance with good environmental practices. For instance, car manufacturers display "Clean Diesel" to guarantee sustainable business practices. Similarly, the "Forest Stewardship Council" label is meant to reduce illegal logging and improve forest governance. Regarding ethical concerns, certifiers such as Fairtrade International grant the Fair Trade logo to cooperatives and farmers that meet standards promoting sustainable development and more equitable distribution of wealth in developing countries.

However, social goodwill is a product attribute with specific features that makes it hard to verify: it is intangible and public, in the sense that every individual's enjoyment of social goodwill does not preclude its enjoyment by other individuals. Certification may fail to display credible information about the social conscience of a business. The Volkswagen emissions scandal in 2015 has raised doubts on both compliance with environmental standards and the trustworthiness of clean certification in the automobile industry. Similarly, non-governmental organizations have long warned about illegal logging in China, Peru and Romania by companies carrying the Forest Stewardship Council label.[1] The same issue has arisen with Fair Trade certification[2] or the Marine Stewardship Council granted to sustainable fisheries by a nongovernmental organization.[3]

This paper addresses the question of watching the watchers formulated long ago by the

[1] See 20/02/2018 Greenwashed Timber: How Sustainable Forest Certification Has Failed.

[2] Weitzman, Hal (9 September 2006). "The bitter cost of Fair Trade coffee": https://www.ft.com/content/d191adbc-3f4d-11db-a37c-0000779e2340

[3] Jacquet et al. (2010) raises concern about potential conflicts of interest involved in this certification.

poet Juvenal in his Satires: "Quis custodiet ipsos custodes?".[4] We argue that independent monitoring has a key role to play in self-regulation in a context without trustworthy certification of firms' social goodwill. Our model shows that, although imperfect, monitoring helps firms abide by their pledges of social goodwill by ensuring the credibility of the signal sent by firms through prices.

The issue of misleading certification is central to a strand of theoretical literature. Feddersen and Gilligan (2001) question the honesty of third-party certification in a model where a certifier biased toward environmental protection is responsible for sending misleading messages to consumers. Hamilton and Zilberman (2006) show that markets for environmentally-friendly products suffer from fraudulent labeling. Baksi and Bose (2007) demonstrate that some firms make false claims or display spurious labels. Mahenc (2017) shows that third-party certification turns out to be misleading when the certifier is driven more by profit than by social welfare.

Since the work of Darby and Karni (1973), products with a desirable private attribute, whose utility is difficult if not impossible for consumers to ascertain, have been termed "credence" products in the economic literature. In the present paper, we extend this terminology to the public attribute of social goodwill. Like every credence attribute, asymmetric information about social goodwill provides firms with incentives for fly-by-night behaviors that mislead consumers. This may raise a twofold problem, compounded by misleading certification: a problem of adverse selection— products tied in with social goodwill are not attracted to the market because this attribute is hidden from consumers (see Akerlof, 1970)— and a problem of moral hazard— firms may break with their pledges of social goodwill if they are neither observable nor verifiable by consumers. —.

The adverse selection issue has been widely investigated for private attributes of experience goods (Nelson, 1970) in the framework of price signaling models (Milgrom and Roberts, 1986; Bagwell and Riordan, 1991; Mahenc, 2008, Daughety and Reinganum, 2008; Janssen and Roy, 2010). Building on Spence's (1973) pioneering work, these models assume that Nature selects the firms' types, which summarize the private attributes of a product such as quality standards resulting from past investments. If repeat purchases and reputation are not an issue, a firm may want to signal a high-quality product with prices higher than those justified by market power under full information. This is due to the credibility requirement inherent to separating equilibria: the high-quality firm must deter its low-quality counterpart from fooling consumers by mimicking. In some circumstances, the least costly means of preventing the fly-by-night strategy is for the high-quality firm to include the forgone profit from cheating in the product price, thereby incurring further costs for informational purpose. Prohibitive signaling costs may explain why market prices conceal rather than reveal information about quality, and also why firms do not find it profitable to put high-quality products on the market.

The aforementioned examples of fly-by-night behaviors in the automobile and logging

[4]Satire VI, lines 347–348.

industries show that the adverse selection problem also occurs for products tied in with a public attribute such as social goodwill. If consumers cannot rely on trustworthy certification, then products labeled differently depending on firms' claims of social goodwill are sold at the same "pooling" prices, thus making them susceptible to Akerlof's (1970) lemon problem. That pooling prices prevail over separating prices in markets for credence products is a theoretical possibility. As demonstrated by Mahenc (2017), this occurs when the certifier is driven more by profit than by social welfare. Recent experimental field and lab studies provide evidence of fraud in markets for credence goods and services that are mainly private in nature (Kerschbamer and Sutter, 2017). The impossibility of credible signaling is in turn a disincentive for firms to tie social goodwill in with products, even though consumers would be willing to pay a premium for it. There is a moral hazard problem behind the adverse selection problem. Anticipating that signaling cannot be achieved, firms do not find it worthwhile to spend resources on social goodwill.

To address the moral hazard issue, we extend the standard signaling model of asymmetric information with hidden knowledge to a model of symmetric information with hidden actions, in which imperfect monitoring by a third-party is allowed. Hidden actions are the technology decisions made by two firms whether to stick to business as usual or to switch to production with an added value of social goodwill that is unverifiable to consumers. In other terms, we consider that firms select their type before they send a price signal to consumers. The choice of type determines the degree of differentiation between two products available for consumption and hence the intensity of subsequent competition in price between firms. Bayesian consumers try to infer from the prices set by firms whether or not social goodwill is tied in with the product they are purchasing.

First, we examine a reduced version of the model, in which certification is not trustworthy and price signaling is the sole source of information for consumers. In this scenario, a firm that would have invested in social goodwill fails to prevent fly-by-night behavior in the form of misleading prices with the result that there exists no separating equilibrium in price. Faced with the inability to signal social goodwill, firms renege on any pledge to provide the credence attribute. Thus, an additional and complementary source of information is needed for consumers concerned with social goodwill.

We extend the model by allowing for imperfect monitoring activities performed by an independent third-party auditor. Consumers make purchase decisions using both the auditor's report and firms' price signaling to gather information. We assume that monitoring improves consumers' perception of social goodwill by correcting for the arbitrary beliefs held after observing a deviation from the price equilibrium.

Our main result is that monitoring, albeit imperfect, allows the existence of price separating equilibria. The reasons for this are that monitoring, on the one hand, raises the cost of cheating consumers with misleading prices, and on the other hand, motivates firms to reveal the truth about their type. As monitoring makes credible the price signals firms send about their type, tying social goodwill in with the product becomes worthwhile. As a result, there

exists a reasonable probability that a firm switches to production with a pledge of social goodwill that can be fulfilled. In a mixed strategy equilibrium, this probability coincides with the probability that a rival firm mimics the price sent as a signal of social goodwill. We conclude that the auditor's monitoring helps solve the two problems firms encounter: the adverse selection problem raised by signaling costs, and the moral hazard problem of pledging social goodwill.

The rest of this paper is organized as follows. Section 2 discusses a brief review of the literature. Section 3 introduces the model of symmetric information with firms' hidden actions in which consumers have two sources of information: independent monitoring and firms' price signaling. Section 4 characterizes the possible Perfect Bayesian Equilibria of the whole game called the "signaling game with hidden actions". In Section 5, we analyze a baseline model with no monitoring and we establish the conditions under which there exists no separating equilibrium in the signaling subgame. In Section 6, we analyze the full scenario with monitoring and we show the existence of equilibria that involve separation in the signaling subgame. Our concluding remarks appear in Section 7.

2 Related literature

Our models builds on the signal theory initiated by Spence (1973) and further developed in a strand of literature on asymmetric information devoted to the analysis of markets in which firms use price as the only means of signaling product quality. Usually, these markets are imperfectly competitive because firms must have enough control over prices to influence the amount of information disclosure. Signaling through prices can occur in monopolistic markets (as in Bagwell and Riordan, 1991), as well as in oligopolistic markets (see Daughety and Reinganum, 2008; Janssen and Roy, 2010).

In these models, firms, like the worker in Spence's job-market model, are privately informed about their type. Generally, type refers to an economic ability— firms' product quality or the worker's productivity —resulting from past activities that remain outside the scope of the model. Hence, several researchers have taken type to be exogenously given. For simplicity, they have often distinguished between a "good" and a "bad" type that differ in their incentives to reveal the truth about their type. A standard feature of these models, known as the single-crossing property, guarantees that the good type is more willing than the bad type to use price as a costly signal. Firms earn a higher profit margin with lower quality, regardless of whether the market is monopolistic or oligopolistic. Therefore, firms of higher quality are less afraid of losing consumers by raising price to signal their type. This property is necessary for the existence of separating equilibria.

We take this signaling approach one step further by allowing firms to choose their type. As with product quality, there is vertical differentiation between the two product variants obtained with and without social goodwill in the present setting. The choice to tie social goodwill in with the product differentiates the good type from the bad one. Another difference

from models of price signaling quality is that our model combines two sources of information: price (endogenous) and imperfect monitoring (exogenous).

One important insight in the signaling literature is that the bad and good types experience two different trade-offs related to cheating and truth telling, respectively. A credibility constraint reflects the trade-off faced by the bad type. A bad firm may be tempted to use the fly-by-night strategy of setting the same price as the good type to trick uniformed consumers into mistaking the bad type for its good counterpart. However, by doing so, the bad type sacrifices the profit obtained by revealing the truth in the separating equilibrium. Usually, the credibility constraint is binding in the least costly separating equilibrium outcome on which the literature has focused a great deal of attention. Credibility sometimes requires that the good type incurs a positive signaling cost by distorting the price upward relative to the full information case. This signaling distortion may be a burden for both a monopoly (Bagwell and Riordan, 1991) and an oligopoly, albeit to a lesser extent (Daughety and Reinganum, 2008; Janssen and Roy, 2010). Mahenc (2008) obtains the same result when the product quality is environmental in nature: an upward distortion in the monopoly price is needed to signal cleaner products, unless cleaner products are cheaper to produce.

In the present setting, price distortion may restrict sales to such an extent in the absence of monitoring that the good type refrains from deterring fly-by-night strategies. It turns out that monitoring is needed for prices to be informative. We also show that monitoring reduces the size of the signaling distortion; this reduction increases as the accuracy of monitoring improves.

There is also a trade-off for the good type firm when revealing the truth. Although the signaling price yields more profit from uninformed consumers, the good type also foregoes the profit earned in the best worst-outcome in which consumers mistake it for the bad type. This trade-off is formalized by the individual-rationality constraint of the good type. In the present setting, the profit accruing to the good type in the best worst-outcome depends on the probability that the bad type is cheating. We show that the good type succeeds in signaling social goodwill provided that the bad type is not too likely to cheat consumers.

The price signaling literature leaves open the moral hazard issue raised by the failure of signaling; firms may refrain from making the good-type-specific investment if prices provide no evidence for such an investment. In the present setting, we further examine the decision of whether to incur the sunk costs needed to be the good type, assuming that this decision is unobservable. For this, we consider that the probability of cheating at the time of signaling is exactly the probability of choosing the good type one step backward, in a mixed strategy over the set of types. In backward induction, the existence of a mixed strategy equilibrium requires that the probability of choosing the good type allows the firm first to overcome the opportunity cost of signaling this type (the individual-rationality constraint) and second to prevent the bad type from cheating (the credibility constraint). We find that there exists such a mixed strategy equilibrium, as well as two pure strategy equilibria in which either firm invests in social goodwill.

Dynamic models of quality premia (Klein and Leffler, 1981; Shapiro, 1983) highlight the role played by repeat purchases and reputation in mitigating the moral hazard problem. In these models, firms choose both quality and price every period of an infinite horizon, while consumers cannot observe quality directly. However, the price and quality provided by firms in the past serve as signals of current quality. If quality has been observed to be high in previous periods, then consumers infer that firms are more likely to provide high quality currently. Moreover, consumers punish undue rents for low-quality products by ceasing to purchase them. Hence, firms have an incentive to develop a reputation for high quality. For this mechanism to work in equilibrium, high quality must command a rent, i. e., a quality premium associated with a high price, such that the cost of losing repeat purchases exceeds the cost savings of cheating consumers. Our analysis deals with credence goods rather than experience goods. Therefore, we abstract from the reputation mechanism to focus on the role of imperfect monitoring in inducing firms to honor promises of high quality. Another difference from Klein and Leffler (1981) is that, in our analysis, competition between firms is imperfect, while it is perfect in their analysis.

Daley and Green's (2014) research is closely related to the present paper. These authors investigate the consequences for costly signaling of using grade as another instrument for information transmission about the sender's type in Spence's (1973) canonical model. In their model, the informed party is a worker who relies both on his education level and his performance on a test (a grade) to signal his ability to potential employers. The authors demonstrate that the presence of sufficiently informative grades dismisses separating equilibria as being less plausible than pooling equilibria. As usual, the worker's choice of education level serves as a signal that allows employers to update beliefs. These beliefs in turn influence the amount of information conveyed by the grades, which are used by employers as a redundant signal about the worker's ability. A high-ability worker finds signaling *via* education less costly than does a low-ability worker. However, above a certain level of education, the information conveyed through grades is more beneficial to the low-ability worker. This deters the high-ability worker from investing further in education to fully reveal information, and thus pooling prevails over separating. In the context of Daley and Green (2014), it is reasonable to assume that the use of education as a signal affects the information conveyed through grades because the caliber of grades obtained by a worker is likely to depend on the quality of that worker's past performance as a student. In contrast, there is no reason to assume that the two instruments available for information transmission are interlinked in the context of our paper. Rather, we assume that the information conveyed by the auditor's monitoring is not affected by the price signals sent by firms to formalize the idea that the two sources of information are independent from each other. Unlike grades in Daley and Green (2014), monitoring in our setting is used by the informed party to correct for arbitrary perceptions in case of deviation from the equilibrium path. These perceptions improve when the auditor's report is more accurate because information released in such a way is complementary rather than redundant. Monitoring can be seen as a refinement of Bayesian equilibria in that

it reduces the leeway in specifying off-the-equilibrium path beliefs. Our existence result of separating equilibria due to the presence of monitoring starkly contrasts with the prevalence of pooling equilibria due to grades in Daley and Green (2014). This is because independent monitoring strengthens the incentive of the good type to deter cheating with a costly signal, while grades weakens this incentive in the context of education.

3 The model

The protagonists in the economy are two firms, consumers and an independent third-party auditor. Both firms are price-takers for a conventional product. This product is a bundle of observable characteristics that can be properly verified by inspection and therefore are perfectly known to consumers. The product value can be enhanced by the social goodwill a firm chooses to tie in with the product. This is a new attribute that is public in nature and reflects the firm's concern for the social and environmental externalities of its economic activity. The new attribute has the following features:

(i) it is vertical in the sense that, everything else being equal, all consumers agree that the product is more valuable with than without the attribute;

(ii) it entails additional costs for the firm that chooses to invest in all the factors needed to provide the attribute, including labor, managers and capital;

(iii) consumers cannot directly observe whether the product has or has not the attribute, either prior to or subsequent to consumption;

(iv) a specific label is intended to guarantee that the product has the new attribute, but this certification may not be credible.

