Greenwashing or Green Innovation? Signaling Games under Environmental Monitoring

March 15, 2026

Abstract

This paper incorporates revenue-based greenwashing monitor into a signaling model of ecolabeling to study firms' certification and clean technology investment decisions. A monopolistic firm, either clean or dirty, chooses whether to adopt an ecolabel and whether to invest in clean production, while consumers update beliefs based on observed labeling behavior. Three equilibrium regimes emerge depending on the penalty rate. A high penalty sustains a separating equilibrium in which only the clean firm certifies. A low penalty leads to pooling, and greenwashing persists. A novel intermediate regime arises when the penalty is moderate: too weak for separation, yet strong enough to induce genuine clean investment by the dirty firm. This innovation regime dominates separation in welfare terms when environmental damage is large. Certification fees and penalties act as policy substitutes. Calibrated to the EU Green Claims Directive, simulations show the current four percent rule achieves separation only in a narrow parameter region, suggesting it is insufficient as a standalone instrument.

Key words: Signaling game; Greenwashing; Green innovation; Environmental regulation; Game theory

1 Introduction

Green labels are increasingly used by firms to signal their environmental performance. However, the growing prevalence of greenwashing—where polluting

firms adopt environmental labels without making real improvements—has raised concerns about the credibility of such signals. Although certification agencies verify compliance at the time of labeling, misreporting and non-compliance may persist without effective regulatory enforcement. In response, governments in several regions, including the EU and the UK, have introduced revenue-based penalties to deter greenwashing. Whether such penalties are sufficient to sustain credible signaling, however, remains an open question.

Ecolabels are designed to help environmentally conscious consumers quickly and directly identify products produced by environmentally friendly firms. By providing clear and observable information, ecolabels aim to reduce information asymmetry in the market, and their adoption is therefore widely encouraged. Well-known examples include the Dolphin-Safe label in the tuna industry, the EU Ecolabel for consumer products, forest and fishery certifications such as the Forest Stewardship Council (FSC) and the Marine Stewardship Council (MSC), and the Energy Star program for energy-efficient appliances. In recent years, firms have increasingly adopted environmental labels in response to rising consumer demand for sustainability. At the same time, the credibility of such labels has been increasingly questioned. In 2021, the European Commission investigated misleading green claims by multinational firms and found that more than 40% of sustainability claims were potentially exaggerated or unfounded. These developments highlight the need for regulatory mechanisms that can distinguish genuinely green firms from those engaging in strategic misrepresentation.

Regulatory authorities worldwide have recently tightened enforcement against greenwashing. In the European Union, under the upcoming Green Claims Directive and related legislation, fines for misleading environmental claims can reach at least 4% of a firm's annual turnover within the EU. In the UK, the Competition and Markets Authority (CMA) has been granted enhanced enforcement powers under the Digital Markets, Competition and Consumers Act 2024, allowing it to directly impose fines of up to 10% of a firm's global turnover. In Japan, since October 2024, firms making false or unsubstantiated environmental claims may be ordered to cease such practices and, in some cases, face surcharges of

up to 3% of the relevant sales. Notably, these regulatory frameworks commonly link penalties to firms' revenue or sales. Motivated by this institutional reality, this paper introduces a revenue-based penalty for greenwashing into the model.

Whether such penalties work, and through which channel, is an open question. A penalty can deter greenwashing in two distinct ways. First, it can make mimicry too costly for dirty firms, inducing honest disclosure and restoring label credibility, which is the separation channel. Second, even when a penalty falls short of achieving separation, it can make genuine investment in clean technology worthwhile for dirty firms, turning strategic misrepresentation into real environmental improvement, that is the innovation channel. Standard greenwashing models focus exclusively on the first channel. This paper argues that the second is equally important, and that conflating the two leads to misleading policy prescriptions.

We develop a signaling game between a privately informed firm and a regulator. A monopolist knows whether its production technology is clean or dirty. It chooses whether to seek an ecolabel by paying a certification fee, and whether to invest in upgrading its technology. A regulator audits certified firms with some probability and imposes a revenue-based penalty on those whose claims are found to be false. Consumers observe the label but not the underlying type, and update their beliefs accordingly. We characterise the full set of Perfect Bayesian Equilibria satisfying the Intuitive Criterion of Cho and Kreps (1987), which eliminates implausible off-equilibrium beliefs.

Three regimes emerge. When the penalty rate is high, it deters mimicry entirely: only genuinely green firms certify, the label is credible, and consumers can trust it. When the penalty is low, deterrence fails: both types pool under the ecolabel, which becomes uninformative. Between these two cases lies a novel intermediate regime in which the penalty is too weak to induce separation but strong enough to make clean investment worthwhile for dirty firms. In this regime, greenwashing gives way to genuine innovation, and the ecolabel, though still uninformative about types, reflects real environmental improvement. Importantly, we show that this innovation regime can deliver higher social welfare

than separation when environmental damage is severe, because clean investment eliminates the underlying pollution rather than merely signaling its absence.

These results have direct implications for the design of environmental monitoring policy. We show that the penalty rate and the certification fee act as policy substitutes: a regulator with limited enforcement capacity can compensate by raising the certification standard, and vice versa. Calibrating to the EU Green Claims Directive, we find that the 4% revenue threshold achieves label separation only when certification fees are already high, suggesting the rule is insufficient as a standalone instrument across most market configurations. A more effective policy combines moderate penalties with subsidised certification to shift firms towards the innovation regime when full separation is infeasible.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the model and analyzes equilibria in the absence of penalties. Section 4 incorporates the audit mechanism and derives the conditions under which greenwashing is deterred and clean investment is encouraged. Section 5 studies incentives for investing in green technology under the greenwashing audit. Section 6 presents the policy analysis and numerical simulations. Section 7 concludes.

2 Related Literature

This paper sits at the intersection of three bodies of work: the literature on signaling and information asymmetry, the ecolabeling and certification literature, and recent work on greenwashing fraud and green investment. We review each in turn and situate our contribution.

Firstly, the theoretical foundation of this paper draws on the canonical signaling framework of Spence (1978) and Spence (2002), in which an informed party can credibly transmit private information to uninformed parties through costly actions. Fudenberg and Tirole (1991) formalise the equilibrium concepts of separating, pooling, and semi-separating equilibria in this setting, while Cho and Kreps (1987) propose the Intuitive Criterion to discipline off-equilibrium

beliefs that the refinement we adopt throughout. Green markets are a natural application of this framework, since environmental attributes are credence goods: consumers cannot verify them even after purchase, and certifiers face significant information asymmetry about firms' true production processes (Nygaard (2023)). Real-world examples illustrate this clearly. The EU Ecolabel system was found in 2020 to have awarded certifications to firms not meeting their own standards, and the Forest Stewardship Council was accused in 2019 of certifying illegal logging operations in the Brazilian Amazon. These failures show that certification alone is insufficient and that post-certification auditing and penalties are a necessary complement. Ecolabeling and certification. A substantial literature examines how ecolabels affect firms' pricing and production decisions under various market structures. Amacher et al. (2004) show that a green producer can signal product quality through a higher price even in the absence of certification, and find that green fraud can arise under competitive conditions. Hamilton and Zilberman (2006) analyse equilibrium fraud in green markets, showing how eco-certification interacts with firm incentives to misrepresent environmental quality. Ibanez and Grolleau (2008) examine whether ecolabeling schemes can preserve the environment in competitive settings, identifying conditions under which labels may be counterproductive. Podhorsky (2010) studies how the design of ecolabel standards affects their informativeness and firms' compliance incentives, showing that stricter standards do not always improve welfare. Heyes et al. (2020) study green markets under incomplete information and find that the effectiveness of certification depends critically on the credibility of the certifying body. Our paper complements this literature by shifting focus from the design of certification standards to the role of post-certification enforcement, specifically, revenue-based penalties, which none of these papers considers. Green investment and clean technology. A separate strand examines firms' incentives to invest in cleaner technology, particularly when green consumers are present. Sengupta (2012) analyses how emission liability stringency affects investment in clean technology in monopoly settings with environmentally conscious consumers, and Sengupta (2015) extends this

to competitive environments. Sengupta (2024) compares firms' incentives to invest in green R&D in markets with and without ecolabels, finding that label availability can either stimulate or crowd out clean investment depending on market structure. Wu et al. (2020) examine how information transparency and corporate social responsibility influence greenwashing behaviour, finding that greater transparency reduces misrepresentation only when consumer awareness is sufficiently high. These models typically treat certification fees as negligible and abstract from revenue-based penalties. Our paper introduces both, showing how their interaction generates a novel innovation regime that this literature has not identified.

