Algorithmic Explainability and Obfuscation under Regulatory Audits

Xavier Lambin*, Adrien Raizonville†

March 21, 2022

Abstract

The best-performing and most popular algorithms are often the least explainable. In parallel, there is growing concern and evidence that sophisticated algorithms may engage, autonomously, in profit-maximizing but welfare-damaging strategies. Drawing on the literature on self-regulation and following recent regulatory proposals, we model a regulator who seeks to encourage algorithmic compliance through the threat of (costly and imperfect) audits. Firms may invest in "explainability" to better understand their own algorithms and reduce their cost of compliance. We find that, when audit efficacy is not affected by explainability, audit regulation always induces investment in explainability. Mandatory disclosure of the explainability level makes regulation even more effective, because it allows firms to signal compliance. If, instead, explainability facilitates regulatory audits a firm may attempt to hide a potential misconduct behind algorithmic opacity. Because of regulatory opportunism, mandatory disclosure may further deter investment in explainability. In these cases, regulatory audits may be counterproductive and laissez-faire or minimum explainability standards should be envisaged.

Keywords: Explainability, Algorithmic decision-making, Self-regulation, Audits, Output regulation.

JEL codes: D21, D83, K13, K24, K42

*ESSEC Business School and THEMA, 3 Av. Bernard Hirsch, B.P. – 50105, Cergy, 95021, France. xavier.lambin@essec.edu

†Telecom Paris, i3, Institut Polytechnique de Paris, adrien.raizonville@telecom-paris.fr

"Algorithms must not be a black box and there must be clear rules if something goes wrong."
Ursula von der Leyen, State of the Union, September 2020.

1 Introduction

An increasing number of decisions involving important business activities such as price-setting, advertising, loan granting processes, are delegated to artificial intelligence algorithms.[1] Over the past decade, these technologies have made staggering progress that may bring sizeable benefits to firms, consumers, and society as a whole. There is, however, mounting concern that these algorithms may take undesirable or even illegal decisions on behalf of their makers–sometimes without their knowing. With suspicion regarding the secrecy and lack of explainability of advanced machine-learning tools rising, firms making extensive use of algorithms have come under increased scrutiny, and many institutions are laying the foundations for algorithmic regulation and auditing (see e.g., European Commission (2020a,b, 2021)).

Explainability is generally defined as the extent to which an algorithm can be explained in human terms. When choosing their technology, firms face a trade-off between performance and explainability; often, the best-performing methods (e.g., deep neural networks) are the least explainable, while the most explainable (e.g., linear regressions or simple decision trees) are the least accurate. This phenomenon, illustrated in Figure 1, is well-known to data scientists[2] but generates a somewhat new concern among economists and regulators: as algorithmic sophistication increases algorithmic misconducts become increasingly difficult to detect and demonstrate. When the lack of explainability is detrimental to social welfare, regulatory intervention may be warranted.

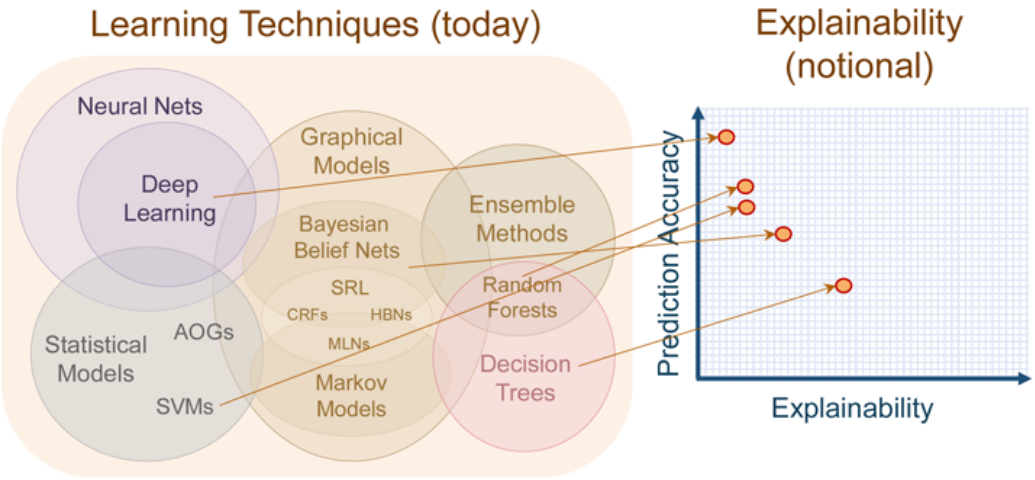
This paper models the interaction between a firm and a regulator. The firm operates a technology that may autonomously engage in "misconduct" (i.e., harm social welfare). The regulatory environment penalizes misconduct through regulatory audits, inducing the firm to intensify its compliance efforts (e.g., by implementing self-audit procedures). Explainability, which comes at the firm's cost,[3] helps the firm understand the behavior of its own technology, and therefore identify

[1] A possible definition of artificial intelligence may be found in Acemoglu and Restrepo (2020): "[Artificial Intelligence] refers to the study and development of 'intelligent (machine) agents', which are machines, software, or algorithms that act intelligently by recognising and responding to their environment."

[2] While recent research has made advances in developing interpretable machine-learning models, Barredo Arrieta et al. (2020) and Bertsimas et al. (2019) note that algorithmic interpretability comes at the cost of accuracy.

[3] In the absence of an appropriate regulatory framework (or effective consumer reactions), the firm always favors

Figure 1: The explainability vs accuracy trade-off



Source: Gunning and Aha (2019)

and remedy a possible misconduct prior to regulatory intervention. More specifically, in our model, explainability lowers the marginal cost of compliance.

We compare a firm's equilibrium explainability in two distinct regulatory regimes: when the explainability level is firm's private information technology-neutral regulation arises naturally. In that case, the regulator maintains the same audit frequency for all firms, irrespective of the observed level of explainability. In a second paradigm we acknowledge that the regulator may request that the explainability level be observed publicly (e.g. through a regulatory requirement for transparency) which makes technology-specific regulation possible: the regulator observes the explainability level before auditing begins and may adjust the audit frequency accordingly. In this case, we will show that explainability acts as a signalling and commitment device of strong compliance.

Out first set of results concerns the effect of imperfectly efficacious regulatory audits on equilibrium explainability. When the efficacy of regulatory audits is not affected by explainability, regulatory audits always prompt firms to voluntarily invest in explainability. This is in line with intuition, as firms invest in compliance following the introduction of audits, and consistently strive to reduce their compliance costs through increased explainability. However, we note that explainability may, in some circumstances, render regulatory audits more efficacious, which deters firms

the efficiency of a technology over its explainability. McKinsey & Company survey of 2,360 company respondents, each answering questions about their organizations. While 39% of respondents recognize the risk associated with "explainability", only 21% say they are actively addressing this risk (Cam et al., 2019). Examples of technology companies that operate "responsible artificial intelligence" divisions include Facebook and Google.

from investing in explainability. We find that when explainability strongly affects audit efficacy our first result may be overturned: audits may not induce investment in explainability at all and firms may even actively obfuscate their algorithms so as to hide behind less transparent processes.

Our second set of results concerns the effect of mandatory disclosure of the explainability level. Two additional effects may come into play. First, the regulator rationally anticipates that a firm which invests robustly in explainability is more likely to be compliant. As a consequence, the regulator rationally lowers the frequency of auditing to explainable technologies. This mechanism strengthens the firm's incentives to invest in explainability, so as to signal compliance and induce less audit. When explainability does not make audits more efficacious this is the only effect of mandatory disclosure, which unequivocally *promotes* investment in explainability. However, when explainability makes audits more efficacious a second effect emerges. Indeed, a regulator may increase the audit frequency to explainable technologies, where its audits are more likely to be successful. This behavior, which we call "regulatory opportunism", weakens firms' incentives to invest in explainability. When this effect is too strong, our conclusion is reversed and mandatory disclosure of the explainability level *reduces* equilibrium investment in explainability.