Although the spectrum of externalities associated with the public attribute encompasses all kinds of social concerns, political, ethical and environmental, we consider that the new attribute is environmental for the sake of illustration. Hence, we refer to the product with the new attribute as being the "green" type, and to the conventional product as being the "brown" type.

The market for the brown product is in a long-run competitive equilibrium, representing business as usual. A firm can command a price premium over its product by switching to green production. However, the decision of whether to switch is unobservable and firms use prices only to signal their type. The brown rival in turn may counteract both product differentiation and information disclosure with the fly-by-night strategy of tricking consumers into believing it is also selling a green product. Our goal is to concentrate on this strategic interaction. The absence of strategic behaviors within the brown industry avoids the complexity of simultaneous signaling recognized by Mailath (1988). We assume that the brown

technology has constant return to scale, i. e., firms have access to perfectly elastic supplies of all the factors needed for brown production.[5]

Switching from brown to green production requires investment in real and financial assets, which involves some fixed setup cost $F > 0$[6] and an additional marginal cost $c(e) = e$, where e is an indicator of the extent to which the product value is increased by the new attribute.

If either firm chooses to stick to the brown type of product, it will be indexed by $t = b$, or else by $t = g$ if it invests in the green type. The firm $i = 1, 2$ of type $t \in T = \{b, g\}$ sells its product at price p_{it}.

The total number of consumers is normalized to unity. Each consumer has exogenous wealth of w and purchases at most one unit of either type of the product. Consumers have heterogeneous preferences that differ according to a taste parameter x for the new attribute, which is assumed to be continuously and uniformly distributed over the interval $[0, l]$. Thus, the indirect utility of a consumer with taste x for the attribute valued e_t, who purchases one unit of the type-t product at price p_{it}, is given by

$$V_t\left(p_{it}, e, x\right) = w + xe_t - p_{it}, t \in T, i = 1, 2, \tag{1}$$

where $e_g = e$ and $e_b = 0$.

Demands.—Market demands are determined from a critical value of the taste distribution. The preference level of the consumer who is indifferent between purchasing the brown product and its green substitute is found by setting

$$w + xe_g - p_{ig} = w - p_{jb} \tag{2}$$

and solving for x. Doing so yields

$$X = \min\{\frac{p_{ig} - p_{jb}}{e_g}, 1\}. \tag{3}$$

All consumers with values of x that satisfy $x \geq X$ purchase the green product and the remaining consumers purchase the brown product. This implies linear demand functions for both products. Generally, the demand for firm i of type g, when it charges price p_{ig} and the rival firm j charges p_{jg} will be denoted by $D_{ig}(p_{ig}, p_{jg})$. This function can take two different

[5]For instance, we could use the model of monopolistic competition developed by Salop (1979) to formalize business as usual in our setting. Under this framework, the market for the brown product is represented by a circle of unit length on which consumers are uniformly distributed. Each point on the circle corresponds to one consumer's most preferred variant of the brown product. A large number of firms are symmetrically located around the circle. Each firm produces a single variant of the brown product with an identical linear technology. There is free entry into the brown market so that firms continue to enter until profits are driven to zero. If t is the transport cost per unit of distance from an ideal location, c is the constant unit product cost and f is the fixed cost of entry, the market price for the brown product in the symmetric equilibrium is $p_b = c + \sqrt{tf}$. These additional variables would make calculations more cumbersome and risk clouding the issue.

[6]F includes unsalvageable expenditures in developing the new attribute and specific entrepreneurial skills, sunk investments in specialized machinery for green production, fuel-saving equipment or long-term rental contracts that cannot be resold, as well as the fee paid to have the product certified.

forms depending on whether firm j is brown or green (see Appendix 1).

The profit earned by firm i in the green market is

$$\pi_{ig}\left(p_{ig}, p_{jb}\right) = \left(p_{ig} - e\right) D_{ig}(p_{ig}, p_{jg}). \tag{4}$$

One possibility in the green market equilibrium is that the cost of providing the attribute is so high that consumers have zero demand for the green product, even when it is sold at marginal cost. Let p_b^c denote the market price in the long-run competitive equilibrium. We substitute p_{ig} and p_{jg} for e and p_b^c, respectively, into (45) given in Appendix 1. As a result, the following inequality ensures that demand for the green product is strictly positive when it is sold at marginal cost

$$\left(l - 1\right) e + p_b^c > 0. \tag{5}$$

Roughly, this assumption says that consumer heterogeneity is sufficient for the existence of a green market.

Till now, we have presented the full information framework in which the new attribute is observable. We now turn to the issue of information transmission over the sale period in the green market. We assume that green certification is not trustworthy due to "greenwash" or "launder" trafficking in illegal products. Consumers are unsure whether a product carrying the seal "green" truly has the new attribute. However, consumers are likely to form perceptions about the type of the product they are purchasing, either by observing prices, or by reading reports and test results about the value of the green product, released by the third-party auditor.

The third-party auditor.—The auditor has the technology to monitor whether the product in the green market really has the new attribute. The monitoring is achieved by observing differences between the firm's claims and the truth about its type. For example, the auditor measures all polluting emissions generated by the product and compares them with the standard required to be green: if the polluting emissions are observed to be lower than the standard, then the product passes the test. The auditor's monitoring is not perfect: it succeeds in learning the true type of a firm with the probability $\alpha, 0 \leq \alpha \leq 1$; thus, α is a parameter that indicates the accuracy of the monitoring technology.

The auditor's observation takes place over the sale period simultaneously with firms' pricing. Moreover, firms and the auditor are independent players, so they do not influence each other regarding their information transmission. At the end of the monitoring process, the auditor releases a report that supplements the information consumers infer upon observing firms' prices. Consumers make their purchase decision using the two sources of information: the auditor's report and firms' price signaling.

The moral hazard issue.—Bayesian consumers form perceptions of a firm's type based on observed prices. This is a standard signaling issue in the spirit of Spence (1973), that has been widely investigated using incomplete-information models. In these models, Nature selects the types of the signal sender according to some exogenous probability distribution.

The type usually refers to an economic ability resulting from past investment made by the signal sender. For instance, the good type in signaling models of quality represents firms eager to truthfully signal quality because they have previously chosen to provide improved quality, which is both more costly to produce and more valued by consumers. What is hidden in signaling games is information about the firm's type, which implicitly assumes that the firm's exogenous choice of becoming the good type was not observed by the other players of the game.

Rather, in the present framework, we investigate firms' quality choice when this decision is unobservable to the other players, i. e., consumers, the firm's rival and the auditor. For this, we allow firms rather than Nature to select their type and we tackle the problem of moral hazard raised by firms' hidden actions.

The main questions are: Is it worthwhile for a firm to switch to green production if the switch is not observable? What if the expected costs of signaling the switch *via* prices are prohibitive? Does the auditor's monitoring help the industry to convert toward green production under these circumstances?

To answer these questions, we transform the signaling model of asymmetric information with hidden knowledge into a model of symmetric information with hidden actions. Unobserved actions are the technological decisions of whether to switch to green production, and the observed actions are both the prices set by firms and the report released by the auditor. A standard assumption in signaling games is that there exists a fraction σ of good-type firms versus a fraction $1-\sigma$ of bad-type firms, and this exogenous distribution is common knowledge in the economy. We reinterpret this probability distribution over the set of types as a mixed strategy over the set of pure strategies for either type of firm in our model. This mixed strategy captures public uncertainty about what a firm does regarding its type. Hence, the distribution of types must emerge as the (Nash) equilibrium randomization of a firm's decision to switch to green production or to stick to brown production.

The timing.—The whole game is a three-stage game that proceeds as follows:

(i) In the first stage, the two firms simultaneously choose their types. If a firm decides to switch to green production, it pays the setup cost F.

(ii) In the second stage, firms simultaneously post their prices and the auditor releases its report. Consumers observe these actions.

(iii) In the third stage, consumers update their beliefs about the firms' types upon seeing prices, supplement this information with that conveyed by the auditor's report and, finally, decide from which firm to buy. Consumers use Bayes' rule whenever possible to form posterior beliefs from the observed prices.

The game begins with the technology decisions made by firms regarding their type: each firm $i = 1, 2$, selects a type t from the set $T = \{b, g\}$ and may randomize over these pure strategies. Firms are committed to their technology decisions until the end of the game: they

will not change their types in the next stages because the appropriate production facilities have a second-hand value lower than their initial value. The technology decision is private information to each firm: its type is unknown to every other protagonist— the rival firm, consumers and the auditor —This implies that the setup cost F specific to the green production is not observable either. A mixed strategy for firm i, $\sigma_i : T \to [0, 1]$, assigns probability $\sigma_i(t)$ that it chooses the type t, where $\Sigma_{t \in T} \sigma_i(t) = 1$. This probability distribution summarizes all the information publicly available to all but the perfectly-informed firm i at the end of the first stage.

In stage 2, the setup cost F is sunk, if ever. The probability distributions $\sigma_i(t), i = 1, 2$, are the only statistics for past play; hence they become the prior beliefs of the uninformed protagonists at the beginning of stage 2**.** With these beliefs in mind, firms charge prices $p_{it} \geq 0, i = 1, 2$. The sequential order between the decisions regarding technology and price captures the notion that a price can in practice be varied at will, unlike the commitment to a type.

It may happen that one type of product is unavailable on the market if both firms decide to supply the other type. In that event, Bertrand competition in the single product market fully reveals information about the firms' type. Asymmetric information is a trickier issue in the event that two differentiated products are available on the market. The subgame starting at every "informational node" involving two distinct types works like a standard signaling game, in which the probability distributions $\sigma_i(t), i = 1, 2$ are common knowledge. In other words, $\sigma_i(t)$ is the probability assigned by everyone but firm i to the event that firm i has chosen type t, given the presence of two differentiated products on the market. In particular, firm i's rival will use this probability distribution to predict how firm i price discriminates between the two potential types in a separating equilibrium. Firms' unobserved randomization over types requires that subsequent signaling *via* prices be credible: in equilibrium, the green-type firm must deter its brown-type counterpart from cheating about its type.

In stage 3, consumers draw inferences about the firm's types and cross-check this information with that released by the auditor's report. Finally, consumers make their purchase decisions to maximize their expected payoffs, given their posterior beliefs. The payoff to a consumer is her expected net surplus if she buys, and zero otherwise. The payoff to each firm is its expected profits.

We can focus on firms' strategic interplay in the green market due to the assumption of zero profits (perfect competition) in the brown market. We require that strategies in the green market form a Perfect Bayesian Equilibrium (PBE); that is, strategies must yield a Bayesian equilibrium not only for the whole game, but also for every subgame, including that starting after any possible choice of a type made by firms.

4 The signaling game with hidden actions

We solve the game by backward induction. For this, we first investigate the signaling subgame starting at every informational node where the market is segmented between two differentiated products and the setup cost specific to green production is sunk. When firms get to move at the second stage of the game, every protagonist other than firm i believes that it has chosen the type t with probability $\sigma_i(t)$.

The signaling issue in stage 2 raises the familiar problem of multiplicity of equilibria. To simplify, our analysis of price signaling focuses on pure-strategy separating equilibria in which the brown and the green types choose different prices. Separating prices in equilibrium ensure that green certification is credible by truthfully disclosing information about the actual types. Following Mahenc (2017), the credibility of green certification requires separating prices in the green market, and conversely, pooling prices (in which the firm's price is independent of its true type) undermine the reliability of certification. Were prices to pool in the green market, information would be concealed by market prices, making information disclosed by the green label inconsistent.

In stage 2, firms' pure strategies are vectors of four prices $\{(p_{it})_{i=1,2,t\in T}\}$. Given that firms are completely symmetric at the beginning of the whole game, we assume that firms employ the same pricing rule in equilibrium, and so, $p_{1t} = p_{2t} = p_t$ for each $t \in T$. Every protagonist other than firm i makes the prediction that it will employ the pricing rule ρ_i phrased as follows: "firm i will charge price p_b with probability $\sigma_i(b)$ and price p_g with probability $\sigma_i(g)$", where $\sigma_i(b) = 1 - \sigma_i(g)$. We can simplify the notation $\sigma_i(g)$ and write σ_i instead.

In stage 3, consumer perception of the firms' types builds on information from two sources: firms' prices and the auditor's report. We assume that the information released by firms through price setting does not influence that released by the auditor's monitoring. Therefore, we treat the two events {the firm is green conditional on observing prices, the firm is green conditional on reading the auditor's report} as being both independent and mutually non-exclusive.[7]

To formalize consumer perception of a firm's type based on price, we define a posterior belief function $\mu(t,p) : T \times R^+ \to [0,1]$, that specifies the probability assigned to either firm of being type t in response to a price p observed in the market for the green product. This belief function is the same function for both firms. Along the equilibrium path, consumers use Bayes' rule to update beliefs from the prior distributions $\sigma_i, i = 1,2$, available at the beginning of the signaling game. When consumers observe a price off the equilibrium path, they update beliefs with an arbitrary rule instead of Bayes' rule.

[7] Let A be the event that the firm is green after observing prices, and B the event that the firm is green based on the auditor's monitoring. The probabilities of A and B are $\mu = \Pr(A)$ and $\alpha = \Pr(B)$, respectively. If A and B are independent events, then the joint probability of both occurring is $\Pr(A \text{ and } B) = \Pr(A)\Pr(B)$. If A and B are not mutually exclusive events, then the probability of either occurring is $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$. Thus, if A and B are both independent and mutually non-exclusive events, the probability that the firm is green after observing either prices or monitoring is $\Pr(A \text{ or } B) = \mu + \alpha - \mu\alpha$.