Closest to this paper is Arguedas and Blanco (2025), who study fraud and certification in a monopoly model where firms choose between standard production, uncertified green claims, and certified green production. They show that lower green production costs can paradoxically increase fraud, and that reducing certification fees shifts firms towards certified strategies. Our paper complements theirs by modelling revenue-based penalties consistent with the EU Green Claims Directive, and by introducing an explicit green investment decision, which generates a third equilibrium regime in which moderate penalties induce genuine innovation rather than label separation. Garrido et al. (2020) study whether mandatory certification promotes greenwashing using a signaling approach, showing that when both firm types find certification profitable, mandatory certification can paradoxically expand pooling equilibria in which non-innovative firms certify falsely. We build on these frameworks in two directions. First, we endogenise the firm's technology choice: rather than treating types as fixed, we allow dirty firms to invest in clean production. Second, we model penalties as proportional to revenue, consistent with the EU Green Claims Directive, rather than as fixed amounts. This yields sharper and more policy-relevant comparative statics for the design of environmental monitoring regimes. Taken together, this paper bridges the ecolabeling literature and the signaling literature by incorporating both certification fees and revenue-based penalties in a unified framework. To the best of my knowledge, this is the first

paper to show that a revenue-based greenwashing penalty can induce genuine green innovation within a pooling equilibrium, and to derive the conditions under which a regulator should prefer the innovation regime over label separation.

3 Model

There is a monopolist whose environmental type is assigned by nature: with probability μ the firm is clean, and with probability $1 - \mu$ it is dirty. The type is private information known only to the firm; consumers hold a common prior belief μ that the firm is clean. Let $i \in \{c, d\}$ denote the firm's type.

The clean type produces clean goods at marginal cost m_c. The dirty type produces polluting goods at marginal cost $m_d < m_c$. At Stage 1, the dirty type may invest in clean technology at cost I, which raises its marginal cost from m_d to m_c and allows it to produce clean goods. Let $v \in \{V, NV\}$ denote the dirty type's investment decision, where V denotes green investment and NV denotes no investment.[1]

At Stage 2, each firm decides whether to adopt the ecolabel (L) or not (NL), at certification cost $f > 0$. Both types may apply for the label. Greenwashing occurs when the dirty type adopts the ecolabel without having invested in clean technology.

Consumers observe the labeling decision but not the firm's type. They update their beliefs using Bayes' rule. Let $b = \Pr\{\text{clean} \mid L\}$ and $b' = \Pr\{\text{clean} \mid NL\}$ denote the posterior beliefs upon observing a label and no label, respectively, with $b, b' \in [0, 1]$. Consumers value dirty products at V and clean products at $V + b\Delta$, where $\Delta > 0$ is the green premium reflecting their willingness to pay for environmentally friendly goods. The monopolist sets prices equal to consumer willingness to pay, so the labeled price is $p^L = V + b\Delta$ and the non-labeled price is $p^{NL} = V + b'\Delta$. Normalising quantity to one unit, revenue equals price.

[1] The firm's inherent type does not change: a dirty type that invests produces clean goods but remains the dirty type for all payoff calculations. The clean type has no incentive to invest.

At Stage 4, a regulator audits labeled firms with probability $\phi \in (0, 1)$. If a labeled firm is found to be greenwashing, it pays a penalty equal to $\phi \cdot \zeta \cdot p^L$, where $\zeta \in (0, 1)$ is the penalty rate.[2] A clean type that is audited passes at no cost.

The timeline of the game can be shown as:

- **Stage 0.** Nature assigns the firm's type: clean with probability μ, dirty with probability $1 - \mu$.
- **Stage 1.** The dirty type chooses $v \in \{V, NV\}$. The clean type produces clean goods automatically.
- **Stage 2.** Each firm chooses $s \in \{L, NL\}$ and pays f if it labels.
- **Stage 3.** Consumers observe the labeling decision, update beliefs to $(b,\ b')$ using Bayes' rule, and the firm sets price accordingly.
- **Stage 4.** The regulator audits labeled firms with probability ϕ and imposes the penalty $\phi\zeta p^L$ on any firm found to be greenwashing.

A **separating equilibrium** exists when the two types choose different labeling strategies, so consumers can perfectly infer the firm's type from the label. A **pooling equilibrium** exists when both types make the same labeling decision, so the label carries no type information. We solve the game by backward induction for a Perfect Bayesian Equilibrium (PBE), and apply the Intuitive Criterion (Cho and Kreps, 1987) to eliminate equilibria supported by implausible off-path beliefs.[3]

3.1 Benchmark: Without Greenwashing Penalty

Without the greenwashing penalty, will both types adopt the ecolabel, and does the dirty type have the incentive to invest in the clean technology? Considering

[2] f is paid to a third-party certification body; whether it accrues to the regulator or not does not affect any equilibrium.

[3] The formal application of the Intuitive Criterion to each case is in the Appendix.

the certification cost, the clean type may not be willing to adopt the ecolabel since the dirty type can mimic it costless. Now consider two types of equilibria.

Stage 3. Consumers beliefs.

Holding the prior belief μ, consumers update their beliefs after observing the signal, label or not. If both firms adopt the ecolabel, consumers cannot distinguish the firm's type and generate the posterior belief regarding to Bayes' rule:

$$b = Pr(clean|L) = \frac{\mu \cdot 1}{\mu \cdot 1 + (1-\mu) \cdot 1} = \mu$$

This implies that consumers conceive the labeling firm as the clean type with probability of mu, while as the dirty type with the probability of $(1-\mu)$. If neither type adopts the label, Bayes' rule provides:

$$b' = Pr(clean|NL) = \frac{\mu \cdot 1}{\mu \cdot 1 + (1-\mu) \cdot 1} = \mu$$

When the clean type adopts the label while the dirty type does not, consumers change their mind:

$$b = Pr(clean|L) = \frac{\mu \cdot 1}{\mu \cdot 1 + (1-\mu) \cdot 0} = 1$$

Consumers conceive the labeling type is the clean type certainly. When the clean type does not adopt the ecolabel, while the dirty type greenwashes, Bayes' rule gives:

$$b' = Pr(clean|NL) = \frac{\mu \cdot 0}{\mu \cdot 0 + (1-\mu) \cdot 1} = 0$$

- **Stage 2: Ecolabel decision.**

The two types choose to whether adopt the ecolabel by comparing the payoffs in two cases. Following the model, consumers' willingness to pay for the labeled

product and for the non-labeled product can be expressed by [4]:

$$p^L = V + b\Delta, b = Pr\{clean|L\}$$

$$p^{NL} = V + b'\Delta, b' = Pr\{clean|NL\}$$

For the clean type, the payoffs in labeling and non-labeling can be denoted by

$$\pi_c(L) = V + b\Delta - m_c - f$$

$$\pi_c(NL) = V + b'\Delta - m_c$$

When $\pi_c(L) > \pi_c(NL)$, that is when $0 < f < (b - b')\Delta$, the clean type has a positive incentive to adopt the ecolabel. We can define the condition as

$$C_1 := 0 < f < (b - b')\Delta$$

For the dirty type, the payoffs in two cases are:

$$\pi_d(L) = V + b\Delta - m_d - f$$

$$\pi_d(NL) = V + b'\Delta - m_d$$

Therefore, when $\pi_d(L) > \pi_d(NL)$, that is when $0 < f < (b - b')\Delta$, the dirty type has the positive incentive to mimic the clean type. Since the incentive condition $f < (b - b')\Delta$ is identical for both types, the labeling incentive of the clean type is always matched by the dirty type. We now characterize each candidate equilibrium by imposing Bayes' rule to pin down (b, b') and then checking the IC refinement.

Case 1. Pooling Equilibrium with ecolabel.