To the best of our knowledge, the present paper is the first to model technology explainability as a firm's strategic choice. Also novel is the explicit acknowledgement that explainability affects the firm's ability to comply with regulation as well as the efficacy of regulatory interventions. Our analysis has important policy implications for the regulation of algorithms. First, we observe that firms may reduce explainability strategically when regulatory audits are imperfect, as doing so enables them to evade regulatory monitoring and punishment. Rather counter-intuitively, in this situation audit regulation generates adverse results and laissez-faire should be preferred. Second, we show that in some situations, transparency over the level of explainability should be avoided. This is, again, rather counter-intuitive inasmuch as transparency usually allows efficient regulation and investments. Our model provides guidance for the implementation of the regulatory audits envisaged in recent regulatory proposals for digital markets, digital services, and AI in Europe (see European Commission 2020a,b, 2021, respectively).

The remainder of the paper is organized as follows. In Section 2, we review the literature. In Section 3, we describe the base model. In Section 4, we analyse investments in explainability in the benchmark case when explainability is firm's private information (technology-neutral regula-

tion). These results are compared to the case of mandatory disclosure of the explainability level (technology-specific regulation) in Section 5. In Section 6, we analyse social welfare and confirm that in most relevant cases, a benevolent welfare-maximizing social planner would seek to promote explainability. We conclude in Section 7.

2 Literature review

This paper relates to three streams of literature. The first stream concerns the analysis of the use of algorithms in decision-making. A recent but very active literature shows that competing AI algorithms may engage in welfare-reducing strategies (Calvano et al., 2020; Assad et al., 2020; Brown and Mackay, 2019; Klein, 2019; Abada and Lambin, 2020). Athey et al. (2020) study decisions to delegate decision-making to either a human agent or an algorithm. Similarly, Dogan et al. (2018) studies adoption and utilization of automation in firms with varying organizational structures by developing a theoretical model of organizational design with embedded cheap-talk. In contrast with this literature, we explicitly model the interactions between the firm and a regulator under moral-hazard. The firm is already committed to use an algorithmic solution, and it must now choose the level of explainability of its algorithm. This choice not only affects its own efficiency, but also the efficacy of regulatory instruments.

Second, our paper relates to the economics literature on optimal law enforcement. How a firm responds to changes in enforcement policies has been discussed extensively in the literature. Becker (1968) was the first to formalize the inclusion of economic considerations in studies of law enforcement (see, e.g., Polinsky and Shavell, 2000 and Shavell, 2009 for a survey). The literature has emphasized the several issues related to public enforcement. Bebchuk and Kaplow (1993) acknowledge that all offenders are not equally easy to apprehend. In turn, firms may respond to regulatory oversight by undertaking avoidance activities that make detection more difficult. Malik (1990) and Garvie and Keeler (1994) model the implications of attempts by agents trying to reduce the probability that they will be sanctioned by engaging in evasion, lobbying, or concealment efforts. He shows that larger penalties increase incentives to engage in avoidance activities, so an optimizing enforcement agency may not choose the stiffest possible sanction. We obtain a similar result with our model, where the firm would choose to obfuscate its technology when the

regulatory regime is too stringent. Heyes (1994) expands these ideas to include a case where a regulator trades off the frequency of inspections against their thoroughness. More frequent inspections encourage concealment, while more thorough inspections encourage transparency. In addition, regulators have developed valuable tools, such as leniency and voluntary disclosure programs, to manage enforcement costs and information problems.[4] In this body of literature, however, firms invest in a technology which only purpose is to make the discovery and verification of violations difficult and costly. In contrast, we allow for explainability to affect not only the regulators performance but also the firm's performance and propensity to self-police. The firm may also reduce the regulator's enforcement cost by committing to compliance through its investment in explainability, thereby affecting the regulatory pressure it faces. Our paper provides practical guidance for curbing algorithmic-driven misconduct through auditing schemes that (indirectly) promote explainability.

Finally, our paper contributes to the theoretical literature that studies how decisions to self-police are influenced by regulatory enforcement policy. To detect and deter misconduct, a regulator would traditionally audit a firm with some probability and impose a penalty when misconduct is identified (as proposed in recent regulatory projects European Commission 2020a,b). This solution requires time and financial resources. This is particularly the case when adequate expertise is rare and costly, as is the case for algorithmic audits. To limit regulatory costs, a regulator may want to stimulate self-policing by a firm, which makes the need for an audit less urgent. Indeed, the threat of punishment may be sufficient to induce the firm to seek to identify misconduct prior to taking the algorithm to market – and prior to exposing the firm to a regulatory intervention. Contrary to the literature on self-reporting, firms do not use regulatory mechanisms previously developed by public authorities; they self-regulate strategically to preempt future regulatory actions. Glazer and McMillan (1992) show theoretically that a monopolistic firm that faces the threat of regulation lowers its prices to avert regulation. Maxwell et al. (2000) study whether firms can avert environmental regulation by controlling pollution voluntarily. Suijs and Wielhouwer (2019) study coordination issues and free-riding problems when firms seek to avert regulation. Lyon and Maxwell (2016) characterize strategies deployed when firms signal their type through extensive self-regulation or remain in step with the rest of the industry through modest levels of self-regulation.

[4]The theoretical literatures on leniency programs and on self-reporting are vast (seeMarvão and Spagnolo (2018) for a literature review on the former and see, e.g., Kaplow and Shavell (1994), Innes (1999) for the latter.

Our model is related to that in Maxwell and Decker (2006), who investigate how a regulator may induce voluntary environmental investments. A key difference with Maxwell and Decker (2006), is that in our model the firm's investment may facilitate regulatory audits. Much of the literature on industry self-regulation argues that firms can profitably preempt mandatory regulatory requirements (e.g., Short and Toffel, 2010). We show that this may not be the case when explainability facilitates regulatory audits. To the best of our knowledge, our paper is the first to acknowledge that algorithmic explainability affects the efficiency of regulatory monitoring. This brings new insight into the question regarding how private investments that facilitate both private and public monitoring can be promoted.

3 The model

We consider a game with incomplete information between two strategic (risk-neutral) agents: a profit-maximizing firm and a regulator. The firm uses a technology (the algorithm) that generates a fixed revenue but, with some probability, may also generate a net welfare cost $K \geq 0$ to society. In this case, we say that the technology engages in "misconduct". The regulator seeks to minimize the technology's expected damage. He may audit the firm at frequency $m \in [0, 1]$, in which case the regulator finds whether there was a misconduct with a certain probability, defined below. If found guilty of misconduct, a fine $F \geq 0$ is imposed on the firm. To minimize the expected fine, the firm may therefore endeavor to comply, as is reflected in a choice of compliance probability, or the probability that there is no misconduct, $p \in [0, 1]$. To attain compliance level p the firm endures a cost $\Psi(x, p)$, where $x \in [0, 1]$ is the explainability of the technology. The compliance cost is an increasing and convex function of the compliance probability ($\Psi_p > 0$, $\Psi_{pp} > 0$). While the compliance level p is the firm's private information, the explainability level x may be publicly observed. In our model, explainability is the extent to which a misconduct facilitated by an algorithm can be identified by humans (the manager of the firm that operates the algorithm, or a regulator).[5] More specifically,

[5]Though our definition is very consensual, explainability is often defined in much broader terms, that include our definition. "Explainability" is defined in Cam et al. (2019) as the ability to explain how AI models come to their decisions. The concept of explainability is closely related to the concepts of interpretability and transparency. Transparency sometimes involves asymmetric information between a regulator and a firm. In contrast, with explainability, information is usually symmetric but possibly imperfect.

explainability reduces the total as well as the marginal cost of compliance ($\Psi_x < 0$, $\Psi_{xp} < 0$).[6] The technology has a base explainability level x_0, which is the technology's intrinsic explainability and is publicly known. Deviating from this base level generates a positive and convex cost $C(x)$, which may represent engineering of efficiency costs and finds its minimum in x_0 ($C(x_0) = 0$, $C'(x_0) = 0$, $C''(x_0) > 0$). x_0 is the level of explainability that would arise in case of laissez-faire.