We write $e(\mu) = \Sigma_{t\in T}\mu(t,p)e_t$ for the expected value consumers infer from price p. Consumers supplement their information about a firm's type with the auditor's report. Assuming that firms and the auditor have no influence on each other in providing information, we describe the overall formation of consumer perception by the equation

$$\widetilde{e}_t(\mu) = \alpha e_t + (1-\alpha)e(\mu). \tag{6}$$

Equation (6) gives consumers' expected valuation for a product of type t sold in the green market, given consumer final beliefs. If the auditor's observation provides no information, then consumer perception relies only on inferences from prices. If the auditor's monitoring is perfectly accurate, then consumers learn the true type. Note that equation (6) applies both on and off the equilibrium path, i. e., regardless of whether consumers use Bayes' rule or an arbitrary rule to update beliefs after observing prices. When Bayes' rule applies in equilibrium to beliefs formed from pure-strategy prices (p_b, p_g), posterior beliefs are $\mu(g,p_t) = 1-\mu(b,p_t) = 1$ for $t \in T$, giving consumers correct perceptions whatever the type, since

$$\widetilde{e}_t(\mu(t,p_t)) = \alpha e_t + (1-\alpha)e\mu(t,p_t) = \begin{cases} e \text{ if } t=g, \\ 0 \text{ if } t=b. \end{cases} \tag{7}$$

Hence, consumers align their readings of the auditor's report with updated beliefs when prices reveal the truth. Otherwise, consumers observe a deviation from price equilibrium. Then, given an arbitrary belief μ based on a probability-0 price, consumer perception is

$$\widetilde{e}_t(\mu) = \alpha e_t + (1-\alpha)\mu e = \begin{cases} [\alpha + (1-\alpha)\mu]\, e \text{ if } t=g, \\ (1-\alpha)\mu e \text{ if } t=b. \end{cases} \tag{8}$$

The motivation for this is that consumers use the information provided by monitoring to correct for arbitrary perceptions when something "surprising" occurs. In the extreme case where beliefs based on prices are pessimistic ($\mu = 0$), even if the actual type is green, monitoring reduces the leeway in specifying off-the-equilibrium path beliefs. Moreover, the perception consumers have about the type, i. e., $\widetilde{e}_g(0) = \alpha e$, improves with monitoring accuracy. In some sense, monitoring helps consumers refine the multiplicity of equilibria supported by unrestricted beliefs based on unexpected prices.[8]

[8]The way monitoring restricts beliefs in our setting clearly differs from how beliefs update based on grades in Daley and Green's (2014) paper. These authors assume that how beliefs update after observing a signal and grades follows a two-stage process. In the first stage, receivers observe a signal through the investment in education, and they use Bayes' rule to update beliefs, as consumers do after observing prices in the present paper. This first updating results in "interim beliefs", based on the history h^1 of the game in the first stage. In the continuation game starting at the second stage, receivers observe grades and again update beliefs from their interim beliefs *via* Bayes' rule. This requirement is stronger than simply using Bayes' rule in the usual fashion since it applies to updating from the first stage to the second stage, whether or not h^1 has probability 0, and whether or not the signal sent in the first stage has probability 0. Hence, the likelihood of grades implicitly depends on the observation of prices *via* the second Bayesian updating. Applying Bayes' rule twice captures a form of redundancy in the information transmission, which is not desirable in the context of our paper. The sender influences the decision to give a grade by sending a costly signal in Daley and Green (2014). Rather, we assume that price signaling does not interfere with monitoring.

Finally, consumers make their purchase decisions to maximize their expected payoffs, in accordance with (6).

From now on, we will simplify the notation $\mu(g,p)$ and write $\mu(p)$ instead, so that $\mu(b,p) = 1-\mu(p)$. The market split between products depends on consumer conjectures about the firms' types. Consider the following: (i) firm i of type t is perceived by consumers to be green with probability μ; (ii) the rival firm j of type t' is perceived to be green with probability σ; and (iii) the expected valuation of the product of firm i exceeds the expected valuation of the product of firm j; that is, $\widetilde{e}_t(\mu) > \widetilde{e}_{t'}(\sigma)$. Then, the critical value of the taste distribution that determines market demands becomes

$$\widetilde{X}(\mu,\sigma) = \min\{\frac{p_{it}-p_{jt'}}{\widetilde{e}_t(\mu)-\widetilde{e}_{t'}(\sigma)}, 1\}. \tag{9}$$

We denote by $D_{it}(p_{it},p_{jt'},\mu,\sigma)$ the demand for the product sold by firm i in the green market resulting from the market split at (9).[9] Clearly, this demand depends on whether consumers have an expected valuation for firm i's product that is higher or lower than that for firm j's product.

A firm's profit can be written as a function of its true type, its perceived type and its price, given the perceived type and the price of its rival. We denote the profit for firm i of type t, when it charges price p_{it}, its perceived type is green with probability μ and the rival firm j of type t' is perceived to be green with probability σ and charges $p_{jt'}$, by

$$\pi_{it}\left(p_{it},p_{jt'},\mu,\sigma\right) = (p_{it}-e_t)D_{it}(p_{it},p_{jt'},\mu,\sigma), \text{ for } t,t' \in T, i = 1,2. \tag{10}$$

In the price signaling subgame, a pure-strategy separating PBE consists of a set of price strategies and beliefs $\{(p_t^*)_{t\in T}, \mu(p_t^*)\}$ such that $p_b^* \neq p_g^*, \mu(p_b^*) = 0$ and $\mu\left(p_g^*\right) = 1$; that is, consumer perception of the firms' types is correct after observing separating prices. The separating pricing rule used by firm i in equilibrium is denoted by ρ_i^*; it predicts: "firm i will charge the equilibrium price p_b^* with probability $1-\sigma_i$ and the equilibrium price price p_g^* with probability σ_i". Furthermore, we know that p_b^* is given by p_b^c in the long-run equilibrium.

Generally, we define $E[\pi_{it}(p,\mu)/\rho_j]$ as the expected profits for firm i of the actual type t, perceived to be green with probability μ after observing the price p, where σ_j is firm i's prediction about firm j's separating pricing rule. In particular, $E\left[\pi_{ig}\left(p_g^*,1\right)/\rho_j^*\right]$ is firm i's expected profits resulting from a pure-strategy PBE in the price signaling subgame. In the event that firm j charges p_b^*, firm i's product sold at price p_g^* in the green market is more valuable to consumers than the rival product. Alternatively, if firm j charges the same price p_g^* as does firm i, then consumers perceive the rival products as the the same. In both events, firm i's demand is calculated from the market split (9) by substituting $\mu = 1$ and $\sigma = 0$ (resp. 1) in the former (resp. latter) event. So, we can write

$$E\left[\pi_{ig}\left(p_g^*,1\right)/\rho_j^*\right] = (1-\sigma_j)\,\pi_{ig}(p_g^*,p_b^*,1,0) + \sigma_j\pi_{ig}(p_g^*,p_g^*,1,1). \tag{11}$$

[9]The explicit form of the demand function is given in Appendix 1.

So far, we have restricted our attention to the signaling subgame starting with prior beliefs σ_j that reflect the common prediction about the price selection made by firm j' for its product. This information is derived from that publicly available at the end of the first stage of the game, in which each firm chooses a mixed strategy from the set of probability distributions over T. The randomization over types summarizes public uncertainty about what each firm does at the first stage. As there is nothing that can alter beliefs between the end of stage 1 and the beginning of stage 2, we assume that the information summarized by the mixed strategies over types is the same as the information used to make subsequent predictions about firms' pricing. Firm i's mixed strategy is the belief that its rival will play the pure strategies $\{b, g\}$ with the probabilities $(1-\sigma_j, \sigma_j)$. Whatever firm i might think about the rival's choice in the first stage, firm i's expected profits from sticking to brown production is zero since the market for the brown product is in a long-run competitive equilibrium. Hence, $E\left[\pi_{ib}\left(p_b^*, 0\right)/\rho_j^*\right] = 0$.If firm i now chooses to switch to green production, it pays the setup cost F and earns expected profits

$$E\left[\pi_{ig}\left(p_g^*, 1\right)/\rho_j^*\right] - F. \tag{12}$$

Firm i's expected profits from playing a mixed strategy in stage 1 are the weighted sum of the expected profits for each of the pure strategies $\{b, g\}$, where the weights are the probabilities $(1-\sigma_i, \sigma_i)$

$$\begin{aligned} \Pi_i\left(\sigma_i, \sigma_j\right) &= \left(1-\sigma_i\right) E\left[\pi_{ib}\left(p_b^*, 0\right)/\rho_j^*\right] + \sigma_i\left[E\left[\pi_{ig}\left(p_g^*, 1\right)/\rho_j^*\right] - F\right] & (13) \\ &= \sigma_i\left[\left(1-\sigma_j\right)\pi_{ig}(p_g^*, p_b^*, 1, 0) + \sigma_j \pi_{ig}(p_g^*, p_g^*, 1, 1) - F\right]. & (14) \end{aligned}$$

Our goal is now to characterize firm i's mixed strategies given by the probabilities $(1-\sigma_i^*, \sigma_i^*)$ of choosing a type from T, and separating prices $\left(p_b^*, p_g^*\right)$, as a subgame PBE of the whole game. In any subgame PBE, no firm must want to change its decision about its type given the decision made by the other firm about its type. This requirement must be met when firms anticipate that prices will subsequently reveal their true types. The mixed-strategy profile $\left(\sigma_i^*, \sigma_j^*\right), i \neq j$ is a Nash equilibrium in which firm i chooses to switch to green production if and only if σ_i^* is a best response to firm j's equilibrium mixed strategy $\left(1-\sigma_j^*, \sigma_j^*\right)$ under the individual-rationality constraint of non-negative profits from green production; that is, for every $\sigma_i \in [0, 1]$,

$$\Pi_i\left(\sigma_i^*, \sigma_j^*\right) \geq \max\{0, \Pi_i\left(\sigma_i, \sigma_j^*\right)\}. \tag{15}$$

Indeed, if $\Pi_i\left(\sigma_i, \sigma_j^*\right) < 0$, then firm i chooses the brown type and earns zero profit, which is strictly better than the loss in profit entailed by paying the setup cost F to become green. From (14), inequality $\Pi_i\left(\sigma_i, \sigma_j^*\right) \geq 0$ determines an upper bound for firm j's equilibrium probability of choosing the green type, which must satisfy

$$F \leq E\left[\pi_{ig}\left(p_g^*, 1\right)/\rho_j^*\right]. \tag{16}$$

Once the setup cost of switching to green production is sunk, firms compete for business. Either firm selling its product on the green market maximizes its expected profit with respect to price, given the auditor's report and given the consumer beliefs function. Beliefs are formed from equilibrium prices and the auditor's report, using Bayes' rule for prices with positive probability, and an arbitrary rule otherwise. This is formalized in the following definition. The set of mixed strategies in type, pure strategies in price and beliefs formed from prices $\{(\sigma_i^*)_{i=1,2}, (p_t^*)_{t\in T}, \mu(p_t^*)\}$ is a PBE if the following four conditions are satisfied:

1. The market for the brown product is in a long-run competitive equilibrium

$$p_b^* = p_b^c. \tag{17}$$

2. Consumers form posterior beliefs from prior beliefs $\sigma_i, i = 1, 2$ using Bayes' rule

$$p_g^* \neq p_b^* \text{ and } \mu\left(p_g^*\right) = 1. \tag{18}$$

3. Prices in the green market are optimal given the separating pricing rule ρ_j^*, consumer beliefs and the auditor's report

$$p_g^* = \arg\max_p E\left[\pi_{ig}\left(p, 1\right)/\rho_j^*\right], i \neq j. \tag{19}$$

4. The mixed strategy profile (σ_1^*, σ_2^*) is a Nash equilibrium under the constraint of non-negative profits from green production

$$\sigma_i^* = \arg\max_\sigma \Pi_i\left(\sigma, \sigma_j^*\right) \text{ subject to } (16), i = 1, 2. \tag{20}$$

As previously mentioned, separation can occur in the price signaling subgame only if one firm is green and the other is brown. Henceforth, we consider, without loss of generality, that firm 1 has chosen to be green and firm 2 has chosen to be brown. With this convention, (6) says that consumer expected valuation for firm 2's product is $\widetilde{e}_b(\mu(p_b^*)) = 0$ in equilibrium. This yields the expected profit $E[\pi_{2b}(p_b^*, 0)/\rho_1^*] = 0$, given the prediction that firm 1 employs the pricing rule ρ_1^*.

However, it might be profitable for firm 2 to mimic its green-type rival, given the prediction that firm 1 charges the price p_b^* with probability $1-\sigma_1$. In that case, consumers will definitely perceive firm 1 as being brown. Then, firm 2 has the leeway to set the price p_g^* upon which consumer perception of its product is $\widetilde{e}_b(\mu\left(p_g^*\right)) = (1-\alpha)e$, from (6). By claiming that its product is green, firm 2 can differentiate it from the brown product. Obviously, this claim is worthless if firm 1 also charges the price p_g^*, which occurs with probability σ_1. Indeed, consumers will then correctly value the product sold by firm 1, inferring that $\widetilde{e}_g(\mu\left(p_g^*\right)) = e$, which strictly exceeds $\widetilde{e}_b(\mu\left(p_g^*\right)) = (1-\alpha)e$ for $\alpha > 0$. It turns out that the auditor's monitoring in the green market allows some differentiation between the two products in favor

of firm 1, despite firm 2's fly-by-night strategy. In contrast, with no monitoring ($\alpha = 0$) the two products look the same in consumers' eyes.

Assume that consumers infer $\mu(p) = 1$ upon seeing a price p charged by firm 2. Using (9), we denote firm 2's expected demand by

$$E\left[D_{2b}(p,1)/\rho_1^*\right] = (1-\sigma_1) D_{2b}(p, p_b^*, 1, 0) + \sigma_1 D_{2b}(p, p_g^*, 1, 1). \tag{21}$$

The functional forms of (21) depend on whether α is positive or zero, as shown by (46) in Appendix 1. Setting p_g^* allows firm 2 to expect the following profits

$$E\left[\pi_{2b}\left(p_g^*, 1\right)/\rho_1^*\right] = p_g^* E\left[D_{2b}(p,1)/\rho_1^*\right]. \tag{22}$$

To prevent firm 2 from imitating firm 1, the separating price p_g^* must satisfy the following credibility constraint:

$$E\left[\pi_{2b}(p_b^*, 0)/\rho_1^*\right] \geq E\left[\pi_{2b}\left(p_g^*, 1\right)/\rho_1^*\right]. \tag{23}$$

The left-hand side of (23) is the profit earned by firm 2 in a separating equilibrium, which is zero. The right-hand side is the profit firm 2 can expect if it mimics the green type, thereby tricking consumers into buying at a positive price. Thus, (23) means that the brown firm has nothing to lose from misleading consumers. Everything else being equal, this profit decreases in α because $\frac{\partial E\left[\pi_{2b}\left(p_g^*,1\right)/\rho_1^*\right]}{\partial \alpha} = \frac{p_g^*\left(p_g^* - p_b^*\right)(\sigma_1 - 1)}{el(1-\alpha)^2} < 0$, but it is also increasing in l. Therefore, the more accurate the auditor's report, the lower is the temptation to cheat for the brown type. However, a larger heterogeneity of consumer preferences for the new attribute strengthens firm 2's incentive to deviate from equilibrium.

For all $\alpha \geq 0$, inequality (23) determines a lower bound for the equilibrium price, $\overline{p}_b$, which must satisfy

$$p_g^* \geq \overline{p}_b, \tag{24}$$

where

$$\overline{p}_b = \begin{cases} l(1-\alpha)e + p_b^c \text{ if } \alpha > 0, \\ le + 2p_b^c \frac{1-\sigma_1}{2-\sigma_1} \text{ if } \alpha = 0. \end{cases} \tag{25}$$

Lemma 1: *In any separating equilibrium, $p_b^* = p_b^c$ and $p_g^* \geq \overline{p}_b$.*

Hence, $\overline{p}_b$ defines the threshold below which it would be misleading to set p_g^* for the green product. Condition (24) guarantees that the equilibrium price in the green market is high enough to deter mimicry by firm 2, which then reverts to p_b^*.