Next, we characterize possible equilibrium. Firstly, under the pooling equilibrium where both labeling strategy, the posterior belief $b = \mu$ according to results in stage 3. In this case, the off-path strategy is non-labeling and the

[4] The willingness to pay equals the price according to the monopolist market structure.

off-path belief that the non-labeling firm is the clean type is b'. Therefore, the condition C_1 can be rewritten by

$$f < (\mu - b')\Delta$$

This identical incentive condition shows that if the clean type wants to stay in the pool, so does the dirty type. The pooling equilibrium exists as a PBE for any $0 < b' < \mu - \frac{f}{\Delta}$. We can then apply IC refinement to examine deviation incentives for both types. A type can be ruled out as deviator if, even under the most favorable belief $b' = 1$, deviation is not profitable [5]. For the clean type, when $b' = 1$:

$$\pi_c(NL)|_{b'=1} - \pi_c(L) = (V + \Delta - m_c) - (V + \mu\Delta - m_c - f) = (1 - \mu)\Delta + f > 0$$

This shows that under the most favorable condition of deviation where consumers conceive the non-labeling type as the clean type with the probability of 1, the clean type would deviate. Similarly, when $b' = 1$ for the dirty type,

$$\pi_d(NL)|_{b'=1} - \pi_d(L) = (V + \Delta - m_d) - (V + \mu\Delta - m_d - f) = (1 - \mu)\Delta + f > 0$$

The dirty type is the potential deviator as well. Since neither type can be ruled out as deviator, IC cannot restrict b' to 0 or 1, and b' is free with $b' \in [0, 1)$. Therefore, the pooling equilibrium survives the IC for some admissible b'.

Case 2. Pooling equilibrium without ecolabel.

In the case where neither type labels, the payoffs of for the clean type can be shown as:

[5] Following the Intuitive Criterion (Cho and Kreps, 1987), to determine whether a type can be ruled out as a potential deviator, we evaluate deviation incentives under the most favorable off-path belief $b' = 1$. A type i is equilibrium dominated if deviation is unprofitable even under this best-case belief, i.e., $\max_{b' \in [0,1]}[\pi_i(dev) - \pi_i(equi)] < 0$. Since the maximum of $\pi_i(dev) = \pi_i(NL)$ over b' is achieved at $b' = 1$, checking $b' = 1$ is both necessary and sufficient to determine whether a type can be eliminated as a deviator. If neither type is equilibrium dominated, the Intuitive Criterion places no restriction on b', and the off-path belief remains free.

$$\pi_c(NL) = V + b'\Delta - m_c$$

$$\pi_c(L) = V + b\Delta - m_c - f$$

where $b \in [0, 1]$ is now the off-path posterior belief upon observing L. By comparing the $\pi_c(L)$ and $\pi_c(NL)$, the existence condition of the equilibrium is give by:

$$C_2 := f > (b - \mu)\Delta$$

where $b' = \mu$ according to Bayes' rule in stage 3. Similarly, for the dirty type, when the condition C_2 is met, $\pi_d(NL) > \pi_d(L)$. The incentive conditions are identical for both types, implying that when the certification fee is higher than the difference green premium between labeling and non-labeling, neither type adopts the ecolabel.

If the clean type deviates to choose labeling strategy, the off-path payoff becomes:

$$\pi_c^{off-path}(L) = V + b\Delta - m_c - f$$

where the off-path posterior belief $b \in [0, 1]$. Therefore, with the most favorable condition of belief $b = 1$,

$$\pi_c(NL)|_{b=1} - \pi_c(L) = (V + \mu\Delta - m_c) - (V + \Delta - m_c - f) = (\mu - 1)\Delta + f$$

When $f > (1-\mu)\Delta$, $\pi_c(NL) > \pi_c(L)$, the clean type will stay at the equilibrium without ecolabel. However when $f < (1 - \mu)\Delta$, the clean type has an incentive to deviate. Similarly, when the dirty type deviates to adopt the ecolabel, the payoff can be given by

$$\pi_d^{off-path}(L) = V + b\Delta - m_d - f$$

Thus, with $b = 1$,

$$\pi_d(NL)|_{b=1} - \pi_d(L) = (V + \mu\Delta - m_d) - (V + \Delta - m_d - f) = (\mu - 1)\Delta + f$$

We can find that both types have an incentive when $f < (1-\mu)\Delta$ under $b = 1$, and the IC refinement simply cannot assign b to either type specifically. Also, when $f > (1 - \mu)\Delta$, neither type wants to deviate. In both cases, the Intuitive Criterion places no restriction on b, and the pooling-without-label equilibrium survives the refinement for any admissible $b \in [0, \mu + \frac{f}{\Delta})$.

The separating equilibrium where the only the clean type adopts the label cannot exist since when the clean type adopts the label, the dirty type will always benefit from mimicing by adopting the label as well. In the case where only the dirty type adopts the ecolabel is not stable, since consumers can distinguish the non-labeling type is the clean type. Therefore, it is always beneficial for the dirty type to mimic the clean type. The analysis is shown in appendix A.1.

- **Stage 1: Investment Decision.**

The dirty type may have an incentive to update the production technology to produce clean product. Before comparing investment strategies, we note that the strategy (NL, V) that investing in clean technology while not adopting the ecolabel, is strictly dominated for the dirty type and can be eliminated from consideration. Without the label, the dirty type earns no green premium regardless of its underlying technology, so investment raises production cost from m_d to m_c without any revenue benefit:

$$\pi_d(NL, V) \;=\; V + b'\Delta - m_c - I \;<\; V + b'\Delta - m_d \;=\; \pi_d(NL, NV).$$

We therefore restrict attention to three strategies for the dirty type: (L, V), (L, NV), and (NL, NV).

If the dirty type chooses to invest in green technology with the cost I, the payoffs can be denoted by:

$$\pi_d(I, L) = V + \mu\Delta - m_c - f - I$$

$$\pi_d^G(NI, L) = V + \mu\Delta - m_d - f$$

$\pi_d(I, L)$ denotes the payoff when the labeling dirty type invests in green innovation, and the marginal production cost increases from m_d to m_c. The $\pi_d^G(NI, L)$ is the payoff of greenwashing that it remains dirty with m_c however adopting the label. Since $\pi_d(I, L) < \pi_d^G(NI, L)$ always exists with $m_c > m_d$, without the greenwashing audit, we can conclude that the dirty type will not invest in the clean technology.

Proposition 1 *In the absence of the greenwashing monitor:*

- *A pooling equilibrium exists in which both types adopt the ecolabel when* $f < (\mu - b')\Delta$.
- *A pooling equilibrium exists where neither type adopts the ecolabel when* $f > (b - \mu)\Delta$.
- *A separating equilibrium does not exist.*
- *The dirty type never invests in clean technology.*

Proof. See Appendix.

The key insight is that, without audit, the certification fee f is the only cost distinguishing labeling from non-labeling behavior. Since f is identical for both types, it creates no asymmetry in their labeling incentives, whenever the clean type finds it worthwhile to adopt the ecolabel. When the green premium $(\mu - b')\Delta$ exceeds f, the dirty type finds mimicry equally worthwhile, since it faces the same fee but enjoys a strictly lower production cost $m_d < m_c$. The ecolabel thus functions as a revenue-enhancing tool rather than a credible environmental signal. The dirty type therefore has no incentive to invest in clean technology either, since it can obtain the same pooling green premium $\mu\Delta$ by

simply attaching the label without incurring the investment cost I. The market consequently fails to deliver either credible information or genuine environmental improvement, which motivates the introduction of a greenwashing audit in Section 4.

4 Greenwashing under Monitor

We now introduce the audit regime into the model. The regulator audits labeled firms with probability ϕ and, if a firm is found to be greenwashing, imposes a penalty $\phi\zeta p^L$ proportional to its revenue. The key change relative to the benchmark is that the dirty type's labeling payoff now includes this expected penalty cost, while all other payoffs and all consumer beliefs are unchanged.

- **Stage 3. Consumers' Beliefs.**

Consumer beliefs (b, b') are determined by Bayes' rule exactlyas in Section 3.