Definition 1. *When $x > x_0$, we say that there is investment in explainability. When $x < x_0$, there is obfuscation.*

Crucially, we allow for explainability to also affect the efficacy $\eta(x)$ of regulatory audits ($\eta'(x) \geq 0$), which is the probability that the audit correctly identifies the misconduct.[7] In doing so, we explicitly acknowledge that explainability is a double-edged sword. It helps the firm ensure that its algorithm is compliant but it may also make regulatory audits more efficacious. In sum, the firm chooses the explainability x of its technology and the compliance effort p to minimize its expected cost:

$$\min_{x,p} \mathcal{FC}(x,p) = C(x) + \Psi(x,p) + (1-p)m\eta(x)F \tag{1}$$

This cost function allows for a very flexible representation of the firms private costs and benefits of explainability and compliance. The first two terms represent respectively the direct costs of providing explainability x and compliance p. We note that the first term may represent not only the loss in accuracy of the algorithm following the introduction of explainability (see Figure 1), but also the developer's time needed to alter the base explainability of an algorithm. Importantly, it may also cover any intrinsic private benefit or cost of explainability: opaque algorithms may generate higher or lower profits than transparent algorithms, but also affect public acceptance, as consumers may value transparency in and of itself. The second term reflects the cost of compliance, in the form of the costs of an internal audit. Explainability allows to reduce this cost. Note that this term may also cover any expected private cost (benefit) of compliance per se, in case a misconduct benefits (harms) the firm. Allowing for such a cost (benefit) of compliance would translate in a

[6] Though we 'black-box' the compliance technology, we believe these to be plausible restrictions to place on it, and we restrict our analysis to technologies which satisfy them. These are consistent with the seminal papers on imperfect inspectability such as Bebchuk and Kaplow (1993); Garvie and Keeler (1994); Heyes (1994) among others.

[7] For simplicity we assume that the fine is conditional on the regulator's having proved the infringement : we allow only for type-2 errors (false negatives).

larger (smaller) sensitivity of compliance costs to the compliance level p. The last term corresponds to the expected fine. It is paid if and only if there was a misconduct, which happens with probability $1-p$, and the firm was successfully audited, which happens with probability $m\eta(x)$.

For simplicity, we assume that compliance efforts and regulatory audits occur before the technology is deployed in large-scale operations: the damage occurs if and only if neither the firm nor the regulator identifies technology misconduct. Thus, the expected social cost of the technology is $(1-m\eta(x))(1-p)K$, where $(1-p)$ is the probability of a misconduct and $(1-m\eta(x))$ is the probability that the regulator fails to identify the misconduct. In addition to minimizing the expected damage, the regulator is also concerned with its monitoring and enforcement cost, $\gamma(m)$, which is an increasing and convex function of the audit probability m ($\gamma_m > 0$ and $\gamma_{mm} > 0$). In sum, the regulator chooses its audit policy m to minimize the following objective function:[8]

$$\min_m \mathcal{RC}(m) = (1-m\eta(x))\,(1-p)K + \gamma(m) \tag{2}$$

With the objective functions (1) and (2), we can now proceed to our benchmark.

4 Benchmark: explainability is firm's private information

We first examine a benchmark case with "technology-neutral regulation", which means the regulator chooses a uniform audit frequency for all technologies (i.e., regardless the explainability level). This case occurs when explainability is not observed prior to the audit: the firm first chooses its explainability level x, which remains private information. Then, it chooses its effort level p, to minimize the expected costs of compliance (1) and, simultaneously, the regulator determines its audit frequency m to minimize (2). We assume that agents form rational expectations. As stated earlier, the regulator's ability to detect the technology misconduct may increase with explainability (i.e., $\eta'(x) \geq 0$). We obtain the following lemma:

Lemma 1 (technology-neutral regulation)**.** *When explainability is firm's private information, any Nash equilibrium requires that investment in explainability, compliance, and audit frequency derive*

[8] In order to identify the tensions between firms and social interest as neatly as possible, this objective function is biaised against the firm. Including the firm's profits in the regulator's objective alters the quantitative results, but our main insights remain intact.

from the following relations:

$$C'(x) = -\Psi_x - \eta'(x)(1-p)mF \tag{TN:x}$$

$$\Psi_p = m\eta(x)F \tag{TN:p}$$

$$\gamma_m = (1-p)\eta(x)K \tag{TN:m}$$

Proof. This results from the differentiation of objective function (1) with respect to x and p, and the differentiation of objective function (2) with respect to m. □

The first relation in Lemma 1 means that the firm chooses the explainability level such that the marginal cost of explainability, which corresponds to the direct efficiency cost, equals the marginal benefit. The marginal benefit is composed of two terms: on the one hand, explainability reduces the cost of compliance (first term). On the other hand, it may also result in a higher probability that the regulator finds a misconduct (second term). The second relation means that the marginal cost of compliance equals the expected penalty for noncompliance. Finally, the third condition means that the regulator chooses its audit policy such that the marginal cost of regulation coincides with the marginal social benefit of detecting a misconduct. Solving (TN:x), (TN:p), and (TN:m) simultaneously and assuming rational expectations yields a Nash equilibrium candidate solution for m, p, and x. We denote these as x^n, p^n, m^n, where the superscript n stands for "technology-neutral".[9]

Importantly, we observe in expression (TN:x) that when explainability positively affects a regulators' audit efficacy, equilibrium explainability decreases. In contrast, Section 6 will show that the welfare-maximizing level of explainability *increases* when explainability positively affects a regulators' audit efficacy. Proposition 1 describes the environments in which voluntary explainability or obfuscation may occur.

Proposition 1 (obfuscation or explainability)**.** *When explainability is firm's private information, the firm makes voluntary investments in explainability ($x^n > x_0$) if and only if explainability strongly reduces compliance costs $\Psi(x^n, p)$ but does not strongly affect audit accuracy $\eta(x)$. Otherwise, they*

[9]The existence and uniqueness of the Nash equilibrium is not our central concern. It is, however, easily obtained under the rather mild conditions that $\eta''(x)$, $\eta'(x)$ and Ψ_{xp} be not too large (in absolute value) relative to $C'''(x)$ and Ψ_{xx}.

engage in obfuscation ($x^n < x_0$).