Let us now turn to firm 1's expected profits under different consumer perceptions of its type. Given consumer beliefs $\mu = \mu(p)$ and the pricing rule ρ_2^*, we denote firm 1's expected demand by

$$E\left[D_{1g}(p,\mu)/\rho_2^*\right] = (1-\sigma_2) D_{1g}(p, p_b^*, \mu, 0) + \sigma_2 D_{1g}(p, p_g^*, \mu, 1). \tag{26}$$

Again, firm 1's demand is derived from the market split (9) by substituting $\sigma = 0$ or 1, depending on whether firm 2 signals its true type or uses the fly-by-night strategy. Appendix 1 provides in-depth analysis of firm 1's expected demand with and without monitoring. When firm 1 charges p_g^*, consumer perception of its type is correct, that is, $\widetilde{e}_g(\mu\left(p_g^*\right)) = e$. Given that firm 2 uses the separating pricing rule ρ_2^*, firm 1 predicts that firm 2 charges the price p_b^* for the brown product with probability $1-\sigma_2$. Then, differentiation between the brown and the green products generates a demand for the green product equal to $D_{1g}(p_g^*, p_b^*, 1, 0) = \frac{le+p_b^*-p_g^*}{le}$. If, in contrast, firm 2 chooses the price p_g^*, which occurs with probability σ_2, consumers correctly infer that firm 1's product is more valuable than firm 2's product sold at the same price, provided that $\alpha > 0$, since $\widetilde{e}_b(\mu\left(p_g^*\right)) = (1-\alpha)e < e$. The auditor's monitoring makes all consumers switch towards firm 1's product and the demand for the green product is $D_{1g}(p_g^*, p_g^*, 1, 1) = \frac{le-p_g^*}{le}$. However, when firm 2 uses the fly-by-night strategy, differentiation between the products vanishes in the absence of monitoring ($\alpha = 0$).

Substituting $\mu = 1$ in (26), we can write firm 1's expected profit when it charges p and consumer perception is $\widetilde{e}_g(\mu(p)) = e$ as follows

$$E\left[\pi_{1g}\left(p,1\right)/\rho_2^*\right] = (p-e)E\left[D_{1g}\left(p,1\right)/\rho_2^*\right], \tag{27}$$

where $E\left[D_{1g}\left(p,1\right)/\rho_2^*\right]$ is given by (50) in Appendix 1. Maximizing these expressions with respect to p, we compute firm 1's best response to the pricing rule ρ_2^* to obtain

$$p_{1g}\left(1\right) = \begin{cases} \frac{e(1+l)+p_b^c(1-\sigma_2)}{2} \text{ if } \alpha > 0, \\ \frac{e(1+l)}{2} + p_b^c\frac{1-\sigma_2}{2-\sigma_2} \text{ if } \alpha = 0. \end{cases} \tag{28}$$

The maximized profit can be written as

$$E\left[\pi_{1g}\left(p_{1g}\left(1\right),1\right)/\rho_2^*\right] = \begin{cases} \frac{\left[e(l-1)+p_b^c(1-\sigma_2)\right]^2}{4le} \text{ if } \alpha > 0, \\ \frac{\left[e(l-1)(2-\sigma_2)+2p_b^c(1-\sigma_2)\right]^2}{8le(2-\sigma_2)} \text{ if } \alpha = 0. \end{cases} \tag{29}$$

We see from (28) that the optimal price $p_{1g}\left(1\right)$ may be too low to satisfy the credibility constraint (24). This occurs when $p_{1g}\left(1\right) < \overline{p}_b$, in which case signaling the green type becomes costly, or even impossible, due to the loss of consumers switching to the brown substitute.

As previously mentioned, several candidates for p_g^* may satisfy (24). In order to avoid this, we focus on least costly separating equilibria, which are robust to the Cho-Kreps intuitive criterion (Cho and Kreps, 1987). If $p_{1g}\left(1\right) < \overline{p}_b$ in the present setting, the least costly separation can be achieved by setting the price $\overline{p}_b$ for the green product. The minimum signaling cost is measured by the price differential

$$\overline{p}_b - p_{1g}\left(1\right) = \begin{cases} \frac{e}{2}\left(l-1-2l\alpha+p_b^c\left(1+\sigma_2\right)\right) \text{ if } \alpha > 0, \\ \frac{e}{2}\left(l-1\right) + p_b^c\frac{2-\sigma_2(3-\sigma_1)}{(2-\sigma_1)(2-\sigma_2)} \text{ if } \alpha = 0. \end{cases} \tag{30}$$

The upward distortion in price includes the forgone profit from fly-by-night strategies, thereby forestalling misleading prices. It is unsurprising that the signaling cost decreases as the auditor's monitoring accuracy increases because the cheating profit $E\left[\pi_{2b}\left(p_g^*,1\right)/\rho_1^*\right]$ decreases in α, as previously shown. One can check that $\overline{p}_b > p_{1g}(1)$ for all $\alpha < \overline{\alpha} = \min\{1, \frac{l-1}{2l} + p_b^c \frac{1+\sigma_2}{2el}\}$.

Lemma 2: *In any separating equilibrium, the least costly way of signaling the green type is to set*

$$p_g^* = \begin{cases} p_{1g}(1) \text{ if } \alpha \geq \overline{\alpha}, \\ \overline{p}_b \text{ otherwise.} \end{cases} \tag{31}$$

If the auditor's monitoring is not sufficiently accurate, signaling the green type *via* prices becomes costly due to the brown firm's mimicry. With sufficient accuracy, monitoring saves the cost of signaling the green type. Note, however, that the minimum signaling cost increases in l when $\alpha < \overline{\alpha}$: it may be more difficult to signal the green type in economies with very heterogeneous preferences for the new attribute because the temptation to cheat is stronger for the brown type.

When $p_g^* = \overline{p}_b$, firm 1's expected profits are

$$E\left[\pi_{1g}\left(\overline{p}_b,1\right)/\rho_2^*\right] = \begin{cases} \frac{\left[p_b^c + e(l(1-\alpha)-1)\right]\left(el\alpha - p_b^c\sigma_2\right)}{le} \text{ if } \frac{p_b^c\sigma_2}{el} < \alpha < \overline{\alpha}, \\ \max\{0, p_b^c \frac{\left[2p_b^c(2-\sigma_1)+e(l-1)(2-\sigma_2)\right](\sigma_2-\sigma_1)}{le(2-\sigma_2)^2}\} \text{ otherwise.} \end{cases} \tag{32}$$

Straightforward calculations give $\frac{\partial E\left[\pi_{1g}(\overline{p}_b,1)/\rho_2^*\right]}{\partial\alpha} = e(l-1) - 2el\alpha + p_b^c\sigma_2$, which is positive for all $\alpha < \overline{\alpha}$. Hence, more accurate monitoring strengthens the green firm's incentive to signal its type with an upward distortion in price.

We now examine firm 1's expected profits under the worst belief that consumers might form from a price, $\mu = 0$, that yields consumer perception $\widetilde{e}_g(0) = \alpha e$. Consider that firm 1 sets the price p for the green product. Given the separating pricing rule ρ_2^*, firm 1 predicts that firm 2 charges the price p_b^* for the brown product with probability $1-\sigma_2$. In this event, differentiation between the brown and the green products gives rise to a demand for the green product equal to $D_{1g}(p, p_b^*, 0, 0) = \frac{l\alpha e + p_b^* - p}{l\alpha e}$, provided that $\alpha > 0$. Clearly, this demand is nil in the absence of monitoring since there is no differentiation between the products ($\alpha = 0$).

Alternatively, firm 1 predicts that firm 2 chooses the price p_g^* for the brown product with probability σ_2. Consumers draw inferences from observing two prices in the green market, i. e., p and p_g^*. It follows that consumer valuations are, respectively, αe for firm 1's product, and $\widetilde{e}_b(1) = (1-\alpha)e$ for firm 2's product. Indeed, consumers infer from p_g^* that firm 2's product is green with probability 1 and the auditor's monitoring supports this belief with probability $1-\alpha$. Then, in consumers' eyes, whether firm 1's product is more or less valuable than firm 2's product depends on whether α is higher or lower than $\frac{1}{2}$, respectively. The market split (9) calculated for $\mu = 0$ and $\sigma = 1$ results in demand $D_{1g}(p, p_g^*, 0, 1)$ for firm 1's product. The functional form of this demand given by (52) in Appendix 1 shows that

better accuracy in monitoring within the range $\left[0, \frac{1}{2}\right]$ decreases product differentiation, while it increases product differentiation as soon as α exceeds the threshold $\frac{1}{2}$. As previously mentioned, monitoring somewhat corrects consumer misperceptions of firm 1's product off the equilibrium path. Again, the picture is different with no monitoring, because the two products are no longer differentiated.

In summary, the expected demand of firm 1 when falsely perceived to be brown by consumers is $E\left[D_{1g}\left(p,0\right)/\rho_2^*\right]$ given by (54) in Appendix 1. From this, firm 1's expected profits are

$$E\left[\pi_{1g}\left(p,0\right)/\rho_2^*\right] = (p-e)E\left[D_{1g}\left(p,0\right)/\rho_2^*\right]. \tag{33}$$

Let $p_{1g}\left(0,\sigma_2\right)$ denote firm 1's best response to the pricing rule ρ_2^*. If firm 1 deviates from the price equilibrium to set $p_{1g}\left(0,\sigma_2\right)$, monitoring will correct consumer misperception based on this price. The reduced-form function $\widetilde{\Pi}_{1g}\left(\sigma_2\right) = E\left[\pi_{1g}\left(p_{1g}\left(0,\sigma_2\right),0\right)/\rho_2^*\right]$ represents the spectrum of best worst-outcomes for the green type: this is the best that the green firm can do when it believes that its rival will use the fly-by-night strategy to trick consumers into buying at a positive price by setting p_g^* with probability σ_2. Everything else being equal, the profit $\widetilde{\Pi}_{1g}\left(\sigma_2\right)$ grows with firm 2's probability of cheating, thereby weakening firm 1's incentive to signal its type. Obviously, this incentive also depends on the auditor's monitoring accuracy.

Figure 1 maps out $\widetilde{\Pi}_{1g}\left(\sigma_2\right)$ as a function (in grey) of α, given $l = 3$ and $p_b^c = 0$, in two limit cases: $\sigma_2 = 0$ and $\sigma_2 = 1$; that is, firm 1 predicts with confidence ($\sigma_2 = 0$) that its brown rival will truthfully signal its type with p_b^* and firm 1 definitely expects its rival to cheat consumers ($\sigma_2 = 1$). The light grey area between both curves depicts the whole spectrum of best worst-outcomes for the green type. Figure 1 also depicts $E\left[\pi_{1g}\left(p_g^*,1\right)/\rho_2^*\right]$ as a function (in black) of α, given the same parameter configuration.

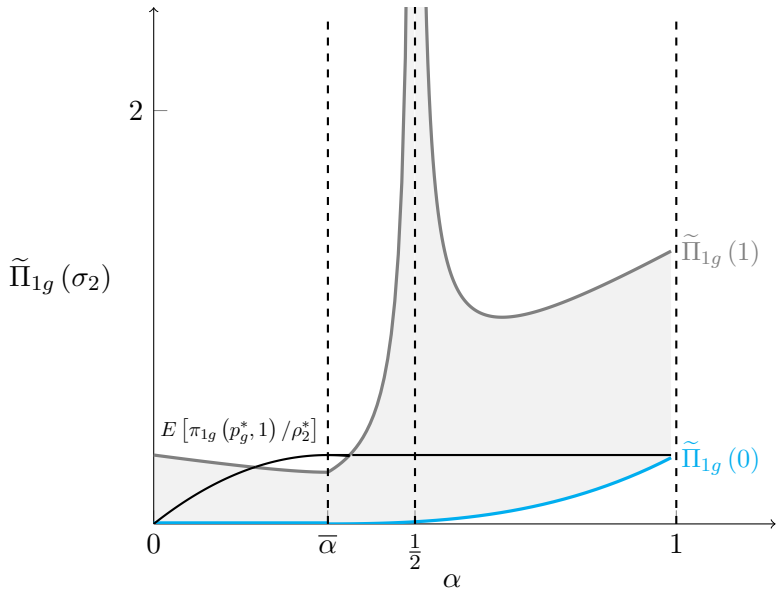

Figure 1: *Spectrum of best worst-profits*

The following table presents the calculation results for firm 1's outcomes.

Firm 1's outcomes for $l=3$ and $p_b^c=0$	Expressions
Firm 1's best worst-profit when $\sigma_2=0$	$\widetilde{\Pi}_{1g}(0)=\frac{e(1-3\alpha)^2}{12\alpha}$ if $\alpha\geq\frac{1}{3}$, and 0 otherwise
Firm 1's best response to ρ_2^* when $\sigma_2=0$	$p_{1g}(0,0)=\frac{e(1+3\alpha)}{2}$ if $\alpha\geq\frac{1}{3}$, and e otherwise
Firm 1's sales volume when $\sigma_2=0$	$E\left[D_{1g}\left(p_{1g}(0,0),0\right)/\rho_2^*\right]=\frac{(3\alpha-1)}{6\alpha}$ if $\alpha\geq\frac{1}{3}$, and 0 otherwise
Firm 1's best worst-profit when $\sigma_2=1$	$\widetilde{\Pi}_{1g}(1)=\begin{cases}\frac{e(1-3\alpha)^2}{3(2\alpha-1)} \text{ if } \alpha\geq\frac{1}{2},\\ \frac{e}{12(1-2\alpha)} \text{ if } \alpha\in\left(\overline{\alpha},\frac{1}{2}\right),\\ \frac{e(2-3\alpha)^2}{12(1-2\alpha)} \text{ if } \alpha\leq\overline{\alpha}=\frac{1}{3}.\end{cases}$
Firm 1's best response to ρ_2^* when $\sigma_2=1$	$p_{1g}(0,1)=\begin{cases}3\alpha e \text{ if } \alpha\geq\frac{1}{2},\\ \frac{3e}{2} \text{ if } \alpha\in\left(\overline{\alpha},\frac{1}{2}\right),\\ \frac{e(4-3\alpha)}{2} \text{ if } \alpha\leq\overline{\alpha}.\end{cases}$
Firm 1's sales volume when $\sigma_2=1$	$E\left[D_{1g}\left(p_{1g}(0,1),0\right)/\rho_2^*\right]=\begin{cases}\frac{e(3\alpha-1)}{3(2\alpha-1)} \text{ if } \alpha\geq\frac{1}{2},\\ \frac{1}{6(1-2\alpha)} \text{ if } \alpha\in\left(\overline{\alpha},\frac{1}{2}\right),\\ \frac{(2-3\alpha)}{6(1-2\alpha)} \text{ if } \alpha\leq\overline{\alpha}.\end{cases}$
Firm 1's least-costly signaling profit	$E\left[\pi_{1g}\left(p_g^*,1\right)/\rho_2^*\right]=\begin{cases}\frac{e}{3} \text{ if } \alpha\geq\overline{\alpha},\\ e\alpha(2-3\alpha) \text{ otherwise.}\end{cases}$
Firm 1's least-costly signaling price	$p_g^*=\begin{cases}p_{1g}(1)=2e \text{ if } \alpha\geq\overline{\alpha},\\ \overline{p}_b=3(1-\alpha)e \text{ otherwise.}\end{cases}$
Firm 1's sales volume with p_g^*	$E\left[D_{1g}\left(p_g^*,1\right)/\rho_2^*\right]=\begin{cases}\frac{1}{3} \text{ if } \alpha\geq\overline{\alpha},\\ \alpha \text{ otherwise.}\end{cases}$

When $\sigma_2 = 0$, $\widetilde{\Pi}_{1g}(0)$ increases with α when profit margin and demand are positive, which happens when α exceeds $\frac{1}{3}$. This is because better monitoring accuracy increases consumer valuation for firm 1's product.