- **Stage 2. Ecolabel Decision.**

Since the audit targets only labeled firms, the penalty does not affect the clean type's payoffs or the non-labeling payoff of the dirty type. Only the dirty type's labeling payoff changes:

$$\pi_c(L) = V + b\Delta - m_c - f,$$

$$\pi_c(NL) = V + b'\Delta - m_c,$$

$$\pi_d(L) = V + b\Delta - m_d - f - \phi\zeta(V + b\Delta),$$

$$\pi_d(NL) = V + b'\Delta - m_d.$$

Comparing payoffs, the clean type prefers L when

$$C_0: \quad f < (b - b')\Delta,$$

which is the same condition as in the benchmark. The dirty type prefers L when

$$D_0: \quad f < (b - b')\Delta - \phi\zeta(V + b\Delta).$$

Since $D_0 \subset C_0$ (the penalty term makes D_0 strictly tighter), whenever the dirty type is willing to label, so is the clean type, but not vice versa. This asymmetry is the central difference from the benchmark: the penalty breaks the symmetric incentives and opens the possibility of a separating equilibrium.

We characterize each candidate equilibrium by imposing Bayes' rule to pin down (b, b'), then verifying the IC refinement.

- **Case 1. Pooling equilibrium with ecolabel.**

Suppose both types label. Bayes' rule gives $b = \mu$ on the equilibrium path. For the clean type, the payoffs for labeling and non-labeling are still:

$$\pi_c^L = V + \mu\Delta - m_c - f, (b = \mu)$$

$$\pi_c^{NL} = V + b'\Delta - m_c$$

The incentive condition for the clean type of labeling is $C_1 : f < (\mu - b')\Delta$. However, the dirty type will pay for the penalty if it does not genuinely produce in a green way.

$$\pi_d(L) = V + \mu\Delta - m_d - f - \phi \cdot \zeta \cdot (V + \mu\Delta), (b = \mu)$$

If it does not adopt the ecolabel, it will not be audited and the payoff is

$$\pi_d^{NL} = V + b'\Delta - m_d$$

Therefore, under the auditing framework, the dirty type prefers ecolabel over

non-label when $(\pi_d(L) - \pi_d(NL)) > 0$, that is when

$$0 < \zeta < \frac{(\mu - b')\Delta - f}{\phi(V + \mu\Delta)}$$

Define the condition $D_1 := 0 < \zeta < \frac{(\mu-b')\Delta-f}{\phi(V+\mu\Delta)}$, and $\zeta^* = \frac{(\mu-b')\Delta-f}{\phi(V+\mu\Delta)}$. Therefore, there exists the pooling equilibrium with both labeling when C_1 and D_1 are met. Since D_1 is stricter than C_1, we can bind the incentive condition as D_1. With the most favorable condition $b' = 1$, the clean type wants to deviate since

$$\pi_c(L) - \pi_c^{off-path}(NL)|_{b'=1} = (\mu - 1)\Delta - f < 0$$

And the dirty type wants to deviate, holding $b' = 1$, since

$$\pi_d^G(L) - \pi_d(NL)|_{b'=1} < 0$$

Both types can benefit from deviation so the IC refinement cannot restrict b' for 0 or 1 to either type. Therefore, the equilibrium can survive with free $b' \in [0, 1)$, when the D_1 can be met.

In this case, the audit penalty is too low to deter the dirty type from mimicking the clean type, although there exist the certification cost. The ecolabel loses its informational content that consumers cannot distinguish types from labeling behavior alone, and the label functions merely as a revenue-enhancing device rather than a credible environmental signal. This is precisely the market failure that motivates stricter enforcement.

- **Case 2. Pooling equilibrium without ecolabel.**

The pooling equilibrium without ecolabel does not survive the Intuitive Criterion. When ζ is sufficiently high, the dirty type no longer benefits from deviating to label (the expected penalty deters it), while the clean type still does benefit. The IC therefore assigns $b = 1$ to any off-path labeler, and the clean type strictly prefers to deviate, unravelling the equilibrium. The full derivation

is in Appendix A.2.

- **Case 3. Separating equilibrium–only the clean type adopts.** Recall $b = 1$ according to the Bayes' rule, the payoffs of the clean type are

$$\pi_c(L) = V + \Delta - m_c - f$$

$$\pi_c(NL) = V - m_c, (b' = 0)$$

Therefore, the clean type prefers L over NL when $C_3 := \Delta > f$. For the dirty type, the payoffs are

$$\pi_d(NL) = V - m_d$$

$$\pi_d(L) = V + \Delta - m_d - f - \phi \cdot \zeta(V + \Delta)$$

Thus,

$$\pi_d(NL) - \pi_d(L) = f - \Delta + \phi \cdot \zeta(V + \Delta)$$

Define $\zeta^{**} = \frac{\Delta - f}{\phi(V+\Delta)}$. When $D_3 := \zeta > \frac{\Delta - f}{\phi(V+\Delta)}$, $\pi_d(NL) > \pi_d(L)$, and the dirty type does not adopt the ecolabel. Therefore, C_3 and D_3 constitute the incentive conditions for the existence of the equilibrium.

To check whether the equilibrium should be ruled out under the IC refinement, we compare the off-path equilibrium to the on-path equilibrium. For the dirty type, the off-path payoff can be denoted by:

$$\pi_d^{off-path}(L) = V + b\Delta - m_d - f - \phi \cdot \zeta(V + b\Delta)$$

At the most favorable condition $b = 1$, we can compare

$$\pi_d^{off-path}(L)|_{b=1} - \pi_d(NL) = \Delta - f - \phi \cdot \zeta(V + \Delta)$$

So when D_3 is met, the dirty type cannot benefit from the deviation.

Under this condition, the IC refinement assigns $b = 1$ to the labeling deviation, but the dirty type still does not deviate.

For the clean type, the deviation of non-labeling produces:

$$\pi_c(NL)^{off-path} = V + b'\Delta - m_c$$

where $b' \in [0, 1]$ denotes the off-path posterior belief upon observing no label. Since the dirty type is already choosing NL on the equilibrium path, NL is not a deviation for the dirty type. The clean type is therefore the only possible deviator to NL. Consequently, the IC refinement assigns $b' = 0$ upon observing no label, and consumers correctly infer that the non-labeler must be the dirty type.

Under this belief assignment $b' = 0$, the clean type's deviation payoff becomes:

$$\pi_c(NL)^{off-path}\big|_{b'=0} - \pi_c(L) = (V - m_c) - (V + \Delta - m_c - f) = f - \Delta < 0$$

where the inequality holds under condition $C_3 := \Delta > f$. Therefore, under the IC-refinement belief $b' = 0$, the clean type does not benefit from deviating. The equilibrium survives the IC refinement whenever C_3 and D_3 are both satisfied.

- **Case 4. Separating equilibrium-only the dirty type labels.**

This equilibrium does not exist under any parameter values. With $b = 0$, the ecolabel carries no green premium, so the dirty type pays f and the expected penalty $\phi\zeta V$ for no revenue gain and always prefers NL. The derivation is in Appendix A.2.

Proposition 2 *Under the greenwashing audit regime with penalty rate ζ and audit probability ϕ, the following equilibria survive the Intuitive Criterion:*

(i) ***Pooling equilibrium with ecolabel*** *exists when D_1 is satisfied, that is,*

$\zeta < \zeta^*$. *Both types adopt the ecolabel, greenwashing persists, and the ecolabel remains uninformative.*

(ii) ***Separating equilibrium*** *exists when* C_3 *and* D_3 *are jointly satisfied, that is,* $\Delta > f$ *and* $\zeta > \zeta^{**}$. *Only the clean type adopts the ecolabel, greenwashing is eliminated, and the ecolabel credibly signals environmental quality.*

(iii) The pooling equilibrium without ecolabel and the separating equilibrium where only the dirty type labels do not survive the Intuitive Criterion under any parameter values.

Proof. See Appendix.

We can figure out different areas of ζ for equilibria in the following figure.

Figure 1 illustrates the equilibrium regions as a function of ζ.

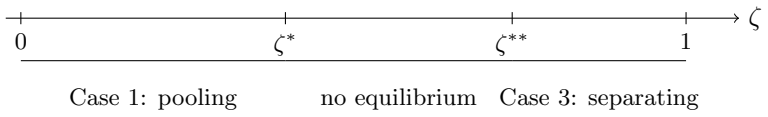


Figure 1: Equilibrium regions as a function of the penalty rate ζ.