Proof. Voluntary investments correspond to situations in which $C'(x) > 0$, which is equivalent to $x > x_0$. Obfuscation corresponds to situations in which $C'(x) < 0$ ($x < x_0$). From equations (TN:x) and (TN:p), we easily derive that firms make voluntary investments in explainability ($x^n > x_0$) if and only if

$$(1-p)\frac{\eta'(x^n)}{\eta(x^n)} \cdot \frac{\Psi_p(x^n, p^n)}{-\Psi_x(x^n, p^n)} < 1 \tag{3}$$

, which proves the proposition. □

Recalling that audit efficacy $\eta(x)$ is a primitive of the model, Proposition 1 can be reformulated as follows: if explainability strongly affects audit accuracy $\eta(x)$ but does not significantly reduce compliance costs $\Psi(x^n, p)$, firms will engage in obfuscation. In that case, audits have a counterproductive effect and laissez faire should be envisaged. We may also observe that when compliance is costly to the firm ($\Psi_p(x^n, p^n)$ is large), which happens when internal self-audits are costly or misconducts grant a large private benefit to the firm, firms are more likely to obfuscate. The results of Proposition 1 are illustrated in Figure 2. Propositions 2 and 3 describe in more detail the rather unfavourable case in which audit efficacy is low, and strongly affected by explainability.

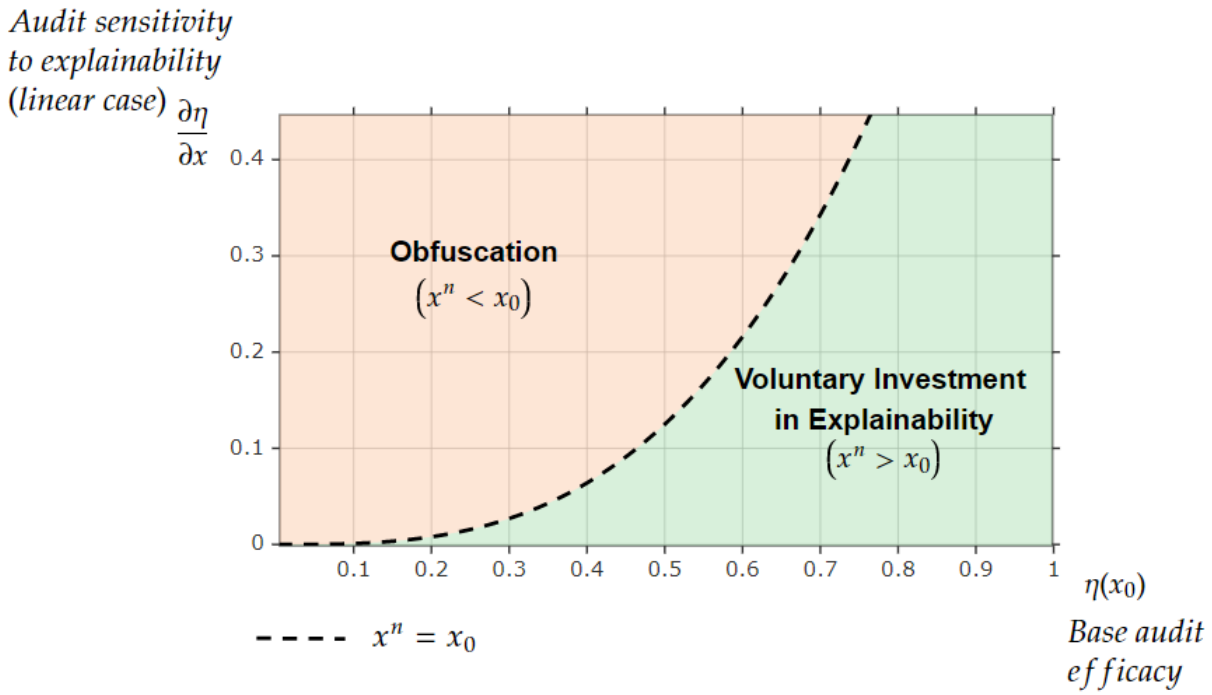
Proposition 2 (black-box algorithms)**.** *If the regulator cannot detect misconduct in a base technology, there is no investment in explainability and no investment in compliance.*

Proof. Using Lemma 1 with $\eta(x_0) = 0$, the unique Nash equilibrium is $x^n = x_0$, $p^n = m^n = 0$. □

This simple result rationalizes the commonly observed empirical fact that most machine-learning algorithms are black boxes over which regulators currently have little power. This allows firms to hide misconduct behind opaque algorithms. If such a case occurs, a minimum explainability standard should be considered.

Proposition 3 (command-and-control regulation)**.** *When explainability strongly increases the efficacy of regulatory audits, there is no voluntary investment in explainability. If implemented, the minimum explainability standard $\underline{x}$ determines the level of technology explainability.*

Figure 2: Obfuscation and explainability as a function of audit efficacy



Graph follows the parametric specification of Appendix A with $K = 1$ and $F = 4$. Audit efficacy is linear in x: $\eta(x) = max\,(0, \eta(x_0) + b(x - x_0))$. The green (orange) area corresponds to settings wheretechnology-neutral regulation generates more (less) explainability than laissez-faire.

Proof. Assume there exists a minimum explainability standard $\underline{x} \geq x_0$. If

$$\Psi_x(\underline{p}, \underline{x}) + C'(\underline{x}) + \eta'(\underline{x})(1 - \underline{p})\underline{m}f > 0,$$

with $\underline{p}$ and $\underline{m}$ derived from (TN:p) and (TN:m) evaluated in $x = \underline{x}$, then the only equilibrium is $x^n = \underline{x}$, $p^n = \underline{p}$ and $m^n = \underline{m}$. □

When the conditions of Proposition 3 are met, output regulation fails to induce explainability and input regulation (i.e. a norm on explainability) should be implemented. We acknowledge, however, that regulation of the inputs (explainability) would require to overcome the hurdle of defining and measuring precisely explainability. This could occur e.g., through audit procedures that allow to identify causal effects in algorithmic decisions, or (much-demanding) technical analysis of the code.

Appendices A and B propose applications of the results of this section to standard cost functions. In particular, Appendix B allows for the firm and the regulator to use symmetric audit technologies. In the next section, we analyze technology-specific regulation which, in some cases, further promotes explainability and compliance.

5 Mandatory disclosure of the explainability level

In this section, we assume that the explainability level x is publicly observed before the regulator chooses its audit policy. This situation would emerge if firms are mandated to disclose some of the characteristics of their decision-making processes, so regulators may infer their explainability prior to conducting the audit. This allows for technology-specific regulation, in which the regulator designs an audit frequency policy that depends on the observed level of explainability. Following standard backward-induction logic, the analysis starts in the last stage of the game (compliance and audit decision, given the observed level of explainability) and proceeds to the first stage (firm chooses explainability level). Variables of this section are denoted with a superscript s, which stands for "specific".

Stage 2: Compliance and Audit decisions

The firm selects its effort p to minimize its costs (1), given rationally anticipated audit frequency $m(x)$ and, simultaneously, the regulator determines its audit frequency $m(x)$ to minimize (2), given rationally anticipated audit frequency $p(x)$. This leads in equilibrium to the following relations, much like in equations (TN:p) and (TN:m):

$$\Psi_p = m(x)\eta(x)F \tag{TS:p}$$

$$\gamma_m = K(1-p(x))\eta(x) \tag{TS:m}$$

The interpretation of these two equations is similar to that of the benchmark of Section 4. The difference is that the decision variables p and m, now depend explicitly on explainability x. We denote $p^s(x)$, $m^s(x)$ as the solutions to (TS:p) and (TS:m). By implicit differentiation of $m^s(x)$ and $p^s(x)$ with respect to x, we obtain the following relations:

$$m_x^s = \frac{\Psi_{xp}K\eta + \eta_x K\left((1-p)\Psi_{pp} - m\eta F\right)}{KF\eta^2 + \gamma_{mm}\Psi_{pp}} \tag{4}$$

$$p_x^s = \frac{-\Psi_{xp}\gamma_{mm} + \eta_x F\left((1-p)\,K\eta + m\gamma_{mm}\right)}{KF\eta^2 + \gamma_{mm}\Psi_{pp}} \tag{5}$$

The first term of the numerator in both equations represents the "commitment to comply" that stems from explainability. This increases compliance (Equation 5), as a marginal increase in explainability decreases the marginal cost of compliance ($\Psi_{xp} < 0$). This also tends to decrease audit pressure in Equation (4), as regulators rationally anticipate that higher explainability generates greater compliance. The second term in the numerator in both equations represents the "opportunistic auditing policy" effect of explainability. This effect always increases compliance but has ambiguous effects on the auditing policy. On the one hand, a regulator may strategically audit firms with higher explainability, as doing so makes audit success more likely. This effect is captured by the term $(1-p)\Psi_{pp}$. On the other hand, high audit accuracy makes actual auditing less necessary. This effect is captured by the term $m\eta F$. Overall, the effect of explainability on audit frequency is ambiguous.