When $\sigma_2 = 1$, $\widetilde{\Pi}_{1g}(1)$ is the dullest outcome for the green type: its product is falsely perceived as brown by consumers who, furthermore, falsely believe the brown firm to be green upon seeing the price p_g^*. We must distinguish two cases depending on whether α is higher or lower than $\frac{1}{2}$.

If $\alpha < \frac{1}{2}$, monitoring slightly corrects consumer misperceptions, and firm 1 can earn positive profits due to product differentiation. However, after reading the auditor's report, consumers find the truly green product less valuable than the false green one. If, moreover, monitoring is so bad that $\alpha < \overline{\alpha}$, the brown firm must distort the price p_g^* upward to cheat consumers; this distortion relaxes the pressure put on the green firm to deter cheating. Figure 1 shows that $\widetilde{\Pi}_{1g}(1)$ decreases with α over the interval $[0, \overline{\alpha}]$ because the fly-by-night strategy of mimicking p_g^* becomes more aggressive as better accuracy in monitoring reduces the price distortion. When α exceeds the threshold $\overline{\alpha}$, there is no longer a need to distort prices for signaling. Then, improved accuracy in monitoring increases consumer valuation for the green product and boosts sales volume without affecting the price $p_{1g}(0,1) = \frac{3e}{2}$ charged for the green product. Therefore, $\widetilde{\Pi}_{1g}(1)$ increases with α over the interval $\left[\overline{\alpha}, \frac{1}{2}\right]$.

If $\alpha > \frac{1}{2}$, consumers find the green product more valuable than its false substitute. Figure 1 shows that $\widetilde{\Pi}_{1g}(1)$ is a convex function of α over the interval $\left[\frac{1}{2}, 1\right]$, with a minimum at $\frac{2}{3}$. Losses and gains of $\widetilde{\Pi}_{1g}(1)$ can be explained by the impact of monitoring accuracy on the green firm's response to the rival fly-by-night strategy. When α is slightly above $\frac{1}{2}$, the two products are close substitutes in consumers' eyes, and the green firm takes the lion's share of the market by attracting consumers with a price $p_{1g}(0,1) = 3\alpha e$ far below $p_g^* = p_{1g}(1)$. In these circumstances, improved monitoring causes significant losses in sales volume, which are not offset by price increases— note that $p_{1g}(0,1)$ increases with α and exceeds $p_{1g}(1)$ when α reaches $\frac{2}{3}$ —. Finally, the green firm can respond less aggressively to the brown firm's cheating only when monitoring is fairly accurate.

In a nutshell, monitoring plays a dual role when the brown firm is cheating consumers: first, it corrects consumer misperceptions; second, it progressively relaxes pressure on the green firm to deter cheating.

Furthermore, firm 1 may be discouraged from signaling its true type with p_g^* if it earns more by deviating to the price $p_{1g}(0, \sigma_2)$. Therefore, the price p_g^* should satisfy the following constraint to be a separating equilibrium

$$E\left[\pi_{1g}\left(p_g^*, 1\right)/\rho_2^*\right] \geq \widetilde{\Pi}_{1g}(\sigma_2). \tag{34}$$

This condition guarantees that it is worthwhile for firm 1 to use the signal p_g^* rather than to be perceived as brown by consumers. The right-hand side of (34) can be interpreted as the green firm's opportunity cost of signaling its true type. As shown in Figure 1, the emergence

of monitoring raises this cost, which makes it harder for firm 1 to reveal the truth about its type. Condition (34) is necessary and sufficient for a least costly separating equilibrium with $p_g^* \geq \overline{p}_b$, and supporting beliefs $\mu^*(p) = 0$ when $p < p_g^*$, and $\mu^*(p) = 1$ when $p \geq p_g^*$. Thus:

Proposition 1: *There exists a least costly separating equilibrium where* $p_b^* = p_b^c$ *and* $p_g^* = \max\{p_{1g}(1), \overline{p}_b\}$ *if and only if* p_g^* *satisfies* (34).

Figure 1 shows that there is an interval of α within which $\widetilde{\Pi}_{1g}(1) < E\left[\pi_{1g}\left(p_g^*, 1\right)/\rho_2^*\right]$; thus requirement (34) is met whatever σ_2, in the specific case where $l = 3$ and $p_b^c = 0$. For any α outside this interval, there exists a whole range of σ_2 strictly below 1, for which requirement (34) is met.

We first examine a baseline model in which there is no monitoring by a third-party auditor. Afterwards, we extend the model to allow for the possibility of imperfect monitoring.

In the scenario with no monitoring, we find that there exists no separating equilibrium under the following circumstance: signaling costs are so high that the green firm has no incentive to deter the brown firm from cheating consumers. In that case, credible price signaling cannot support the switch to green production in the absence of monitoring. This serves as a benchmark for further comparison with the scenario of monitoring. The full analysis shows that monitoring, albeit imperfect, is likely to ensure the credibility of price signaling, and in turn make it worthwhile for a firm to switch to green production.

4.1 Price signaling with no monitoring: the case $\alpha = 0$

As a benchmark, we examine the issue of price signaling in the green market with no monitoring: $\alpha = 0$. This presumes that the market is segmented between the brown product and the product carrying the green seal. As a firm may falsely claim that the brown product is green, the credibility of green certification requires separating prices. These are meant to counteract the brown firm's incentive to cheat consumers. Moreover, we focus on the least costly separating equilibrium: $p_b^* = p_b^c$ and $p_g^* = \max\{p_{1g}(1), \overline{p}_b\}$, from Proposition 1.

In the case where $\alpha = 0$, we also know from the credibility condition (23) that the minimum price candidate for deterring mimicry is

$$\overline{p}_b = le + 2p_b^c \frac{1 - \sigma_1}{2 - \sigma_1}. \tag{35}$$

We suppose again that firm 1 is the green type and firm 2 is the brown type. In the separating equilibrium, consumer perception is correct for both types: $\widetilde{e}_b(\mu(p_b^*)) = 0$ and $\widetilde{e}_g(\mu(p_g^*)) = e$ after observing p_b^* and p_g^* , set by firm 2 and firm 1, respectively.

Consider firm 1's behavior and suppose that firm 1 charges p_g^*. Given that firm 2 uses the separating pricing rule ρ_2^*, firm 1 predicts that firm 2 will charge the prices p_b^* and p_g^* with probabilities $1 - \sigma_2$ and σ_2, respectively. Consumers will draw two different inferences depending on whether firm 2 charges p_b^* or p_g^*. Consumers either infer that the two products are imperfect substitutes from observing two distinct prices or consumers infer that the rival

products are the same from observing the same price p_g^* set by both firms, because no product differentiation can be detected in the absence of monitoring.

From (27) and (50) given in Appendix 1, firm 1's expected profit is

$$E\left[\pi_{1g}\left(p_g^*,1\right)/\rho_2^*\right]=(p_g^*-e)\frac{\left(el-p_g^*\right)(2-\sigma_2)+2p_b^c(1-\sigma_2)}{2le}, \tag{36}$$

which is maximized at

$$p_{1g}(1)=\frac{e(1+l)}{2}+p_b^c\frac{1-\sigma_2}{2-\sigma_2}. \tag{37}$$

Comparing (35) and (37) reveals that firm 1 may be forced to distort upward p_g^* in order to deter its brown rival from cheating consumers. Therefore, separation is potentially costly for the green type. The following lemma states the parameter conditions under which $\overline{p}_b$ exceeds $p_{1g}(1)$.

Lemma 3: *Assume* $(5),\alpha=0$ *and* $p_b^c\frac{\sigma_2(3-\sigma_1)-2}{(2-\sigma_1)(2-\sigma_2)}<\frac{e}{2}(l-1)$. *In any separating equilibrium,* $p_b^*=p_b^c$ *and* p_g^* *must be distorted upward relative to* $p_{1g}(1)$.

It may happen that the threat of a fly-by-night strategy entails positive signaling costs for the green firm. Price distortion allows the green firm to prove that it is less reluctant than its brown rival to restrict sales volume. The logic is the same as that recognized in most models of price signaling quality in markets of experience goods. In Bagwell and Riordan (1991), a monopolist signals high quality by raising prices up to the level where the loss of sales for the low-quality type is not worth the rent from cheating. Thus, the cost of signaling high quality is determined by the forgone rent from cheating. Compared to the monopoly regime, price competition in oligopolistic markets reduces the degree to which firms distort prices for the purpose of signaling (as shown by Daughety and Reinganum, 2008; Janssen and Roy, 2010).

In the present context, a novel insight is that price separation fails if p_g^* is distorted to the point that firm 1's expected sales volume falls to zero. From (36), we see that $\overline{p}_g=le+2p_b^c\frac{1-\sigma_2}{2-\sigma_2}$ is the maximum price for which demand for the green product is positive. Thus, $\overline{p}_g\leq\overline{p}_b$ when $p_b^c\frac{\sigma_2-\sigma_1}{(2-\sigma_1)(2-\sigma_2)}\geq 0$. In that case, credible price signaling is too demanding to be attractive with no monitoring. Following Mahenc (2017), green certification cannot be credible if prices fail to signal the green type. In such circumstances, no firm will pay the setup cost F to switch to green production. This serves as a baseline model to further investigate the case of imperfect monitoring. It turns out that the signaling failure occurs with no monitoring in the limit case where $p_b^c=0$, and so there is no added value in performing calculations for cases where $p_b^c>0$. In the remainder of the paper, we normalize p_b^c to zero whenever appropriate for calculations.

Corollary 1: *In the absence of auditor's monitoring, if* $p_b^c=0$, *then there exists no separating equilibrium in which the green firm can credibly signal its type. In any subgame perfect equilibrium of the three-stage game, no firm will switch to green production with the aim of signaling its true type.*

As in Akerlof (1970), there is an adverse selection problem in the absence of monitoring; namely, there is no incentive for a given firm to provide anything but the brown product with minimum verifiable quality. As a result, the market for the green product is overrun by business as usual.

Corollary 1 also implies that there is a moral hazard problem: a firm does not commit to switching to green production because there is no credible way of signaling the switch. This is closely related to the moral hazard problem initially pointed out by Klein and Leffler (1981): a firm may refrain from producing high quality products because it would lose all consumers at the minimum price needed to signal high quality. Klein and Leffler show that the threat of losing future business is likely to prevent a firm from reneging on its promise to enhance product quality.

The analysis of the full scenario below shows that the auditor's monitoring can not only mitigate the adverse selection problem, but also solve the moral hazard problem.

4.2 Price signaling with monitoring: the case $\alpha > 0$

We now introduce the auditor to the previous benchmark. The auditor inspects the firm claiming that its product is green in order to detect the presence of the new attribute. The auditor's monitoring reveals the true type of the firm with probability α. A lack of accuracy in monitoring is perfectly known to consumers. They read the auditor's report while observing the price signals sent by firms. Using both channels of communication, consumers infer information about the firms' types and, finally, make their purchase decisions.

Unlike the scenario with no monitoring, the auditor's monitoring allows consumers to differentiate products when the brown firm tricks them into buying at price p_g^*.

From (31), the least costly separating price is distorted upward if the monitoring accuracy falls below the threshold $\alpha = \overline{\alpha}$. Assuming that firm 1 is the green type, the expected profit earned in the least costly separating equilibrium depends on the monitoring accuracy. From (29) and (32), we have

$$E\left[\pi_{1g}\left(p_g^*,1\right)/\rho_2^*\right] = \begin{cases} \frac{\left[e(l-1)+p_b^c(1-\sigma_2)\right]^2}{4le} \text{ if } \alpha \geq \overline{\alpha}, \\ \frac{\left[p_b^c+e(l(1-\alpha)-1)\right]\left(el\alpha-p_b^c\sigma_2\right)}{le} \text{ if } \frac{p_b^c\sigma_2}{el} < \alpha < \overline{\alpha}. \end{cases} \tag{38}$$

Proposition 1 states that firm 1 has no incentive to defect from p_g^* if and only if the opportunity cost of signaling its true type with p_g^*, $\widetilde{\Pi}_{1g}(\sigma_2)$, does not exceed $E\left[\pi_{1g}\left(p_g^*,1\right)/\rho_2^*\right]$.

Let us now examine in detail condition (34).[10] The maximized profit $\widetilde{\Pi}_{1g}(\sigma_2)$ depends on the probability σ_2 that firm 2 uses the fly-by-night strategy of charging p_g^*, unlike $E\left[\pi_{1g}\left(p_g^*,1\right)/\rho_2^*\right]$ when we normalize p_b^c to zero. Setting $\widetilde{\Pi}_{1g}(\sigma_2) = 0$ defines a threshold $\underline{\sigma}$ for σ_2, above which the margin and demand of firm 1 are positive at the same time. It turns out that $\underline{\sigma} < 1$ for all $\alpha \in [0,1]$.

[10] All the calculations and proofs for the full scenario with monitoring can be found in Appendix 2. To reduce the number of cases to review, we have assumed that $l \geq 2$. As previously mentioned, p_b^c is normalized to zero in order to allow a straightforward comparison with the baseline model without monitoring.

This can be seen in Figure 1, where $\widetilde{\Pi}_{1g}(1) > 0$ for all $\alpha \in [0,1]$, given $l = 3$ and $p_b^c = 0$. In this figure, the light grey area between the curves $\widetilde{\Pi}_{1g}(0)$ and $\widetilde{\Pi}_{1g}(1)$ depicts the spectrum of $\widetilde{\Pi}_{1g}(\sigma_2)$ such that $\sigma_2 \geq \underline{\sigma}$. If firm 2's probability of cheating falls below $\underline{\sigma}$, then firm 1's opportunity cost of signaling its true type is reduced to zero. As a preliminary conclusion, we can state that a least costly separating equilibrium exists for all $\sigma_2 \leq \underline{\sigma}$, provided that $\underline{\sigma} > 0$.