The figure illustrates how the penalty rate ζ determines the equilibrium outcome. When $\zeta < \zeta^*$, the expected penalty is too low to deter the dirty type from mimicking the clean type, and a pooling equilibrium with greenwashing persists. When $\zeta > \zeta^{**}$, the penalty is sufficiently stringent to sustain the separating equilibrium: only the clean type finds labeling profitable, and the ecolabel recovers its informational content. In the intermediate region $\zeta^* < \zeta < \zeta^{**}$, neither pure-strategy equilibrium is sustained; analysing mixed-strategy equilibria in this region is left for future work. The two thresholds ζ^* and ζ^{**} depend on different sets of parameters: ζ^* depends on the pooling green premium $\mu\Delta$ and the off-path belief b', while ζ^{**} depends on the full separating green premium Δ. In general, $\zeta^* < \zeta^{**}$ whenever the separating premium exceeds the pooling premium, which holds when $\mu < 1$. A key implication,

explored in Section 5, is that while a penalty $\zeta > \zeta^{**}$ eliminates greenwashing, it does not automatically induce the dirty type to invest in clean technology.

5 Incentives for Green Innovation

Without the ecolabel audit in the benchmark, there are two pooling equilibria, both adopt the label and neither adopts the label. However, the dirty type always chooses to greenwash rather than update production technology. Under the framework of ecolabel audit, we again check the dirty type's innovative incentives.

Based on conclusions in stages 3 and 2, there exist a pooling equilibrium where both type adopt the ecolabel, and the separating equilibrium where only the clean adopts the label. So the following analysis for the stage 1 will be expanded in two cases.

5.1 Incentives when both types adopt the ecolabel.

Under this equilibrium, the dirty type mimics the clean type, however, whether it greenwashes or not depends on payoffs. If the dirty type invests in green technology with the cost I, it will have the payoff:

$$\pi_d(V, L) = V + \mu\Delta - m_c - f - I$$

where the production cost increases from m_d to m_c and the punishment of audit does not appear where $\zeta = 0$. If the dirty type does not invest in innovation and greenwashes, the payoff becomes:

$$I_1 := \pi_d^G(NI, L) = V + \mu\Delta - m_d - f - \phi \cdot \zeta \cdot (V + \mu\Delta)$$

The dirty type prefers genuine innovation over greenwashing when $\pi_d(V, L) > \pi_d^G(NV, L)$, which simplifies to:

$$I < \phi\zeta(V + \mu\Delta) - (m_c - m_d)$$

We define the innovation threshold:

$$\zeta^{***} := \frac{I + (m_c - m_d)}{\phi(V + \mu\Delta)}$$

Whether the dirty type can be encouraged to invest in technology depends on the level of ζ^{***}. Recall the incentive condition in the case 1 where the dirty type has an incentive to adopt the label, with the threshold $\zeta^* = \frac{(\mu - b')\Delta - f}{\phi(V + \mu\Delta)}$. When $\zeta^* > \zeta^{***}$, the equilibrium sustains and the dirty type can be motivated to invest in technology when $\zeta \in (\zeta^{***}, \zeta^*)$. Otherwise, when $\zeta^* < \zeta^{***}$, the condition that $\zeta > \zeta^{***}$ breaks down the equilibrium.

In the first case, when $\zeta^* > \zeta^{***}$ and when the threshold I_1 is met, the pooling equilibrium is still sustained. It is worth noting that $\zeta^* > \zeta^{***}$ implies that

$$I + (m_c - m_d) + f < (\mu - b')\Delta$$

The left-hand-side denotes the investment cost I, labeling cost f, and the increase of marginal production cost from updating technology. The right-hand side represents the gree premium from consumers. When the total cost from investment and labeling is lower than the green premium, the dirty type has positive incentive to update technology. This contributes that when $\zeta \in (\zeta^{***}, \zeta^*)$, both types adopt the ecolabel, and the dirty type can be encouraged to invest in innovation rather than greenwash.

In the second case, if

$$I + (m_c - m_d) + f > (\mu - b')\Delta$$

where the total cost of labeling and investing is higher than the green premium, $\zeta^* < \zeta^{***}$. Therefore, under the pooling equilibrium, the dirty type does not have an incentive to update technology.

These two conclusions provide policy implication on settings of labeling cost f. Given consumers' attitude Δ, and firm's cost I, m_c, m_d, the relatively low f set by the third-party agency or the government, can effectively encourage the dirty type transfers to produce genuine clean products and does not greenwash. Define the level

$$f^* = (\mu - b')\Delta - I - (m_c - m_d)$$

Below the f^*, the dirty type is motivated to invest, while above the f^*, the dirty type does not.

5.2 Incentives when only the clean type adopts the ecolabel.

In the separating equilibrium characterized in Case 3, the dirty type does not adopt the ecolabel. Since consumers only pay the green premium Δ to labeled firms, investing in clean technology yields no revenue benefit for the dirty type. Formally:

$$\pi_d(I, NV) = V - m_c - I < V - m_d = \pi_d(NV)$$

since $m_c > m_d$ and $I > 0$. The dirty type therefore has no incentive to invest in green innovation when excluded from the ecolabel market. This because investment without labeling yields no revenue benefit, because consumers only pay the green premium to labeled firms. So ζ strong enough to create separation also cuts off the dirty type's return to investment entirely. The penalty 'works' on the signaling dimension but fails on the environmental dimension.

Proposition 3 *Under the greenwashing audit regime:*

1. *In the separating equilibrium ($\zeta > \zeta^{**}$), the dirty type never invests in clean technology.*

2. *In the pooling equilibrium ($\zeta < \zeta^{*}$), the dirty type invests if and only if*

$f < f^*$ and $\zeta > \zeta^{***}$, where

$$\zeta^{***} := \frac{I + (m_c - m_d)}{\phi(V + \mu\Delta)}, \qquad f^* := (\mu - b')\Delta - I - (m_c - m_d).$$

Otherwise the dirty type greenwashes even under the penalty.

3. *No single penalty rate ζ simultaneously achieves label credibility and green innovation.*

The first conclusion shows that strong penalties eliminate greenwashing but cut off the dirty type's return to investment: without the label, investment yields no green premium regardless of technology. The second conclusion shows that innovation is only possible in the pooling equilibrium, and only when the certification fee is low enough ($f < f^*$) that the total cost of labeling and upgrading is covered by the green premium. The last one is the central policy implication: the penalty instrument alone cannot achieve both goals simultaneously, which motivates the two-instrument analysis in Section 6.

- Proof. See in Appendix.

6 Policy Analysis

This section translates the equilibrium results into policy guidance. We first identify the regulator's objectives and the trade-off between them, then characterize how the thresholds respond to key parameters, and finally evaluate the effectiveness of the EU's current 4% penalty through numerical simulation.

6.1 The Regulator's Problem

The regulator controls two instruments: the penalty rate ζ and the certification fee f. We assume the regulator pursues two goals. Firstly, the government aims to sustain the separating equilibrium so that the ecolabel is an effective signal that each type can reveal themselves correctly, i.e. $\zeta > \zeta^{**}$. Secondly,

the regulator aims to induce the dirty type to invest in clean technology, i.e. $\zeta^{***} < \zeta < \zeta^{*}$ and $f < f^{*}$.

A central finding of the model is that these two goals cannot be achieved simultaneously with the penalty instrument alone. Recall from Propositions 2 and 3 that the first goal requires $\zeta > \zeta^{**}$, which sustains the separating equilibrium. However, in the separating equilibrium the dirty type does not adopt the ecolabel and therefore receives no green premium from investing, making innovation strictly unprofitable regardless of ζ. The second goal, by contrast, requires the pooling equilibrium to be active ($\zeta < \zeta^{*}$), which is incompatible with the first goal. The regulator therefore must choose which goal to prioritise or rely on an additional instrument.

6.2 Numerical Simulation

The analytical results characterise the conditions for each equilibrium but do not quantify which parameter configurations are empirically relevant. We now use numerical simulation to illustrate the structure of the equilibrium regions across the (Δ, f) parameter space. We first map the minimum penalty rate ζ^{**} required to achieve separation for all feasible (Δ, f) combinations, showing how the effective region expands as ζ increases. We then evaluate the EU Green Claims Directive's 4% penalty as a specific case, identifying the parameter configurations under which it achieves label credibility (Goal 1) and green innovation (Goal 2) respectively. Throughout we fix $V = 1$, $\mu = 0.5$, $\phi = 0.5$, and $b' = 0.3 (b' < \mu)$ as benchmark values.