Stage 1: Explainability decision

The firm chooses explainability by rationally anticipating $p^s(x)$ and $m^s(x)$. We rewrite the objective function of the firm (1):

$$\min_x \mathcal{FC}(x, p^s(x)) = C(x) + \Psi(x, p^s(x)) + (1 - p^s(x))m^s(x)\eta(x)F \tag{1'}$$

Recalling from the resolution of the second stage that in equilibrium $\Psi_p(x^s, p^s) = m^s(x^s)F\eta(x^s)$, we obtain an implicit formulation of the firm's equilibrium investment in explainability, x^s:

$$C'(x) = -\Psi_x - (1 - p^s)F\left(\eta'(x)m^s + \eta m_x^s\right) \tag{TS:x}$$

This expression is the technology-specific counterpart of equation (TN:x). We summarize these findings in the following lemma.

Lemma 2 (technology-specific regulation)**.** *With mandatory disclosure of the explainability level, any subgame-perfect Nash equilibrium has the equilibrium investment in explainability, compliance, and audits derive from Equations (TS:x), (TS:p) and (TS:m).*

Proof. Derives from previous developments. □

From Lemma 2, we derive the following proposition:

Proposition 4 (audits unaffected by explainability)**.** *When explainability does not affect the quality of audits, i.e., $\eta'(x) = 0$, mandatory disclosure of the explainability level always favors equilibrium explainability and compliance.*

Proof. This derives from the comparison of equations (TN:x), (TN:p) and (TN:m) on the one hand, and equations (TS:x), (TS:p) and (TS:m) on the other hand. See Appendix C.1 for detailed developments. □

The intuition of this proposition is rather straightforward. When audit efficacy is not affected by explainability, the only effect of explainability on regulatory audits is the "commitment to comply": explainability allows the firm to use its investment in explainability as a signalling and

commitment device to signal its compliance efforts. Because this effect reduces the likelihood that it is audited, the firm raises its investment in explainability relative to the equilibrium investment under technology-neutral regulation (i.e., $x^s > x^n$). As we noted with regard to the benchmark of Section 4, however, the explainability of machine-learning techniques may strongly affect the efficacy of audits. In turn, this may affect the relative efficacy of technology-neutral and technology-specific regulations, as we show in Proposition 5:

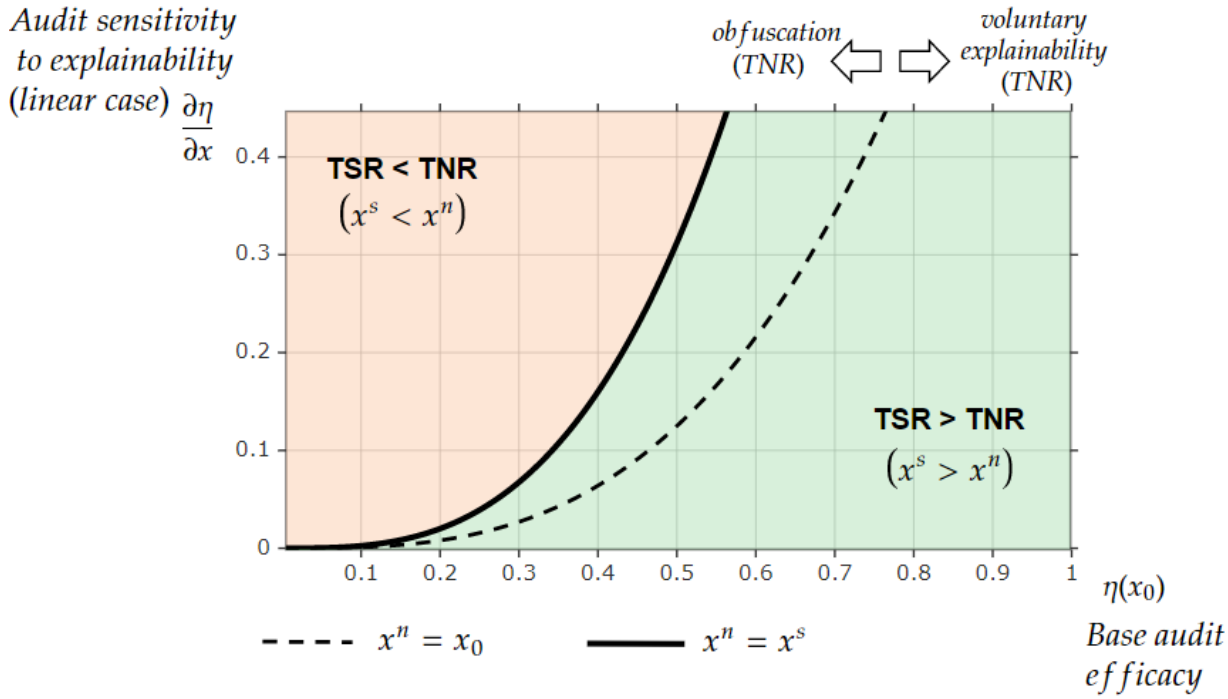
Proposition 5 (regulatory opportunism)**.** *When the fine F is small, explainability strongly affects the quality of audits ($\eta'(x)$ is large) and does not strongly affect the cost of compliance (Ψ_{xp} is small), mandatory disclosure of the explainability level induces less robust investment in explainability and compliance.*

Proof. This derives from the comparison of equations (TN:x), (TN:p) and (TN:m) on the one hand, and equations (TS:x), (TS:p) and (TS:m) on the other hand. See Appendix C.2 for detailed developments. □

The effect of technology-specific regulation on equilibrium explainability is ambiguous. It depends on the sign of m_x^s. If the opportunistic regulatory response dominates the commitment effect ($m_x^s > 0$), explainability facilitates inspection by the regulator: the firm lowers its investment in explainability relative to the investment under technology-neutral regulation. Conversely, if ($m_x^s < 0$), technology-specific regulation promotes explainability. Figure 3 illustrates the results derived from Lemma 2 and Proposition 5 based on the specification detailed in Appendix A. We observe that, when audit accuracy is low and highly sensitive to explainability ($\eta(x_0)$ is small and $\eta'(x_0)$ is large), firms are more likely to engage in obfuscation and technology-neutral regulation should be preferred over technology-specific regulation. In other words, when there is a risk of regulatory opportunism, transparency over explainability should be avoided.

To promote explainability under technology-specific regulation, the regulator should make a credible commitment to eschewing opportunism. A possible solution is that the legislator modifies the regulator's objective function (2) so it does not explicitly depend on $\eta(x)$.

Figure 3: Best regulatory regime as a function of audit efficacy



This figure uses the same parametric assumptions as Figure 2. The green (orange) area corresponds to settings where technology-specific regulation induces more (less) explainability than technology-neutral regulation. The dotted line is as in Figure 2.