We now introduce the function $\widetilde{\Delta}_{1g}(\sigma_2) = E\left[\pi_{1g}\left(p_g^*, 1\right)/\rho_2^*\right] - \widetilde{\Pi}_{1g}(\sigma_2)$, and assume that $\sigma_2 > \underline{\sigma}$. Under these circumstances, we have $\widetilde{\Pi}_{1g}(\sigma_2) > 0$. Furthermore, setting $\widetilde{\Delta}_{1g}(\sigma_2) = 0$ defines a critical $\overline{\sigma}$ for the probability of cheating, below which a least costly separating equilibrium exists, provided that $\overline{\sigma} > \max\{0, \underline{\sigma}\}$. If so and inequality $\underline{\sigma} > 0$ is satisfied as well, then firm 1's opportunity cost of signaling its true type becomes positive at $\sigma_2 > \underline{\sigma}$ and increases with σ_2 until this probability reaches a "no-defect" threshold $\min\{\overline{\sigma}, 1\}$. In Figure 1, $\overline{\sigma}$ is determined according to α at every point where the black curve $E\left[\pi_{1g}\left(p_g^*, 1\right)/\rho_2^*\right]$ crosses the light grey area.

We have previously seen that, in the best worst-situation the green type may face, a green product mistaken for brown is more or less valuable than the brown product mistaken for green, depending on whether α is higher or lower than $\frac{1}{2}$, respectively. For the sake of clarity, we henceforth refer to $\alpha > \frac{1}{2}$ as a "good" accuracy for monitoring.

Furthermore, price signaling happens to be costless or costly, depending on whether α exceeds or falls short of $\overline{\alpha}$, respectively. Therefore, we must also distinguish between both of these cases. For this, we refer to $\alpha \in \left[\overline{\alpha}, \frac{1}{2}\right]$ and $\alpha < \overline{\alpha}$, respectively, as "intermediate" and "bad" accuracy for monitoring.

Calculations yield two limit functions for α, defined as follows: $\alpha_0(l) = \frac{5l - \sqrt{l(32-7l)}}{16l}$ and $\alpha_1(l) = \min\{\overline{\alpha}, \frac{5l + \sqrt{l(32-7l)}}{16l}\}$ are, respectively, the lowest and highest values of α for which $\overline{\sigma} \geq 1$. One can check that these functions are ranked in the following order: $\alpha_0(l) < \alpha_1(l) < \frac{1}{2}$ for all $l \geq 2$. The upper boundary for the probability of cheating that is consistent with the existence of a price separating equilibrium is $\overline{\sigma}$ or 1, depending on the parameter values of l and α. The following lemma lays the groundwork for the parameter values under which there exists a separating equilibrium.

Lemma 4: *Assume that $l \geq 2$ and $p_b^c = 0$. For all $\alpha < \frac{1}{2}$,*

$$\max\{0, \underline{\sigma}\} < \overline{\sigma}.$$

The minimum threshold above which $\widetilde{\Pi}_{1g}(\sigma_2) > 0$ satisfies

(i) $0 < \underline{\sigma}$ *for all* $\alpha < \frac{1}{l}$,

(ii) $\underline{\sigma} < 1$ *whatever* α.

The maximum threshold below which $\widetilde{\Delta}_{1g}(\sigma_2) > 0$ satisfies

(i) $0 < \overline{\sigma}$ *for all* $\alpha \geq \frac{1}{2}$,

(ii) $\overline{\sigma} \geq 1$ *for all* $\alpha \in [\alpha_0(l), \alpha_1(l)]$.

Proof: see Appendix 2.

We are now ready to specify the conditions for parameters l, α and σ_2, under which separation of the brown and the green type can be achieved *via* prices.

Proposition 2: *Assume that* $l \geq 2$ *and* $p_b^c = 0$. *With imperfect monitoring, a least costly separating equilibrium exists for all* $\alpha \in (0, 1]$ *if and only if* $\sigma_2 \leq \sigma^{defect}$, *where*

$$\sigma^{defect} = \begin{cases} \overline{\sigma} \text{ when monitoring accuracy is good,} \\ \min\{\overline{\sigma}, 1\} \text{ when monitoring accuracy is intermediate or bad.} \end{cases} \tag{39}$$

This result contrasts with the non-existence of separating equilibria established in Corollary 1. This shows that imperfect monitoring operates as a mechanism for assuring the credibility of price signaling. In the least costly separating equilibrium, the price of the green product may include an upward distortion (when monitoring accuracy is bad) or not (when monitoring accuracy is good or intermediate), due to the forgone profit from the brown firm's fly-by-night attempt to mislead consumers. Thus, monitoring with good or intermediate accuracy saves the signaling costs for the green firm. Furthermore, when the critical level σ^{defect} falls short of 1, which may occur for every level of monitoring accuracy, the green firm finds it worthwhile to disclose full information about its type, unless the probability of cheating exceeds σ^{defect}.

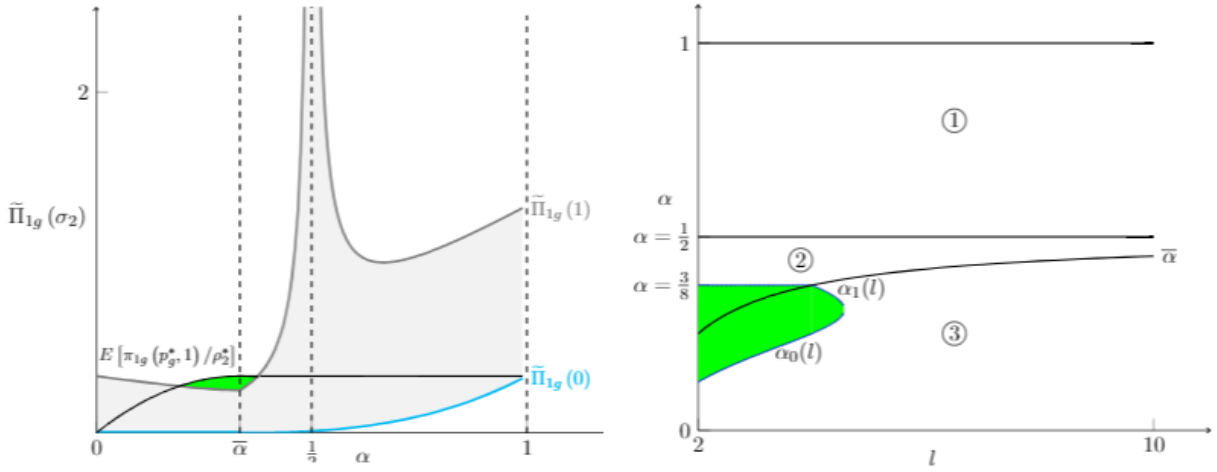

Figure 2: *Existence of Separating Equilibria*

The right-hand picture in Figure 2 divides the (l, α) space into three regions, in which the least costly separating equilibrium exists.

In Region 1, monitoring accuracy is good. There is no extra cost for signaling the green type, so $p_g^* = p_{1g}(1)$. The brow firm cannot trick consumers into buying at this price

provided that σ_2 remains sufficiently below 1; namely, $\sigma_2 \leq \overline{\sigma}$. Therefore, price signaling succeeds unless the brown firm is too likely to cheat. When σ_2 exceeds $\overline{\sigma}$, the weight on the fly-by-night strategy in the best worst-profit $\widetilde{\Pi}_{1g}(\sigma_2)$ is too heavy for the green firm to use p_g^* as a signal of its type. Therefore, the green firm prefers to be mistaken for brown for all $\sigma_2 > \overline{\sigma}$.

In Region 2, monitoring accuracy is moderate, though still sufficient to save signaling costs for the green firm. Thus, the price signaling the green type is still $p_g^* = p_{1g}(1)$. As in Region 1, separation can occur if only if the brown firm is not too likely to cheat: $\sigma_2 \leq \overline{\sigma}$ when $\alpha > \max\{\overline{\alpha}, \frac{3}{8}\}$. In this parameter configuration, the best worst-profit $\widetilde{\Pi}_{1g}(1)$ may be very high when α is close to $\frac{1}{2}$. As previously seen, the reason for this is that monitoring accuracy boosts sales volume for the green firm without affecting the price p_g^*, if the brown firm ever tries to cheat consumers. The probability of cheating must be sufficiently below 1 to reduce $\widetilde{\Pi}_{1g}(\sigma_2)$ to the point where the green firm foregoes this profit and instead signals its true type. When α falls below $\frac{3}{8}$ (dark grey area), we know that $\widetilde{\Pi}_{1g}(1)$ decreases due to a loss in sales volume for the green firm when faced with the rival's fly-by-night strategy. This provides the green firm with a stronger incentive to signal its type, even if it is sure that the brown firm is cheating consumers. As a result, the least costly separating equilibrium exists regardless of σ_2.

In Region 3, monitoring accuracy is bad and cheating is more attractive to the brown firm than in Region 2. Therefore, an upward-distorted price is needed to signal the green type. As a result, price signaling for the green type is $p_g^* = \overline{p}_b$. This is a credible strategy for the green firm because it suffers less than its brown rival from the consequent loss of sales volume, due to the gap in production costs. Moreover, the green firm is better off signaling its true type than being mistaken for brown, even if it is sure that the rival is mimicking $\overline{p}_b$, as long as α lies inside $[\alpha_0(l), \alpha_1(l)]$ (dark grey area). When α is outside this area, the fly-by-night strategy becomes less aggressive. Consequently, the best worst-profit $\widetilde{\Pi}_{1g}(1)$ increases to the point that the green firm is worse off with the price $\overline{p}_b$ than with the lower price $p_{1g}(0,1)$ at which it is mistaken for the brown type. σ_2 must fall short of $\overline{\sigma}$ to reduce $\widetilde{\Pi}_{1g}(\sigma_2)$ enough for the green firm to separate with $\overline{p}_b$.

Solving the game by backward induction, we now examine how firms choose a type at the first stage of the game. Setting $p_b^c = 0$ in (13) gives firm 1's expected payoff from playing a mixed strategy as

$$\Pi_1(\sigma_1, \sigma_2) = \begin{cases} \sigma_1 \left[(1-\sigma_2) \frac{(e(l-1))^2}{4el} - F \right] & \text{if } p_g^* = p_{1g}(1), \\ \sigma_1 \left[(1-\sigma_2)(l(1-\alpha)-1))e\alpha - F \right] & \text{if } p_g^* = \overline{p}_b. \end{cases} \tag{40}$$

Consider now the mixed strategy subgame PBEs of the model satisfying the constraint (16). Define σ^{IR} as the critical σ_2 above which firm 1 does not recover the cost of switching

to green production

$$\sigma^{IR} = \begin{cases} 1 - \frac{F4el}{(e(l-1))^2} \text{ if } p_g^* = p_{1g}(1), \\ 1 - \frac{F}{(l(1-\alpha)-1)e\alpha} \text{ if } p_g^* = \overline{p}_b. \end{cases} \tag{41}$$

Inequality $0 < \sigma^{IR}$ holds for all $F < \frac{(e(l-1))^2}{4el}$ and all $\alpha \in (\overline{\alpha} - \frac{\sqrt{(e(l-1))^2 - 4elF}}{2el}, 1]$. Let us assume this to be the case.

Straightforward calculations yield $\frac{\partial \sigma^{IR}}{\partial \alpha} = F\frac{l(1-2\alpha)-1}{e\alpha^2(l(1-\alpha)-1)^2} > 0$ for all $\alpha < \overline{\alpha}$, such that $p_g^* = \overline{p}_b$. Hence, σ^{IR} increases as α increases. In the case of bad monitoring, cheating problems are less severe with more accurate monitoring.

Using the notation $\widehat{\sigma} = \min\{\sigma^{defect}, \sigma^{IR}\}$, we can now construct firm 1's best response function. If $\sigma_2 < \widehat{\sigma}$, then firm 1's unique best response is $t = g$, while if $\sigma_2 \geq \widehat{\sigma}$, then firm 1's unique best response is $t = b$ because it would incur the loss of F by changing its decision to switch to green production. In summary, firm 1's best response function is

$$\sigma_1^*(\sigma_2) = \begin{cases} 1 \text{ if } \sigma_2 < \widehat{\sigma}, \\ \text{any } \sigma_1 \in [0,1] \text{ if } \sigma_2 = \widehat{\sigma}, \\ 0 \text{ if } \sigma_2 > \widehat{\sigma}. \end{cases} \tag{42a}$$

By symmetry, we can find firm 2's best response function. The best response functions of both firms are depicted in Figure 3. The set of mixed-strategy Nash equilibria corresponds to the set of intersections of the best response functions in this figure.

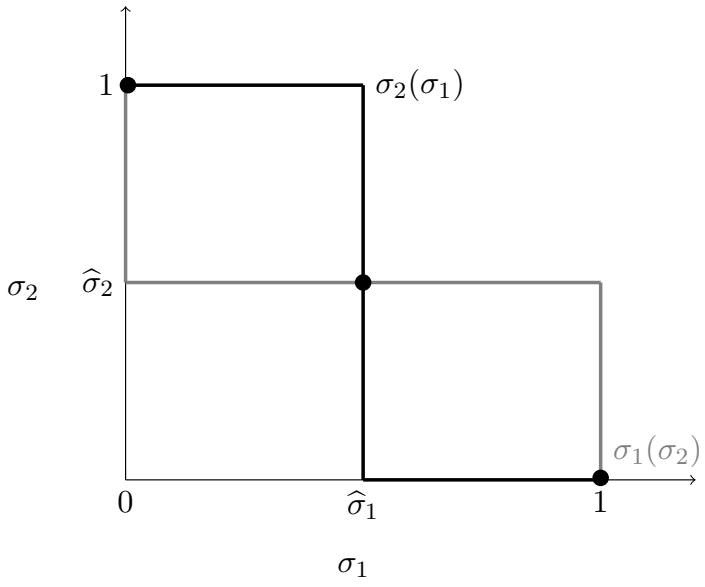

Figure 3: *Best response functions and Nash Equilibria*

Proposition 3: *Assume that* $l \geq 2, p_b^c = 0, F < \frac{(e(l-1))^2}{4el}$ *and* $\alpha \in (\overline{\alpha} - \frac{\sqrt{(e(l-1))^2 - 4elF}}{2el}, 1]$. *With imperfect monitoring, there exists three PBEs in which* $(\sigma_1^*, \sigma_2^*) = (0,1), (1,0)$ *and* $(\widehat{\sigma}, \widehat{\sigma})$.

In the first two equilibria, firms choose asymmetric pure strategies: one firm switches to green production with probability 1, while the other sticks to brown production with probability 1. In these equilibria, all the protagonists of the game, including consumers, are certain that the firms put two differentiated products on the market. The auditor's monitoring mitigates the threat of fly-by-night strategies on the brown firm's side. If consumers observe such a deviation from price equilibrium in the signaling subgame, monitoring partly corrects the misperception, thereby maintaining some degree of differentiation between the products. The cheating profit resulting from Bertrand competition off the equilibrium path decreases sufficiently to make the fly-by-night strategy unattractive. Thanks to the auditor's report, the green firm can rely on price to curtail cheating and hence credibly signal its type. Simultaneously, the brown firm becomes fully convinced that cheating is worthless. If either firm is sure that its rival switches to green production, then it is better off sticking to brown production. A dilemma remains in the pure strategy equilibria: there is no explanation for how firms know which equilibrium will play out.