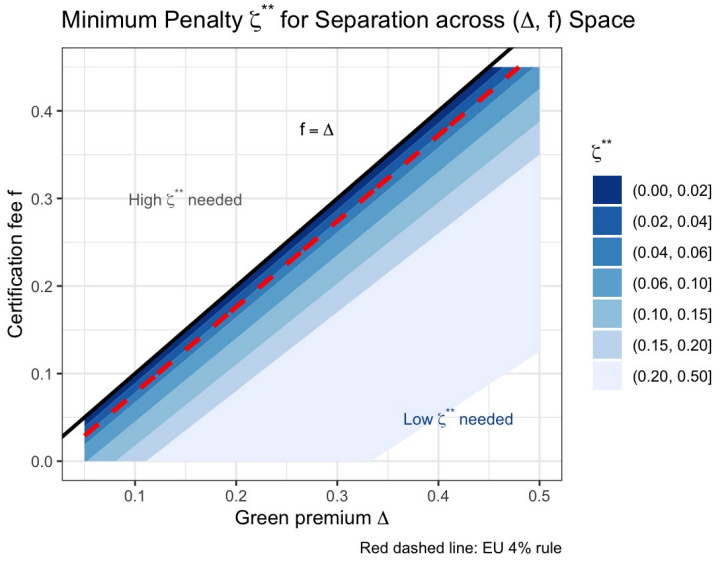


Figure 2: EU 4% sufficiency across the feasible (Δ, f) set.

Figure 2 illustrates how the minimum penalty threshold $\zeta = (\Delta - f)/[\phi(V + \Delta)]$ varies across the feasible parameter space. The feasible region is bounded above by the diagonal $\Delta = f$(solid black line): certification is only credible when the green premium strictly exceeds the certification fee. Within this region, the colour gradient indicates the value of ζ required to sustain a separating equilibrium. Specificlly, darker shades (near the diagonal) correspond to low ζ, while lighter shades (bottom right) correspond to high ζ. Intuitively, when f is close to Δ, the dirty type's net gain from mimicry is small, so even a mild penalty suffices to deter greenwashing. Conversely, when the green premium Δ is large and the certification fee f is low, the mimicry incentive is strong and a higher penalty rate is required.

The red dashed line marks the EU Green Claims Directive threshold $\zeta = 0.04$. Any parameter combination (Δ, f) lying above the red dashed line (i.e. in the dark blue band) satisfies $\zeta^{**} < 0.04$, meaning the EU rule is *sufficient* to induce separation. Parameter combinations below the red dashed line require a penalty rate exceeding 4% of revenue, suggesting the current EU policy may be

inadequate in markets where green premiums are large but certification costs are low.

6.3 Welfare Implications.

Since the monopolist extracts full consumer surplus, social welfare equals firm profit plus the environmental benefit from clean production. Let $e > 0$ denote the environmental damage from dirty production not captured in the green premium Δ. Expected social welfare in each regime is:

$$\begin{aligned} W_{\text{pool}} &= \mu\pi_c + (1-\mu)\pi_d^{GW} - (1-\mu)e, \\ W_{\text{pool+innov}} &= \mu\pi_c + (1-\mu)\pi_d^{I}, \\ W_{\text{sep}} &= \mu\pi_c^{sep} + (1-\mu)\pi_d^{NL} - (1-\mu)e. \end{aligned}$$

Under pooling without innovation, the dirty firm greenwashes and continues to cause environmental damage e. Under pooling with innovation, the dirty firm upgrades to clean technology, eliminating the damage entirely. Under separation, the dirty firm is revealed and forgoes the ecolabel, but still produces dirty goods, so damage e persists. The welfare comparison between separation and innovation therefore depends on whether the environmental gain e outweighs the innovation cost $I + (m_c - m_d)$.

We can conclude the above subsections 6.1-6.3 by the following proposition.

Proposition 4 *The policy implication from the analysis can be concluded by three main facts.*

- $W_{pool+innov} > W_{pool}$ *and* $W_{sep} > W_{pool}$ *hold unambiguously. The ranking between the two depends on environmental damage:* $W_{pool+innov} > W_{sep}$ *if and only if* $e > I + (m_c - m_d)$*. Consequently, when environmental damage is sufficiently severe, a regulator should prefer the innovation regime over separation, even though the ecolabel is uninformative in the former.*

- ζ^{**} *is strictly decreasing in* f*:* $\partial\zeta^{**}/\partial f < 0$*. Hence the certification fee*

and the penalty rate are policy substitutes *for achieving separation. So a regulator with limited enforcement capacity can compensate by raising f, and vice versa.*

- *The EU 4% rule achieves separation if and only if*

$$f \;>\; f_{EU} \;\equiv\; \Delta - 0.04\,\phi(V+\Delta).$$

When $f \leq f_{EU}$, the current penalty rate is insufficient and a stricter regime is required.

The first conclusion is derived by the above welfare analysis. The second conclusion can be derived from the relationship between ζ^{**} and f. We can differentiate ζ^{**} and ζ^{***} with respect to the main parameters. The results are summarised in Table 1.

We can find two complementarities from the table. First, f and ζ are *substitutes* for achieving separation: a regulator with limited enforcement capacity can compensate by raising the certification fee, and vice versa. Second, ϕ and ζ jointly determine the expected penalty $\phi\zeta p$; increasing either one lowers the threshold for both goals, making audit frequency a powerful complement to the penalty rate.

Table 1: Comparative statics of the policy thresholds. ↑ denotes the threshold increases; ↓ denotes it decreases; — denotes no effect.

Parameter	ζ^{**} (Case 1)	ζ^{***} (Case 3)	Intuition
Green premium $\Delta \uparrow$	↑	↓	Higher Δ makes greenwashing more attractive, requiring a stronger penalty for separation; it also raises the revenue benefit of genuine innovation.
Certification fee $f \uparrow$	↓	—	A higher fee already deters mimicry, so a lower penalty rate suffices for separation.
Audit probability $\phi \uparrow$	↓	↓	More frequent audits amplify the effective expected penalty, easing both goals simultaneously.
Investment cost $I \uparrow$	—	↑	Costlier innovation requires a higher penalty to make investment worthwhile relative to greenwashing.
Cost gap $(m_c - m_d) \uparrow$	—	↑	A larger production cost increase from upgrading technology discourages investment.
Prior belief $\mu \uparrow$	—	↓	A higher prior raises the pooling green premium $\mu\Delta$, increasing the opportunity cost of greenwashing.

The last conclusion in the proposition 4 is concluded from simulations in subsection 6.2. Taken together, the results suggest two concrete recommendations for regulators. First, the EU Green Claims Directive's 4% revenue penalty is insufficient for separation in most empirically relevant configurations of (Δ, f)

(Figure 2); stricter enforcement or higher certification fees are needed. Second, when environmental damage is severe, specifically when $e > I + (m_c - m_d)$, regulators should prioritise inducing green innovation over achieving ecolabel separation, even at the cost of an uninformative label.

7 Conclusions

This paper develops a signaling game to analyse when revenue-based penalties can discipline greenwashing. A monopolist privately knows its environmental type and chooses whether to seek an ecolabel, pay the certification fee, and invest in clean technology. A regulator audits firms and imposes a penalty proportional to revenue on those found to misrepresent their type. We characterise the full set of Perfect Bayesian Equilibria satisfying the Intuitive Criterion.

Three equilibrium regimes emerge. When the penalty rate is sufficiently high, it deters mimicry and a separating equilibrium prevails: only the genuinely green firm certifies, and consumers can trust the label. When the penalty is too low, deterrence fails and both types pool under the ecolabel, rendering it uninformative. Between these two cases, no pure-strategy equilibrium exists. Within the pooling regime, a third outcome arises: a moderate penalty that falls short of inducing separation may still be strong enough to make clean investment worthwhile for the dirty firm, turning greenwashing into genuine innovation. This identifies a previously overlooked channel through which moderate penalties can improve environmental outcomes even when they fail to achieve label credibility.