6 Welfare effects

So far, the analysis focused on the effects of regulatory audits on the equilibrium level of explainability when the explanability level is firm's private information (Section 4) and under mandatory disclosure of the explanability level (Section 5). We acknowledge, however, that explainability is not necessarily a policy target *per se*. In order to formulate effective policy recommendations, it is useful to derive the first-best (i.e. welfare-maximizing) levels of explainability and compliance and compare them to the outcomes of the aforementioned regulations. In line with intuition, this section shows that a social planner interested in maximizing social welfare would generally seek to increase the equilibrium explainability level relative to laissez-faire, and also relative to the level under technology-neutral regulation.

To detect compliance, the social planner may use both the technology of the firm and that of

the regulator. The (unbiaised) social planner's objective is to minimize social costs:

$$\min_{x,p,m} \mathcal{SC}(x,p,m) = C(x) + \Psi(x,p) + K(1-p)(1-\eta(x)m) + \gamma(m) \tag{6}$$

We easily derive that the first-best levels of explainability, compliance, and audits are solutions to:

$$C'(x) = -\Psi_x + \eta'(x)(1-p)mK \tag{SP:x}$$

$$\Psi_p = (1-\eta(x)m)K \tag{SP:p}$$

$$\gamma_m = (1-p)\eta(x)K \tag{SP:m}$$

These conditions uniquely define a minimum of Equation (6) when the problem is concave, which requires that $\Psi_{xp}(x,p)$ be not too large. As shown in Equation (SP:x), the marginal investment in explainability x (in the form e.g. of an efficiency loss) must equal its marginal benefit (in the form of a lower cost of compliance, and an increase in audit efficiency). In this equation, we may also note that social-optimality always requires that there is positive investment in explainability relative to laissez-faire.[10]

Table 1 compares the first-order conditions of the first-best and the two regulatory regimes considered in the paper. A few observations are in order. First, as expected, the social planner chooses its audit frequency in the same way as the regulator in both regulatory regimes (third line). This is because regulators cannot commit to a specific audit frequency: the marginal costs of audits equal their marginal social benefit, taking firm's decisions as given.

Second, the social planner chooses a compliance effort such that the marginal cost of the compliance effort corresponds to the expected social cost, while a firm under technology-neutral or technology-specific regulation chooses its compliance effort in accordance with its expectation of a fine (second line).

Third, the social planner chooses the investment in explainability such that the direct costs equal the marginal social benefit, which is the reduction in the compliance cost, plus the social benefit of an increase in audit efficacy. In contrast, the firm weights the direct costs against the marginal private benefit, which is the reduction in the compliance cost and the private cost associated an

[10] $C'(x) > 0$, except in the extreme case when explainability does not allow to limit misconducts, nor to improve audit accuracy ($\Psi_x = \eta'(x) = 0$).

increase in audit efficacy (first line).

	Social planner	Technology-neutral regulation	Technology-specific regulation
$C'(x) =$	$-\Psi_x + (1-p)K\eta'(x)m$	$-\Psi_x - (1-p)F\eta'(x)m$	$-\Psi_x - (1-p)F\left(\eta'(x)m(x) + \eta(x)m'(x)\right)$
$\Psi_p \;=$	$K(1-\eta(x)m)$	$Fm\eta(x)$	$Fm(x)\eta(x)$
$\gamma_m \;=$	$K(1-p)\eta(x)$	$K(1-p)\eta(x)$	$K(1-p)\eta(x)$

Table 1: First-order conditions in first-best (left), technology-neutral (center) and technology-specific regulation (right)

It is not possible to compare welfare in these three scenarios without specific functional forms. Our conclusions would also be affected by the (exogenous) level of the fine F. Nevertheless, we can state a few general results that confirm that social planners would in real-life cases prefer regulatory regimes that promote explainability. We denote by an asterisk superscript the values set by the social planner (i.e. the solution to the system of equations of the first column in Table 1).

Proposition 6 (Optimum level of explainability)**.** *Assume that the fine F is set such that the equilibrium compliance level is optimal given explainability, or lower than optimal ($p^n \leq p^*$). Assume further that $\eta'(x) > 0$. Then, equilibrium quality is suboptimally low under technology-neutral regulation ($x^n < x^*$).*

Proof. We can easily show that in technology-neutral regulation, $F = 0$ results in $p^n = 0$ and setting F to infinity results in $p^n = 1$. By the intermediate value theorem is thus possible to set $F = F^*$ such that $p^n = p^*$. Then, relation SP:p is equivalent to TN:p, and SP:m is equivalent to TN:m. From the comparison of SP:x and TN:x, recalling the assumption that $\eta'(x) > 0$, and that the direct cost of explainability is increasing and convex, we observe that explainability is suboptimally low under technology-neutral regulation. □

Proposition 6 shows that, in the likely case when equilibrium compliance is optimal or deemed too low given the explainability level (either because the fine is too low or because the regulator's technology is not effective enough in inducing self-regulation), explainability under technology-neutral regulation is too low relative to the first best. We conclude that maximizing welfare requires increasing the equilibrium explainability level.

7 Conclusion and policy implications

Our analysis has highlighted an important trade-off that firms contemplating investment in explainability face: they may choose either to invest in explainability and take advantage of a reduction in compliance costs and reduce the regulatory pressure it experiences, or strategically reduce explainability so as to hide possible algorithmic misconducts behind algorithmic opacity. This decision depends crucially on the regulatory framework in which it operates, and on the regulator's ability to exploit explainable technologies when conducting an audit.

When explainability strongly affects the efficacy of regulatory audits, firms may strategically reduce explainability or even actively obfuscate their technology processes so as to render audits ineffective. This results in very low levels of compliance. In this case, audit regulation is counterproductive and laissez faire or minimum explainability standards may need to be considered. When explainability does not strongly affect audit efficacy, some degree of self-policing may be observed. When the regulator adopts a technology-specific regulation, explainability acts as a signalling and commitment device through which firms may signal compliance efforts. This tends to reduce regulatory pressure and in turn it increases further the marginal returns from increased explainability. When explainability strongly affects audit efficacy, however, another factor comes into play: regulatory opportunism deters investment in explainability. When this effect dominates, mandatory disclosure of explainability decreases equilibrium explainability and harms social welfare. This calls for a careful design of explainability regulatory policy.

Even though it is probably the application where the implications of our work are strongest, our work applies not only to AI technologies. It may indeed apply to any decision-making process which mechanisms can be rendered more or less complex in the eyes of the manager and/or the regulator. As such, it may also apply to more traditional algorithms or even to organizations where final decisions are essentially delegated to a complex hierarchy of managers and decision-makers.

Our work could be extended in several directions. First, we deliberately chose a setting where the regulator bears the burden of the proof. It is important to note that making it the responsibility of firms to demonstrate their innocence would strongly enhance incentives for explainability. Second, we assume the regulator and firms' audit technologies are independent. Allowing for correlation may highlight interesting moral-hazard and signaling effects. Finally, this work would benefit from

the consideration of leniency and voluntary disclosure programs.

References

Abada, I. and Lambin, X. (2020). Artificial intelligence: Can seemingly collusive outcomes be avoided? Available at SSRN 3559308.

Acemoglu, D. and Restrepo, P. (2020). The wrong kind of ai? artificial intelligence and the future of labour demand. Cambridge Journal of Regions, Economy and Society, 13(1):25–35.

Assad, S., Clark, R., Ershov, D., and Xu, L. (2020). Algorithmic Pricing and Competition : Empirical Evidence from the German Retail Gasoline Market. (August).

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020). The allocation of decision authority to human and artificial intelligence. In AEA Papers and Proceedings, volume 110, pages 80–84.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58.