There is no such a dilemma in the symmetric mixed-strategy equilibrium. Both firms choose to switch to green production with the same probability $\widehat{\sigma}$. This probability turns into the probability of cheating in the signaling subgame. Intuitively, cheating is as likely to occur as the switch to green production because cheating boils down to mimicking the green behavior as often as possible. The probability $\widehat{\sigma}$ is sufficiently low in equilibrium to meet the two requirements for curtailing cheating and revealing the truth. Another interpretation of the mixed-strategy equilibrium is that the proportion of firms that choose to switch to green production in the economy must remain reasonably low for price signaling to be credible. If too high a proportion of firms turn into green firms, the likelihood that brown firms will mimic is just as high and this threat may be too strong to afford the cost of revealing the truth.

We have seen that price signaling entails no further cost for the green firm when monitoring accuracy is good or intermediate. In those cases, the problem of cheating is less a matter of concern than that of revealing the truth for the green firm. This firm must forego the profit earned in the best worst-outcome in which consumers mistake it for the brown type. The opportunity cost of signaling the green type decreases as the brown rival becomes less likely to cheat. Therefore, the probability of cheating must offer the green firm a balanced compromise between revealing the truth and being mistaken for the brown type.

In the case where monitoring is bad, the problem of cheating is more severe because price signaling turns to be costly for the green firm. In equilibrium, the probability of cheating must be sufficiently low to make the brown firm indifferent between signaling its true type and using the fly-by-night strategy. Increased accuracy in monitoring mitigates the signaling distortion, thereby reducing the problem of cheating. When cheating is no longer a problem, the probability of switching to green production coincides with the probability of cheating for which the firm exactly covers the cost of switching.

5 Conclusion

When social goodwill enhances product value in consumers' eyes, firms can increase profits by tying social goodwill in with their product. However, social goodwill is a credence attribute of the product that may be difficult to assess and certify. Asymmetric information about social goodwill may raise a twofold problem of adverse selection and moral hazard: firms renege on their pledge of social goodwill and the market is overrun by products carrying spurious labels.

In a simple model of signaling where price alone fails to send a credible signal of social goodwill, we show that independent monitoring, although imperfect, restores the credibility of price signaling, motivates firms to reveal the truth and, finally, helps a firm fulfill its pledge of social goodwill. The argument for this stems from the existence of a mixed-strategy equilibrium. The fly-by-night strategy of mimicking the price signal of social goodwill occurs as often as a firm pledges to tie social goodwill in with its product. Furthermore, a firm can afford the cost of using price as a signal of social goodwill provided that the fly-by-night strategy is not used too often to mislead consumers. Therefore, in a mixed-strategy equilibrium, there may exist a reasonably low probability that a firm will pledge social goodwill. This promise is credible because the firm, aided by monitoring, can find a price aimed at deterring the rival from cheating while making it worthwhile to reveal the truth about its goodwill.

There also exists pure strategy equilibria in which monitoring mitigates the threat of fly-by-night strategies by improving consumer perception based on prices. In these equilibrium outcomes, only one firm pledges social goodwill because, again, it can rely on the signal of social goodwill sent *via* price. Knowing this pledge, the rival firm has no incentive to make the same pledge and any attempt to cheat consumers is unsuccessful.

Our findings highlight the role played by independent monitoring beyond the informational content of the auditor's report. Monitoring underpins the credibility of price signaling, and hence the honesty of certification. When signaling social goodwill through prices, a firm can rely on monitoring to correct for any arbitrary beliefs consumers might hold after observing a deviation from the equilibrium path. In other words, monitoring acts as a refinement of Bayesian equilibrium. It may happen that signaling social goodwill through price is costly due to the fly-by-night strategy of mimicking prices. In that case, monitoring relaxes the pressure that the firm must withstand to deter cheating. It turns out that the signaling cost decreases with enhanced accuracy in monitoring because the cheating profit is lower. Beyond a certain level of accuracy, monitoring saves the full cost of signaling social goodwill.

In the present world, monitoring is increasingly conducted by non-governmental organizations (NGOs). Among various activities, they operate as watchdogs for public and private certification of social goodwill by inspecting and testing products to verify an industry's compliance with certification standards. Their monitoring is constantly improving with advances in information and communications technology. In various recent cases, NGOs have demonstrated skill and ability in disclosing accurate information on fly-by-night practices in

industry.[11] For instance, the International Council on Clean Transportation commissioned a study on emissions discrepancies between European and US models of diesel vehicles that has cast serious doubts on Volkswagen's compliance with environmental standards. Greenpeace and the Environmental Investigation Agency have produced evidence that FSC certification has been granted to logging companies operating with little regard for sustainability or even legality. While some firms see NGOs' monitoring as a threat, others may be willing to invite them to strengthen their commitment to social goodwill. This suggests that firms and independent auditors somehow interact to release information to the public, in a way that can be cooperative or not.

In order to investigate this strategic interaction, it would be worthwhile to endogenize the auditor's behavior in the present model.

[11] Several examples can be found in Delmas and Burbano (2011) and Heyes and Martin (2017).

6 Appendix

6.1 Appendix 1: Demand functions

Demand functions under full information In the case where firm i of type g charges price p_{ig} and the rival firm j of the same type charges p_{jg}, products are undifferentiated. Then, $D_{ig}(p_{ig}, p_{jg}) = D^u_{ig}(p_{ig}, p_{jg})$, where

$$D^u_{ig}(p_{ig}, p_{jg}) = \begin{cases} \frac{le - p_{ig}}{le} \text{ if } p_{ig} < p_{jg} \\ \frac{1}{2}\frac{le - p_{ig}}{le} \text{ if } p_{ig} = p_{jg} \\ 0 \text{ if } p_{ig} > p_{jg} \end{cases} \tag{43}$$

In the case where firm i is green and its rival is brown, the rival products are differentiated. Then, $D_{ig}(p_{ig}, p_{jg}) = D^d_{ig}(p_{ig}, p_{jg})$, where

$$D^d_{ig}(p_{ig}, p_{jb}) = \begin{cases} 1 \text{ if } p_{ig} \leq p_{jb} \\ \frac{le + p_{jb} - p_{ig}}{le} \text{ if } p_{jb} < p_{ig} < le + p_{jb} \\ 0 \text{ if } p_{ig} \geq le + p_{jb} \end{cases} \tag{44}$$

Demand functions under asymmetric information In the case where consumers have higher expected valuation for the product i than for the product j, the demand for firm i resulting from the market split at (9) is given by

$$D_{it}(p_{it}, p_{jt'}, \mu, \sigma) = \begin{cases} 1 \text{ if } p_{it} < p_{jt'} \\ \frac{l\widetilde{e}_t(\mu) - p_{it}}{l\widetilde{e}_t(\mu)} \text{ if } p_{it} = p_{jt'} \\ \frac{l(\widetilde{e}_t(\mu) - \widetilde{e}_{t'}(\sigma)) + p_{jt'} - p_{it}}{l(\widetilde{e}_t(\mu) - \widetilde{e}_{t'}(\sigma))} \text{ if } p_{jt'} < p_{it} < l\left(\widetilde{e}_t(\mu) - \widetilde{e}_{t'}(\sigma)\right) + p_{jt'} \\ 0 \text{ if } p_{it} \geq l\left(\widetilde{e}_t(\mu) - \widetilde{e}_{t'}(\sigma)\right) + p_{jt'} \end{cases}. \tag{45}$$

In the case where firm 1 has chosen to be green while firm 2 has chosen to be brown, the pricing rule ρ^*_i predicts that firm i will charge price p^*_b with probability $1 - \sigma_i, i = 1, 2$.

Firm 2's expected demand Given ρ^*_1 and consumer beliefs $\mu(p) = 1$ after observing the price p^*_g charged by firm 2, its expected demand (21) takes two different forms depending on whether monitoring is active or not; namely,

$$E\left[D_{2b}\left(p^*_g, 1\right)/\rho^*_1\right] = \begin{cases} (1 - \sigma_1)\frac{l(1-\alpha)e + p^*_b - p^*_g}{l(1-\alpha)e} \text{ if } \alpha > 0, \\ (1 - \sigma_1)\frac{le + p^*_b - p^*_g}{le} + \frac{\sigma_1}{2}\frac{le - p^*_g}{le} \text{ if } \alpha = 0. \end{cases}. \tag{46}$$

Firm 1's expected demand when consumer perception is correct If firm 1 charges p^*_g, then consumer perception of its type is correct, i. e., $\mu\left(p^*_g\right) = 1$. From the pricing rule ρ^*_2, two events may occur: firm 2 signals its true type with p^*_b or firm 2 plays the fly-by-night strategy of mimicking p^*_g. In the first event, firm 1's demand is derived from (9)

by substituting $\mu = 1$ and $\sigma = 0$, yielding

$$D_{1g}(p_g^*, p_b^*, 1, 0) = \frac{le + p_b^* - p_g^*}{le}. \tag{47}$$

In the second event, firm 1's demand with monitoring is derived from (9) by substituting $\mu = 1$ and $\sigma = 1$, yielding

$$D_{1g}(p_g^*, p_g^*, 1, 1) = \frac{le - p_g^*}{le}. \tag{48}$$

Demands (47) and (48) are those faced by firm 1 in the presence of monitoring. With no monitoring, firm 1's demand is the same as (47) when firm 2 signals its true type. But when firm 2 plays the fly-by-night strategy, consumers perceive the two rival products as the same, and so firm 1's demand becomes

$$D_{1g}(p_g^*, p_g^*, 1, 1) = \frac{1}{2}\frac{le - p_g^*}{le}. \tag{49}$$

To sum up, when firm 1 charges p_g^*, firm 1's expected demand is

$$E\left[D_{1g}\left(p_g^*, 1\right)/\rho_2^*\right] = \begin{cases} (1-\sigma_2)\frac{le+p_b^*-p_g^*}{le} + \sigma_2\frac{le-p_g^*}{le} \text{ if } \alpha > 0, \\ (1-\sigma_2)\frac{le+p_b^*-p_g^*}{le} + \frac{\sigma_2}{2}\frac{le-p_g^*}{le} \text{ if } \alpha = 0. \end{cases} \tag{50}$$

Firm 1's expected demand when consumer perception is wrong In the case where consumers hold the worst belief $\mu = 0$ after observing a price p set by firm 1, consumer perception of its type is $\widetilde{e}_g(0) = \alpha e$ from (6). Again, it may happen that firm 2 signals its true type with probability $1-\sigma_2$ or plays the fly-by-night strategy with probability σ_2. Then, consumer perception of its type is $\widetilde{e}_b(0) = 0$ in the first event, and $\widetilde{e}_b(1) = (1-\alpha)e$ in the second event.

If firm 2 charges p_b^*, then the market split (9) calculated for $\mu = 0$ and $\sigma = 0$ gives

$$D_{1g}(p, p_b^*, 0, 0) = \frac{l\alpha e + p_b^* - p}{l\alpha e}, \tag{51}$$

when α is positive, and 0 otherwise.

If firm 2 chooses price p_g^* while firm 1 sets p, then the market split (9) calculated for $\mu = 0$ and $\sigma = 1$ yields the following demand for firm 1's product

$$D_{1g}(p, p_g^*, 0, 1) = \begin{cases} \frac{l(2\alpha-1)e-p+p_g^*}{l(2\alpha-1)e} \text{ if } \alpha \geq \frac{1}{2}, \\ \frac{p_g^*-p}{l(1-2\alpha)e} \text{ if } \alpha \in \left(0, \frac{1}{2}\right). \end{cases} \tag{52}$$

If $\alpha = 0$, then the demand for firm 1's product, given that firm 2 chooses the price p_g^* with probability σ_2, is

$$D_{1g}(p, p_g^*, 0, 1) = \begin{cases} 1 \text{ if } p \leq p_g^* - e \\ \frac{p_g^*-p}{le} \text{ if } p_g^* - e < p < p_g^* \\ 0 \text{ if } p \geq p_g^* \end{cases} \quad \text{if } \alpha = 0. \tag{53}$$

When consumers falsely perceive firm 1 to be brown, its expected demand is

$$E\left[D_{1g}\left(p,0\right)/\rho_2^*\right]=\begin{cases}\left(1-\sigma_2\right)\frac{l\alpha e-p}{l\alpha e}+\sigma_2\frac{l(2\alpha-1)e-p+p_g^*}{l(2\alpha-1)e} \text{ if } \alpha\geq\frac{1}{2},\\ \left(1-\sigma_2\right)\frac{l\alpha e-p}{l\alpha e}+\sigma_2\frac{p_g^*-p}{l(1-2\alpha)e} \text{ if } \alpha\in\left(0,\frac{1}{2}\right),\\ \left(1-\sigma_2\right)\times 0+\sigma_2\frac{p_g^*-p}{le} \text{ if } \alpha=0 \text{ and } p_g^*-e<p<p_g^*.\end{cases} \tag{54}$$

6.2 Appendix 2: Proof of Lemma 4

In the best worst-situation, the green firm believes that its rival uses the fly-by-night strategy by setting p_g^* with probability σ_2. The best worst-profit (33) reaches a maximum at $p_{1g}\left(0,\sigma_2\right)$, where it takes the following values, depending on whether α is higher or lower than $\frac{1}{2}$,

$$E\left[\pi_{1g}\left(p_{1g}\left(0,\sigma_2\right),0\right)/\rho_2^*\right]=\begin{cases}\frac{\left(p_g^*\alpha\sigma_2+e\left(\sigma_2-1+\alpha(2+l-\sigma_2(l+3))+2l\alpha^2(\sigma_2-1)\right)\right)^2}{4le\alpha(2\alpha-1)(1-2\alpha+\sigma_2(3\alpha-1))} \text{ if } \alpha\leq\frac{1}{2},\\ \frac{\left(p_g^*\alpha\sigma_2-e\left(1-\sigma_2+\alpha(\sigma_2-l-2)+2l\alpha^2\right)\right)^2}{4el\alpha(1-2\alpha)(1-\sigma_2+\alpha\sigma_2-2\alpha)} \text{ otherwise.}\end{cases} \tag{55}$$

From Lemma 2, the least costly price that signals the green type is $p_g^*=p_{1g}\left(1\right)$ if $\alpha\geq\overline{\alpha}$, and $p_g^*=\overline{p}_b$ otherwise. Substituting $p_{1g}\left(1\right)$ and $\overline{p}_b$ for p_g^* in (55) leads to consider three cases.