The welfare analysis qualifies the conventional preference for separation. While both separation and innovation dominate pure pooling, their relative ranking depends on the severity of environmental damage. When damage is large relative to the cost of innovation, inducing green investment delivers higher social welfare than achieving label separation. This suggests that regulators facing severe environmental damage should target green innovation rather than ecolabel credibility. On the instrument side, the certification fee and the penalty

rate act as policy substitutes, giving regulators a two-dimensional lever. Calibrating to the EU Green Claims Directive, we find that the current four percent revenue penalty achieves separation only in a narrow region of the parameter space where certification fees are already high, suggesting the rule is insufficient as a standalone instrument.

Several extensions remain for future work. The model assumes monopoly; with competing firms, strategic certification decisions may generate additional equilibria and alter the welfare ranking. The penalty rate is treated as exogenous; endogenising the regulator's problem under a budget constraint would enrich the policy analysis. Finally, a dynamic extension could examine how repeated auditing and reputation effects reshape firms' long-run incentives to invest in green technology.

Appendix

A.1 Proof of Proposition 1

We verify each case in turn. Throughout, we work without the greenwashing penalty ($\zeta = 0$).

Case 1: Pooling equilibrium with ecolabel.

Suppose both types adopt the ecolabel. By Bayes' rule, the on-path posterior is $b = \mu$, giving price $p^L = V + \mu\Delta$. The off-path action is NL with belief $b' \in [0, 1]$. The clean type prefers L over NL if and only if

$$V + \mu\Delta - m_c - f \;>\; V + b'\Delta - m_c \quad\Longleftrightarrow\quad f \;<\; (\mu - b')\Delta.$$

The dirty type faces the identical condition:

$$V + \mu\Delta - m_d - f \;>\; V + b'\Delta - m_d \quad\Longleftrightarrow\quad f \;<\; (\mu - b')\Delta.$$

Hence both types prefer to label whenever $f < (\mu - b')\Delta$, which can be satisfied for any $b' \in [0, \mu - f/\Delta)$.

The IC asks whether either type is equilibrium dominated in the deviation to NL: a type is equilibrium dominated if deviation is unprofitable even under the most favourable off-path belief $b' = 1$, i.e. if $\max_{b' \in [0,1]} [\pi_i(NL) - \pi_i(L)] < 0$. Since the deviation payoff $\pi_i(NL) = V + b'\Delta - m_i$ is increasing in b', the maximum is attained at $b' = 1$. Evaluating for each type:

$$\pi_i(NL)\big|_{b'=1} - \pi_i(L) \;=\; (V + \Delta - m_i) - (V + \mu\Delta - m_i - f) \;=\; (1-\mu)\Delta + f \;>\; 0.$$

The deviation is profitable for *both* types under $b' = 1$. Therefore neither type is equilibrium dominated, the IC cannot assign $b' = 0$ or $b' = 1$ to either type specifically, and b' remains free. The pooling equilibrium survives the IC for any admissible $b' \in [0, \mu - f/\Delta)$.

Case 2: Pooling equilibrium without ecolabel.

Suppose neither type labels. By Bayes' rule, $b' = \mu$ on the equilibrium path, giving $p^{NL} = V + \mu\Delta$. The off-path action is L with belief $b \in [0, 1]$. Each type prefers NL over L if and only if

$$V + \mu\Delta - m_i \;\geq\; V + b\Delta - m_i - f \quad \Longleftrightarrow \quad f \;\geq\; (b - \mu)\Delta.$$

This holds for any $b \leq \mu + f/\Delta$, so the equilibrium is supported whenever $f > (b - \mu)\Delta$.

Intuitive Criterion. The off-path deviation is to L with belief $b \in [0, 1]$. The deviation payoff is increasing in b, so we evaluate at $b = 1$:

$$\pi_i(L)\big|_{b=1} - \pi_i(NL) \;=\; (V + \Delta - m_i - f) - (V + \mu\Delta - m_i) \;=\; (1 - \mu)\Delta - f.$$

If $f \geq (1 - \mu)\Delta$, neither type gains from deviation even at $b = 1$; both are equilibrium dominated and IC places no restriction on b. The equilibrium survives. If $f < (1 - \mu)\Delta$, both types benefit from deviation simultaneously. Since neither type can be ruled out as the deviator, IC again cannot assign b to one type specifically, and the equilibrium survives with free $b \in [0,\, \mu + f/\Delta)$.

Case 3: Only the clean type labels. By Bayes' rule, $b = 1$ and $b' = 0$. The clean type prefers L when $f < \Delta$; the dirty type prefers NL when $f > \Delta$. These two conditions require $f < \Delta$ and $f > \Delta$ simultaneously, which is a contradiction. No such equilibrium exists.

Case 4: Only the dirty type labels. By Bayes' rule, $b = 0$ and $b' = 1$. The dirty type's payoff from L is $V - m_d - f$, while its payoff from deviating to NL is $V + \Delta - m_d$. Since $\pi_d(NL) - \pi_d(L) = \Delta + f > 0$ always, the dirty type strictly prefers to deviate. This equilibrium cannot be sustained.

Along any equilibrium path, the dirty type compares investing (L, V) with greenwashing (L, NV):

$$\pi_d(L, V) = V + b\Delta - m_c - f - I \;<\; V + b\Delta - m_d - f = \pi_d(L, NV),$$

since $m_c > m_d$ and $I > 0$. Investment is strictly dominated by greenwashing regardless of beliefs or the certification fee.

A.2 Proof of Proposition 2

We verify each case in turn. Throughout, we work with the greenwashing audit and penalty (ζ).

Case 1: Pooling equilibrium with ecolabel.

Suppose both types adopt the ecolabel. By Bayes' rule the on-path posterior is $b = \mu$, so $p^L = V + \mu\Delta$. Let $b' \in [0, 1]$ be the off-path belief upon observing NL.

The clean type prefers L over NL if and only if

$$\pi_c(L) > \pi_c(NL) \iff V + \mu\Delta - m_c - f \;>\; V + b'\Delta - m_c \iff f < (\mu - b')\Delta. \tag{C_1}$$

The dirty type (greenwashing) prefers L over NL if and only if

$$\pi_d(L) > \pi_d(NL) \iff V+\mu\Delta-m_d-f-\phi\zeta(V+\mu\Delta) \;>\; V+b'\Delta-m_d \iff \zeta < \frac{(\mu - b')\Delta - f}{\phi(V + \mu\Delta)} =: \zeta^*. \tag{D_1}$$

Since $\phi\zeta(V + \mu\Delta) > 0$, condition D_1 is strictly tighter than C_1: whenever D_1 holds, C_1 holds automatically. The binding incentive constraint is therefore D_1, and a PBE exists for any $b' \in [0,\, \mu - f/\Delta)$ satisfying D_1.

The off-path action is NL. Since $\pi_i(NL) = V + b'\Delta - m_i$ is strictly increasing in b', the maximum deviation gain for each type is attained at $b' = 1$. Evaluating:

$$\pi_c^{off}(NL)\big|_{b'=1} - \pi_c(L) = (V + \Delta - m_c) - (V + \mu\Delta - m_c - f) = (1-\mu)\Delta + f > 0,$$

$$\begin{aligned}\pi_d^{off}(NL)\big|_{b'=1} - \pi_d(L) &= (V + \Delta - m_d) - (V + \mu\Delta - m_d - f - \phi\zeta(V + \mu\Delta)) \\ &= (1-\mu)\Delta + f + \phi\zeta(V + \mu\Delta) > 0.\end{aligned}$$

Both types strictly gain from deviation under $b' = 1$, so neither type is equilibrium dominated. The IC cannot assign $b' = 0$ or $b' = 1$ to either type specifically, and b' remains free. The pooling equilibrium therefore survives the IC refinement for any admissible $b' \in [0, 1)$ satisfying D_1.

- **Case 2. Pooling equilibrium without ecolabel.**

Recall that the clean type adopts the ecolabel when $\pi_c(NL) > \pi_c(L))$, that is when the condition $C_2 : f > (\mu - b')\Delta$ is met. For the dirty type, if it adopts the ecolabel, the payoff becomes:

$$\pi_d(L) = V + b\Delta - m_d - f - \phi \cdot \zeta \cdot (V + b\Delta)$$

where $b \in [0, 1]$ denotes the posterior belief on off-path. If it does not adopt the ecolabel, it has $\pi_d(NL) = V + b'\Delta - m_d$. Therefore, the dirty type does not adopt the ecolabel when $\pi_d(NL) > \pi_d(L)$, that is,

$$D_2 := \zeta > \frac{(b-\mu)\Delta - f}{\phi(V + b\Delta)}$$

Therefore, D_2 combined with C_2 consist the incentive compatibility of the equilibrium.