Bebchuk, L. A. and Kaplow, L. (1993). Optimal sanctions and differences in individuals' likelihood of avoiding detection. International Review of Law and Economics, 13(2):217–224.

Becker, G. S. (1968). Crime and punishment: An economic approach. In The economic dimensions of crime, pages 13–68. Springer.

Bertsimas, D., Delarue, A., Jaillet, P., and Martin, S. (2019). The price of interpretability. CoRR, abs/1907.03419.

Brander, J. A. and Spencer, B. J. (1983). Strategic commitment with r&d: the symmetric case. The Bell Journal of Economics, pages 225–235.

Brown, Z. Y. and Mackay, A. (2019). Competition in Pricing Algorithms .

Calvano, E., Calzolari, G., Denicolo, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. American Economic Review, 110(10):3267–97.

Cam, A., Chui, M., and Hall, B. (2019). Global ai survey: Ai proves its worth, but few scale impact. McKinsey.

Dogan, M., Jacquillat, A., and Yildirim, P. (2018). Strategic automation and decision-making authority. Available at SSRN 3226222.

European Commission (2020a). Proposal for a regulation of the european parliament and of the council on a single market for digital services (digital services act) and amending directive 2000/31/ec.

European Commission (2020b). Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act). EUR-Lex, 0374.

European Commission (2021). Proposal for a Regulation on a European approach for Artificial Intelligence. Journal of Chemical Information and Modeling.

Garvie, D. and Keeler, A. (1994). Incomplete enforcement with endogenous regulatory choice. Journal of Public Economics, 55(1):141–162.

Glazer, A. and McMillan, H. (1992). Pricing by the firm under regulatory threat. The Quarterly Journal of Economics, 107(3):1089–1099.

Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. AI Magazine, 40(2):44–58.

Heyes, A. G. (1994). Environmental enforcement when 'inspectability'is endogenous: A model with overshooting properties. Environmental and Resource Economics, 4(5):479–494.

Innes, R. (1999). Self-policing and optimal law enforcement when violator remediation is valuable. Journal of Political Economy, 107(6):1305–1325.

Kaplow, L. and Shavell, S. (1994). Optimal law enforcement with self-reporting of behavior. Journal of Political Economy, 102(3):583–606.

Klein, T. (2019). Autonomous algorithmic collusion: Q-learning under sequential pricing. Amsterdam Law School Research Paper, (2018-15):2018–05.

Lyon, T. P. and Maxwell, J. W. (2016). Self-regulation and regulatory discretion: Why firms may be reluctant to signal green. In Strategy Beyond Markets. Emerald Group Publishing Limited.

Malik, A. S. (1990). Avoidance, screening and optimum enforcement. The RAND Journal of Economics, pages 341–353.

Marvão, C. and Spagnolo, G. (2018). Cartels and leniency: Taking stock of what we learnt. In Handbook of Game Theory and Industrial Organization, Volume II. Edward Elgar Publishing.

Maxwell, J. W. and Decker, C. S. (2006). Voluntary environmental investment and responsive regulation. Environmental and Resource Economics, 33(4):425–439.

Maxwell, J. W., Lyon, T. P., and Hackett, S. C. (2000). Self-regulation and social welfare: The political economy of corporate environmentalism. The Journal of Law and Economics, 43(2):583–618.

Polinsky, A. M. and Shavell, S. (2000). The economic theory of public enforcement of law. Journal of economic literature, 38(1):45–76.

Shavell, S. (2009). Foundations of economic analysis of law. Harvard University Press.

Short, J. L. and Toffel, M. W. (2010). Making self-regulation more than merely symbolic: The critical role of the legal environment. Administrative Science Quarterly, 55(3):361–396.

Suijs, J. and Wielhouwer, J. L. (2019). Disclosure policy choices under regulatory threat. The RAND Journal of Economics, 50(1):3–28.

Appendices

A An application

We now propose an application of our model. Assume that the firm's costs of compliance is $\Psi(x,p) = \frac{p^2}{a+x}$ with $a > 0$. This specification verifies the assumptions formulated in Section 3 ($\Psi_p > 0$, $\Psi_{pp} > 0$, $\Psi_x < 0$, $\Psi_{xp} < 0$). The firm's cost of explainability is $C(x) = \frac{x^2}{2}$. This means that the base explainability is $x_0 = 0$. Finally, the monitoring cost of the regulator is $\gamma(m) = \frac{m^2}{2}$.

A.1 No explainability

The first-order condition for explainability (TN:x) can be rewritten as :

$$A(x) \equiv x + \eta'(x)mF(1-p) - \frac{p^2}{(a+x)^2} = 0 \tag{7}$$

Assume further for simplicity that $\forall x, \frac{\partial A(x)}{\partial x} > 0$ so the firm's cost-minimizing exercise has an interior solution. It suffices for example to impose the condition that η_{xx} not be too negatively large. We have that

$$A(0) = \eta'(0)mF(1-p) - \frac{p^2}{a^2} \tag{8}$$

Assume that $\eta'(0) = 0$ or that $\eta'(0) > 0$ but p is large enough even when there is no explainability. Using, (TN:p), the latter outcome occurs when the regulator's ability to detect technological misconduct in the absence of explainability and the fine are sufficiently large (i.e., $F\eta(0)^2 \gg 0$). In these cases we have that $A(0) < 0$. We note from Equation (7) that $\lim_{x\to+\infty} A(x) > 0$. As $A(x)$ is always continuous and differentiable, there exists a positive level of explainability, $x^* > 0$, such that $A(x^*) = 0$: we conclude that self-policing emerges, even in the absence of a minimum explainability standard, when $F\eta^2(0)$ is large enough or $\eta'(0)$ is small enough. Failing this, $A(0) > 0$ and there is obfuscation: $x^n < 0$.

A.2 Explainability or obfuscation?

By Definition 1, if $x^n > 0$, firms make their algorithms explainable. If $x^n < 0$ they obfuscate them. In equilibrium, solving (TN:p) and (TN:m) allows us to derive equilibrium p and m :

$$m = \frac{2K\eta(x)}{2 + (a+x)KF\eta^2(x)}$$

$$p = \frac{(a+x)KF\eta^2(x)}{2 + (a+x)KF\eta^2(x)}$$

The equilibrium explainability x^n is then derived from (TN:x):

$$x^n = x_0 + \frac{\eta(x^n)KF}{(2 + \eta^2(x^n)KF(x^n + a))^2} \left(\eta^3(x^n)KF - 4\eta'(x^n)\right) \tag{9}$$

We conclude that, if $\eta^3(x^n)KF > 2\eta'(x^n)$, firms invest in explainability ($x^n > x_0$). Otherwise, they obfuscate their algorithms ($x^n < x_0$). Recalling that $\eta(x)$ is a primitive of the model, this conclusion can be reformulated as follows: when the fine and audit accuracy are too low relative to the sensitivity of audit accuracy to explainability, firms strategically make their algorithms less transparent.

For the specific illustration of Figure 2, we assume that $\eta(x) = max\left(0, \eta(x_0) + b(x - x_0)\right)$. b is the sensitivity of audits to explainability. From (9), we find that technology-neutral regulation is more effective than laissez-faire if and only if $b \leq \frac{\eta(x_0)^3 KF}{4}$. This region of the parameters is represented by the green area in Figure 2.