1. Monitoring accuracy is good: $\alpha>\frac{1}{2}$

   For all $\sigma_2>\underline{\sigma}$, the best worst-profit is

   $$\widetilde{\Pi}_{1g}\left(\sigma_2\right)=\frac{\left(e\left(2\left(1-2\alpha\right)\left(l\alpha-1\right)-\sigma_2\left(3\alpha+l\alpha-2\right)\right)\right)^2}{16le\alpha\left(1-2\alpha\right)\left(1-\sigma_2+\alpha\sigma_2-2\alpha\right)}, \tag{56}$$

   and the margin profit is

   $$p_{1g}\left(0,\sigma_2\right)-e=\frac{e\left(2\left(1-2\alpha\right)\left(l\alpha-1\right)-\sigma_2\left(3\alpha+l\alpha-2\right)\right)}{4\left(1-\sigma_2+\alpha\sigma_2-2\alpha\right)}. \tag{57}$$

   In order that margin and demand be positive at the same time for the green firm falsely perceived to be brown, a necessary condition is $\sigma_2>\underline{\sigma}=\frac{2(1-2\alpha)(l\alpha-1)}{3\alpha+l\alpha-2}$. Note that $1-\sigma_2+\alpha\sigma_2-2\alpha<0$ and $3\alpha+l\alpha-2>0$ when $\alpha\geq\frac{1}{2}$, and hence $\underline{\sigma}>0$ for all $\frac{1}{2}<\alpha<\frac{1}{l}$.

   Let us consider

   $$\widetilde{\Delta}_{1g}\left(\sigma_2\right)=0 \tag{58}$$

   This is a quadratic equation in σ_2 with at most two real roots, $\overline{\sigma}^-$ and $\overline{\sigma}$, such that $\overline{\sigma}^-<\overline{\sigma}$ whenever they exist. Furthermore, $\widetilde{\Delta}_{1g}\left(\sigma_2\right)$ is a concave function of σ_2 because its second derivative with respect to σ_2 is negative. Thus, $\widetilde{\Delta}_{1g}\left(\sigma_2\right)>0$ if the discriminant $D\left(l\right)$ of (58) is positive and σ_2 lies inside $[\overline{\sigma}^-,\overline{\sigma}]$. It turns out that $D\left(l\right)>0$ for all $l>\frac{1-\alpha-\alpha^2}{1-3\alpha+3\alpha^2}$, which is satisfied for any $l>1$ because $1-\frac{1-\alpha-\alpha^2}{1-3\alpha+3\alpha^2}=\frac{2\alpha(2\alpha-1)}{1-3\alpha+3\alpha^2}>0$.

   The calculations done by Mathematica produce the following expressions

$$\overline{\sigma}\,(\text{resp. }\overline{\sigma}^{-}) = 2\frac{2-\left(8+l+l^2\right)\alpha+\left(9+3l+4l^2\right)\alpha^2-2\left(1+l+2l^2\right)\alpha^3 \\ +(\text{resp. }-)\sqrt{(l-1)^2(l+1)(1-2\alpha)^2\alpha^2(l-1+(1-3l)\alpha+(1+3l)\alpha^2)}}{((3+l)\alpha-2)^2}.$$

From these expressions, further calculations show that $\overline{\sigma}^{-} < \underline{\sigma} < \overline{\sigma} < 1$ for all $\alpha > \frac{1}{2}$. We can conclude that $\overline{\sigma}$ is the critical σ_2 below which separation is possible.

2. Monitoring accuracy is intermediate: $\alpha \in \left[\overline{\alpha}, \frac{1}{2}\right]$

For all $\sigma_2 > \underline{\sigma}$, the best worst-profit is

$$\widetilde{\Pi}_{1g}(\sigma_2) = \frac{\left(e\left(\left((l+5)\alpha-2-4l\alpha^2\right)\sigma_2+2(2\alpha-1)(l\alpha-1)\right)\right)^2}{16le\alpha(1-2\alpha)(\sigma_2(3\alpha-1)+1-2\alpha)}, \tag{59}$$

and the margin profit is

$$p_{1g}(0,\sigma_2) - e = \frac{e\left(\left(2-(5+l)\alpha+4l\alpha^2\right)\sigma_2-2(2\alpha-1)(l\alpha-1)\right)}{4(\sigma_2(3\alpha-1)+1-2\alpha)}. \tag{60}$$

In order that both the margin and demand be positive for the green firm falsely perceived to be brown, a necessary condition is $\sigma_2 > \underline{\sigma} = \frac{2(2\alpha-1)(l\alpha-1)}{2-(5+l)\alpha+4l\alpha^2}$. Note that $2-(5+l)\alpha+4l\alpha^2 > 0$ and so $\underline{\sigma} > 0$ for all $\alpha < \min\{\frac{1}{2}, \frac{1}{l}\}$ and $\underline{\sigma} < 1$.

Equation (58) is quadratic in σ_2 with at most two real roots, $\overline{\sigma}^{-}$ and $\overline{\sigma}$, such that $\overline{\sigma}^{-} < \overline{\sigma}$ whenever they exist. Furthermore, $\widetilde{\Delta}_{1g}(\sigma_2)$ is a concave function of σ_2 because its second derivative with respect to σ_2 is negative. Thus, $\widetilde{\Delta}_{1g}(\sigma_2) > 0$ if the discriminant $D(l)$ of (58) is positive and σ_2 lies inside $[\overline{\sigma}^{-}, \overline{\sigma}]$. $D(l)$ is a quadratic and convex function of l, such that $D(l) > 0$ when $l \in \left(1, \frac{1}{1-2\alpha}\right)$. The highest root in l of equation $D(l) = 0$ is given by $\frac{\alpha(1-2\alpha)+\sqrt{(8\alpha-3)(1-4\alpha+3\alpha^2)^2}}{-1+7\alpha-15\alpha^2+8\alpha^3}$: it does exist for any $\alpha > \frac{3}{8}$ and falls short of 1. Otherwise, $D(l) > 0$ when $\alpha \leq \frac{3}{8}$. Finally, $D(l) > 0$ when $\frac{(l-1)}{2l} < \alpha \leq \frac{1}{2}$ and so both roots $\overline{\sigma}^{-}$ and $\overline{\sigma}$ exist.

The calculations done by Mathematica produce the following expressions

$\overline{\sigma}\,(\text{resp. }\overline{\sigma}^{-})$

$$= 2\frac{2-\left(10+l+l^2\right)\alpha+\left(15+5l+6l^2\right)\alpha^2-6\left(1+l+2l^2\right)\alpha^3+8l^2\alpha^4 \\ +(\text{resp. }-)\sqrt{(l-1)^2(1-2\alpha)^2\alpha^2(3-11\alpha+9\alpha^2+2l\alpha(1-2\alpha)+l^2(1-7\alpha+15\alpha^2-8\alpha^3))}}{(2-(5+l)\alpha+4l\alpha^2)^2}.$$

From these expressions, further calculations show that, for all $\alpha \in \left[\overline{\alpha}, \frac{1}{2}\right]$, $\overline{\sigma}^{-} < \underline{\sigma} < \overline{\sigma}$, and, moreover, $0 < \overline{\sigma} < 1$ when $\alpha > \max\left\{\overline{\alpha}, \frac{3}{8}\right\}$. We can conclude that $\min\{\overline{\sigma}, 1\}$ is the critical σ_2 below which separation is possible.

3. Monitoring accuracy is bad: $\alpha < \overline{\alpha}$

For all $\sigma_2 > \underline{\sigma}$, the best worst-profit is

$$\widetilde{\Pi}_{1g}(\sigma_2) = \frac{e\left(\left(1-3\alpha+l\alpha^2\right)\sigma_2-(2\alpha-1)(l\alpha-1)\right)^2}{4l\alpha(1-2\alpha)(\sigma_2(3\alpha-1)+1-2\alpha)}, \tag{61}$$

and the margin profit is

$$p_{1g}(0,\sigma_2) - e = \frac{e\left(\left(1-3\alpha+l\alpha^2\right)\sigma_2 - (2\alpha-1)(l\alpha-1)\right)}{2\left(\sigma_2(3\alpha-1)+1-2\alpha\right)}. \tag{62}$$

In order that both the margin and demand be positive for the green firm falsely perceived to be brown, a necessary condition is $\sigma_2 > \underline{\sigma} = \frac{(2\alpha-1)(l\alpha-1)}{1-3\alpha+l\alpha^2}$. Note that $1-3\alpha+l\alpha^2 > 0$ and hence $\underline{\sigma} > 0$ for all $\alpha < \min\{\frac{1}{2}, \frac{1}{l}\}$.

Equation (58) is quadratic in σ_2 with at most two real roots, $\overline{\sigma}^-$ and $\overline{\sigma}$, such that $\overline{\sigma}^- < \overline{\sigma}$ whenever they exist. Furthermore, $\widetilde{\Delta}_{1g}(\sigma_2)$ is a concave function of σ_2 because its second derivative with respect to σ_2 is negative. Thus, $\widetilde{\Delta}_{1g}(\sigma_2) > 0$ provided that the discriminant $D(l)$ of (58) is positive. This turns out to be true when $\alpha < \frac{(l-1)}{2l}$, or, equivalently, $l > \frac{1}{1-2\alpha}$, because, first, $D(l)$ is a quadratic and convex function of l, and second, $\frac{1}{1-2\alpha}$ exceeds $\frac{1}{1-\alpha}$, which turns out to be the highest root in l of equation $D(l) = 0$. Thus, both roots $\overline{\sigma}^-$ and $\overline{\sigma}$ do exist.

The calculations done by Mathematica produce the following expressions

$$\overline{\sigma}\,(\text{resp. } \overline{\sigma}^-) = \frac{1-(5+l)\alpha+\left(6+8l-2l^2\right)\alpha^2+l(11l-18)\alpha^3-4l(5l-3)\alpha^4+12l^2\alpha^5 + (\text{resp. } -)\,2\sqrt{\alpha^3(l-2l\alpha)^2\left(\begin{array}{c}-1+6\alpha-12\alpha^2+9\alpha^3+l(1-8\alpha+24\alpha^2-34\alpha^3+18\alpha^4)\\ +l^2\alpha\left(1-7\alpha+19\alpha^2-22\alpha^3+9\alpha^4\right)\end{array}\right)}}{(1-3\alpha+l\alpha^2)^2}.$$

From these expressions, further calculations show that, for all $\alpha < \overline{\alpha}$, $\overline{\sigma}^- < \underline{\sigma} < \overline{\sigma}$, and $0 < \overline{\sigma} \leq 1$ when α lies outside $\left[\frac{5l-\sqrt{l(32-7l)}}{16l}, \min\left\{\overline{\alpha}, \frac{5l+\sqrt{l(32-7l)}}{16l}\right\}\right]$. Moreover, $\overline{\sigma}$ reaches a minimum of 0 at $\alpha = \frac{3l-\sqrt{l(9l-16)}}{8l} < \frac{1}{4}$ for all $l \geq 2$. We can conclude that $\min\{\overline{\sigma}, 1\}$ is the critical σ_2 below which separation is possible.

References

[1] **Akerlof, G. A.** 1970. The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism, *Quarterly Journal of Economics* 84, 488–500.

[2] **Bagwell, K., and M. H. Riordan.** 1991. High and Declining Prices Signal Product Quality, *American Economic Review* 81, 224–239.

[3] **Baksi, S., and P. Bose.** 2007. Credence Goods, Efficient Labelling Policies, and Regulatory Enforcement, *Environmental and Resource Economics* 37, 2, 411–430.

[4] **Baron, D. P**. 2010. Morally Motivated Self-Regulation, *American Economic Review* 100, 4, 1299–1329.

[5] **Besley, T., and M. Ghatak**. 2007. Retailing Public Goods: The Economics of Corporate Social Responsibility, *Journal of Public Economics* 91, 9, 1645–1663.

[6] **Cho, I-K., and D. Kreps.** 1987. Signalling Games and Stable Equilibria, *Quarterly Journal of Economics* 102, 179–221.

[7] **Daley, B., and B. Green.** 2014. Market Signaling with Grades, *Journal of Economic Theory* 151, 114–145.

[8] **Darby, M. R., and E. Karni.** 1973. Free Competition and the Optimal Amount of Fraud, *The Journal of Law & Economics* 16, 1, 67–88.

[9] **Daughety A. F., and J. F. Reinganum**. 2008. Imperfect Competition and Quality Signalling, *RAND Journal of Economics* 39, 1, 163–183.

[10] **Delmas, M. A., and V. C. Burbano**. 2011. The Drivers of Greenwashing, *California Management Review* 54, 64–87.

[11] **Feddersen, T. J., and T. W. Gilligan.** 2001. Saints and Markets: Activists and the Supply of Credence Goods, *Journal of Economics & Management Strategy* 10, 149–171.

[12] **Hamilton, S. F., and D. Zilberman**. 2006. Green Markets, Eco-Certification, and Equilibrium Fraud, *Journal of Environmental Economics and Management* 52, 627–644.

[13] **Heyes, A. and S. Martin.** 2017. Social Labeling by Competing NGOs: A Model with Multiple Issues and Entry, *Management Science* 63, 6, 1657–2048.

[14] **Jacquet, J., D. Pauly, D. Ainley, S. Holt, and J. Jackson.** 2010. Seafood Stewardship in Crisis, *Nature* 467, 28–29, doi:10.1038/467028a.

[15] **Janssen, C. W. J and S. Roy.** 2010. Signaling Quality Through Prices in an Oligopoly, *Games and Economic Behavior* 68, 1, 192–207.

[16] **Kerschbamer, R., and M. Sutter**. 2017. The Economics of Credence Goods – a Survey of Recent Lab and Field Experiments, *CESifo Economic Studies* 63, 1, 1–23.

[17] **Klein, B., and K. B. Leffler.** 1981. The Role of Market Forces in Assuring Contractual Performance, *Journal of Political Economy* 89, 4, 615–641.

[18] **Lyon, T. P., and J. W. Maxwell**. 2004. Corporate Environmentalism and Public Policy. Cambridge, UK: Cambridge University Press.

[19] **Mahenc, P.** 2008. Signalling the Environmental Performance of Polluting Products to Green Consumers, *International Journal of Industrial Organization* 26, 59–68.

[20] **Mahenc, P.** 2017. Honest versus Misleading Certification, *Journal of Economics & Management Strategy* 26, 2, 454–483.

[21] **Mailath, G., J.** 1988. An Abstract Two-Period Game with Simultaneous Signaling—Existence of Separating Equilibria, *Journal of Economic Theory* 46, 2, 373–394.

[22] **Milgrom, P., and J. Roberts.** 1986. Price and Advertising Signals of Product Quality, *Journal of Political Economy* 94, 796–821.

[23] **Nelson, P.** 1970. Advertising as Information, *Journal of Political Economy* 82, 4, 729–754.

[24] **Salop, S. C.** 1979. Monopolistic Competition with Outside Goods, *The Bell Journal of Economics* 10, 1, 141–156.

[25] **Shapiro, C.** 1983. Premiums for High Quality Products as Returns to Reputations, *The Quarterly Journal of Economics* 98, 4, 659–680.

[26] **Spence, A. M.** 1973. Job Market Signalling, *Quarterly Journal of Economics* 87, 355–374.