Next, the IC refinement can examine the stability. For the clean type, it has an incentive to deviate when $f < (1-\mu)\Delta$, holding the best-case belief that $b = 1$, according to analysis in benchmark.

For the dirt type, without the ecolabel the dirty type's payoff is

$$\pi_d(NL) = V + \mu\Delta - m_d, (b' = \mu)$$

The deviation payoff for the dirty type can be denoted by

$$\pi_d^L(off - path) = V + b\Delta - m_d - f - \phi \cdot \zeta \cdot (V + b\Delta)$$

Thus,

$$\pi_d^L(off - path) - \pi_d(NL) = (b - \mu)\Delta - f - \phi \cdot \zeta(V + b\Delta)$$

Holding the best-case belief that $b = 1$, we can find that the dirty type does not want to deviate since $\pi_d^L(off - path) < \pi_d(NL)$, specifically, when $\zeta > \frac{(1-\mu)\Delta - f}{\phi(V+\Delta)}$. Binding the condition where the clean type wants to deviate, $f < (1-\mu)\Delta$, we can conclude that the IC refinement will assign $b = 1$ to the clean type that can benefit from the deviation. Under this belief assignment, the clean type strictly prefers to deviate, which violates the equilibrium requirement that both types remain on the no-label path. Therefore, the pooling equilibrium without ecolabel **fails** the Intuitive Criterion and is eliminated whenever $f < (1-\mu)\Delta$ and $\zeta > \frac{(1-\mu)\Delta - f}{\phi(V+\Delta)}$.

A.3 Proof of Proposition 3

- **Case 1. Separating equilibrium.**

In the separating equilibrium the dirty type does not adopt the ecolabel. If it deviates to invest in clean technology while remaining non-labelled, its payoff is

$$\pi_d(NL, V) = V - m_c - I.$$

Compared with not investing and not labeling,

$$\pi_d(NL, V) - \pi_d(NL, NV) = (V - m_c - I) - (V - m_d) = -(m_c - m_d) - I \ < \ 0,$$

since $m_c > m_d$ and $I > 0$. Investment is strictly dominated regardless of ζ: without the label, upgrading technology raises costs but generates no additional revenue.

- **Case 2. Pooling equilibrium.**

In the pooling equilibrium the on-path belief is $b = \mu$, giving $p^L = V + \mu\Delta$. The dirty type compares two strategies within this equilibrium: invest and label at true cost (L, V), or greenwash and label (L, NV).
If it invests, it is genuinely clean and faces no penalty:

$$\pi_d(L, V) = V + \mu\Delta - m_c - f - I.$$

If it greenwashes, it faces the expected audit penalty:

$$\pi_d(L, NV) = V + \mu\Delta - m_d - f - \phi\zeta(V + \mu\Delta).$$

The dirty type strictly prefers investment over greenwashing if and only if $\pi_d(L, V) > \pi_d(L, NV)$, which simplifies to

$$\phi\zeta(V + \mu\Delta) > I + (m_c - m_d) \iff \zeta > \zeta^{***} := \frac{I + (m_c - m_d)}{\phi(V + \mu\Delta)}.$$

For the dirty type to also prefer labelling with investment over not labelling at all, we need $\pi_d(L, V) \geq \pi_d(NL)$:

$$V + \mu\Delta - m_c - f - I \ \geq \ V + b'\Delta - m_d \iff f \ \leq \ (\mu - b')\Delta - I - (m_c - m_d) =: f^*.$$

Therefore the dirty type invests if and only if both $\zeta > \zeta^{***}$ and $f < f^*$ hold. If either condition fails, greenwashing strictly dominates investment within the

pooling equilibrium.

Suppose for contradiction that some ζ achieves both Goal (i) (label credibility, $\zeta > \zeta^{**}$) and Goal (ii) (green innovation, $\zeta \in (\zeta^{***}, \zeta^{*})$ and $f < f^{*}$) simultaneously.

Goal (i) requires $\zeta > \zeta^{**}$, which places the equilibrium in the separating region. But Case 1 shows that in the separating equilibrium the dirty type never invests, so Goal (ii) fails.

Goal (ii) requires $\zeta < \zeta^{*}$, which keeps the equilibrium in the pooling region. But when $\zeta < \zeta^{*}$, both types label and the ecolabel is uninformative, so Goal (i) fails.

Since $\zeta^{*} < \zeta^{**}$ whenever $\mu < 1$, the two regions (ζ^{***}, ζ^{*}) and $(\zeta^{**}, 1)$ are disjoint, and no single ζ lies in both. This is a contradiction.

A.4 Proof of Proposition 4

For the first conclusion, comparing $W_{\text{pool+innov}}$ and W_{pool}, the difference is $(1-\mu)[\pi_d^I - \pi_d^{GW} + e] > 0$, since the dirty type innovates only when $\pi_d^I \geq \pi_d^{GW}$ and $e > 0$. Comparing $W_{\text{pool+innov}}$ and W_{sep}, the dirty firm eliminates damage e under innovation but incurs cost $I + (m_c - m_d)$; the net welfare gain is positive if and only if $e > I + (m_c - m_d)$.

For the second conclusion, differentiating $\zeta^{**} = (\Delta - f)/[\phi(V + \Delta)]$ with respect to f gives $\partial\zeta^{**}/\partial f = -1/[\phi(V + \Delta)] < 0$.

At last, setting $\zeta^{**} < 0.04$ and solving for f yields $f > \Delta - 0.04\,\phi(V + \Delta) \equiv f_{EU}$ directly.

References

Amacher, G.S., Koskela, E., Ollikainen, M., 2004. Environmental quality competition and eco-labeling. Journal of Environmental Economics and Management 47, 284–306.

Arguedas, C., Blanco, E., 2025. On fraud and certification of green production. Journal of Economics & Management Strategy 34, 501–524.

Cho, I.K., Kreps, D.M., 1987. Signaling games and stable equilibria. The Quarterly Journal of Economics 102, 179–221.

Fudenberg, D., Tirole, J., 1991. Game theory. MIT press.

Garrido, D., Espínola-Arredondo, A., Munoz-Garcia, F., 2020. Can mandatory certification promote greenwashing? a signaling approach. Journal of Public Economic Theory 22, 1801–1851.

Hamilton, S.F., Zilberman, D., 2006. Green markets, eco-certification, and equilibrium fraud. Journal of environmental economics and management 52, 627–644.

Heyes, A., Kapur, S., Kennedy, P.W., Martin, S., Maxwell, J.W., 2020. But what does it mean? competition between products carrying alternative green labels when consumers are active acquirers of information. Journal of the Association of Environmental and Resource Economists 7, 243–277.

Ibanez, L., Grolleau, G., 2008. Can ecolabeling schemes preserve the environment? Environmental and resource economics 40, 233–249.

Nygaard, A., 2023. Is sustainable certification's ability to combat greenwashing trustworthy? Frontiers in Sustainability 4, 1188069.

Podhorsky, A., 2010. Environmental labeling. Available at the author's website: http://dept. econ. yorku. ca/˜ andrea/research. html .

Sengupta, A., 2012. Investment in cleaner technology and signaling distortions in a market with green consumers. Journal of Environmental Economics and Management 64, 468–480.

Sengupta, A., 2015. Competitive investment in clean technology and uninformed green consumers. Journal of Environmental Economics and Management 71, 125–141.

Sengupta, A., 2024. Communicating clean technology: Green premium, competition, and ecolabels. Journal of Economics & Management Strategy 33, 605–629.

Spence, M., 1978. Job market signaling, in: Uncertainty in economics. Elsevier, pp. 281–306.

Spence, M., 2002. Signaling in retrospect and the informational structure of markets. American economic review 92, 434–459.

Wu, Y., Zhang, K., Xie, J., 2020. Bad greenwashing, good greenwashing: Corporate social responsibility and information transparency. Management Science 66, 3095–3112.