B A special case: the firm and the regulator have the same audit technology

For the sake of generality, the main text assumes very general and flexible formulations for audit efficacy $\eta(x)$ and the cost of compliance $\Psi(x, p)$. In particular, we don't specify how these two functions may relate to each other. However, it may be useful to microfound how an increase in explainability will concurrently facilitate both the undertaking of compliance and the efficacy of regulatory audits. For example, $\eta(x)$ may simply be interpreted as the probability that the regulator

will "see through" the algorithm and perfectly understand its functioning. With probability $1-\eta(x)$ the algorithm remains obscure and no fine can be imposed. Symmetrically, it may be natural to assume that the firm also "sees through" its own algorithm with the same probability, given that both forms of audits (internal or regulatory) involve human beings with arguably similar technical skills. To reflect this, we specify the compliance costs as follows:

$$\Psi(x,p) = \frac{p^2}{\eta^2(x)} \tag{10}$$

This cost function ensures that an increasing and convex effort of $e^2 = \Psi(x,p)$ results in a probability of compliance of $p = \eta(x)e$. In other words, the possibility of increasing compliance is conditional on seeing through the algorithm, which happens with probability $\eta(x)$. Failing that, the effort is wasted.

Applying this specification to Proposition 1, we have that firms invest in explainability if and only if $p^n > 2/3$, or equivalently (using Lemma 1), $\eta^4(x^n)FK > 4$. We retrieve the intuitive result that if regulatory audits are not efficacious, fine is small and environmental damage is small, firms don't invest in explainability.

C Technology-specific regulation

C.1 Proof of Proposition 4: Explainability does not affect audit efficacy

Recall that $\Psi_{px} < 0, \Psi_{pp} > 0$. When audit efficacy does not depend on explainability (i.e., $\eta_x = 0$), we derive the following results by applying the implicit function theorem to (TS:p) and (TS:m) and solving for m_x^s and p_x^s:

$$m_x^s = \frac{K\eta\Psi_{px}}{KF\eta^2 + \gamma_{mm}\Psi_{pp}} < 0 \tag{11}$$

$$p_x^s = \frac{-\Psi_{xp}\gamma_{mm}}{KF\eta^2 + \gamma_{mm}\Psi_{pp}} > 0 \tag{12}$$

The equilibrium investment x^s is solution to this equation:

$$\Psi_x = -C_x - (1-p^{s*})\eta F m_x^{s*} > 0 \tag{13}$$

We now follow the proof derived in Maxwell and Decker (2006), Proposition 1(B). As the authors notice, we cannot in general compare equilibrium investment between technology-neutral and technology-specific regulations directly because doing so involves making comparisons across two separate models. Following Brander and Spencer (1983), though, such a comparison is possible in our model. The proof makes use of the mean value theorem, which we recall here. Let $h(x)$ be a continuously differentiable function defined over the set of real numbers R^2 and let x^* and x^n be two points on this function. Then there exists a point x^c such that

$$\Delta h = h(x^s) - h(x^n) = \frac{\partial h}{\partial x}|_{x=x^c}(x^s - x^n) \tag{14}$$

where $x^c = x^n + \theta(x^s - x^n)$ and $\theta \in (0, 1)$.

Using this, we first define $\Delta x = x^s$, where x^s and x^n are investments in explainability in the responsive and unresponsive cases, respectively. Let $h(x)$ be $\frac{\partial FC^s(x)}{\partial x}$. We then apply 14 as follows: $\Delta \frac{\partial FC^{s*}}{\partial x} = \frac{\partial FC^s}{\partial x}|_{x=x^s} - \frac{\partial FC^{s*}}{\partial x}|_{x=x^n} = \frac{\partial^2 FC^s}{\partial x^2}|_{x=x^c}(x^s - x^n)$. Re-arranging terms we can get

$$(x^s - x^n) = \frac{\frac{\partial E^s}{\partial x}|_{x=x^s} - \frac{\partial FC^s}{\partial x}|_{x=x^n}}{\frac{\partial^2 FC^s}{\partial x^2}|_{x=x^c}} \tag{15}$$

From (13) we know that $\frac{\partial E^s}{\partial x}|_{x=x^s} = 0$ and $\frac{\partial FC^s}{\partial x}|_{x=x^n} < 0$. Therefore, the numerator in (15) is positive. Since cost minimization requires that $\frac{\partial^2 FC^s}{\partial x^2}|_{x=x^c} > 0$, we can conclude that $(x^s - x^n) > 0$. As Brander and Spencer (1983) note, existence and uniqueness are difficult to establish in stage games, so second-order partials derived from such games are difficult to sign in practice. In our case, we find that:

$\frac{\partial^2 FC^s}{\partial x^2}|_{x=x^c} = C_{xx} + \Psi_{xx} - \frac{\partial m^s}{\partial x}\frac{\partial p^s}{\partial x}\eta F + (1 - p^s)\eta F \frac{\partial^2 m^s}{\partial x^2}$

Note that the first three terms in this equation are positive, but the third term is indeterminate because $\frac{\partial^2 m^s}{\partial x^2}$ cannot, in general, be signed. Straightforward differentiation shows that each component of $\frac{\partial^2 m^s}{\partial x^2}$ involves third-order cross partial derivatives from our $\Psi(x, p)$ function. As a result $\frac{\partial^2 m^s}{\partial x^2}$ is ambiguous. In practice, such third -and higher- order effects are reasonably assumed to be relatively small. Because the first two terms in the preceding equation are higher-order effects of the correct sign, we follow Brander and Spencer (1983) and reasonably assume that, overall,

$\frac{\partial^2 m^{s*}}{\partial x^2} < 0$.

C.2 Proof of Proposition 5: Explainability affects audit efficacy

From the comparison of (TN:x) and (TS:x) we observe that technology-specific regulation induces greater explainability and compliance than technology-neutral regulation if and only if explainability induces less robust audit efforts by the regulator ($m_x^s < 0$). Recall that $\Psi_{xp} < 0$, $\Psi_{pp} > 0$. An analysis of the terms in Equation (4) proves Proposition 5. The following lemma provides the specific formal condition for explainability to be greater in technology-specific than in technology-neutral regulation:

Lemma 3 (technology-neutral or technology-specific regulation)**.** *Technology-specific regulation induces more explainability and compliance than technology-neutral regulation if and only if :*

$$\frac{\eta'(x^s)}{\eta(x^s)} \cdot \frac{1}{-\Psi_{xp}(x^s, p^s)} \big((1-p^s)\Psi_{pp}(x^s, p^s) - \Psi_p(x^s, p^s)\big) < 1 \tag{16}$$

Proof. Inserting (TS:p) in Equation (4) with the condition that $m_x^s < 0$, we obtain condition (16). □

The expression of Lemma 3 is not easy to interpret. With the specification of Appendix A, we may derive conditions for technology-specific regulation to dominate technology-neutral regulation that are easier to interpret. From the comparison of (TN:x) and (TS:x) we know that this is the case if and only if $m_x^s < 0$. Using Equation (4) and the expressions of Ψ_{xp} and Ψ_{pp}, this condition reduces to :

$$2b(1-2p^s) < \eta(x^s)^2 m(x^s) F \tag{17}$$

Audit sensitivity to explainability b, as well as the term to the right of the inequality are always nonnegative. We conclude that condition (17) is met if and only if either of these conditions is met:

- **condition 1:** $p^s > 1/2$, which by (TS:m) and (TS:p) is equivalent to $\eta(x^s)^2 FK(x^s + a) > 2$
- **condition 2:** $b < \frac{\eta(x^s)^3 KF}{4} \frac{1-p^s}{1/2-p^s}$

This means that technology-specific regulation outperforms technology-neutral regulation if and only if equilibrium compliance is high (Condition 1 requires that base explainability, social damage, fine, and explainability effect on compliance costs are high), or if the audit efficacy is not too sensitive to explainability (Condition 2). If neither of these conditions is met, technology-neutral regulation outperforms technology-specific regulation.