Regulation and Responsible Innovation*

Nicolas Figueroa† Carla Guadalupi‡ Jorge Lemus§

February 9, 2025

Abstract

Should product harms be investigated by firms via "Responsible Innovation" (RI) or by regulators? Firms could investigate potential harms early on, using resources otherwise allocated to innovation. Regulators, by contrast, typically assess products only after they reach the market. We characterize the efficient solution and show that RI is not always efficient; in some cases, regulators must bear the entire burden of investigating harm. We then examine whether the efficient solution is implementable in a dynamic game where a firm chooses its RI investment, whereas a regulator, who advocates for consumers, can levy fines, ban products, and investigate the harm. Our main insights are twofold. First, the efficiency of regulatory actions requires mild constraints on the regulator when the harm is very small or large, but more restrictive measures when the harm is moderate. Second, efficient ex-post regulation deters firms from engaging in responsible innovation, so to encourage it, regulatory decisions must be distorted. Effective regulatory regimes must balance incentives for responsible innovation with necessary oversight.

JEL Classification: L51, M14, O32, D82
Keywords: Innovation, Regulation, Corporate Responsibility, Technological Innovation, R&D.

*For comments and suggestions we thank participants and discussants at IIOC(2024), ALEA (2024), JEI (2024), Universidad Andrés Bello, TOI Chile (2024).
†Pontificia Universidad Católica, Chile. Email: nicolasf@uc.cl
‡Universidad Andres Bello, Chile. Email: carla.guadalupi@unab.cl.
§University of Illinois at Urbana-Champaign. Email: jalemus@illinois.edu.

1 Introduction

Disruptive technologies can transform industries and significantly enhance welfare. However, they also carry inherent risks that may result in unforeseen and severe social harm.[1] Regulators often struggle to make timely decisions when faced with rapidly emerging technologies, raising the economic question of who should be responsible for investigating the potential harms of an innovation. Firms could engage in "Responsible Innovation" (RI) by proactively investigating and disclosing potential harms early on, or focus exclusively on innovation, leaving the task of investigating risks to regulators.

Recent cases highlight the decisions firms face regarding RI. OpenAI recently disbanded a team that managed long-term risks.[2] The team was promised 20% of the company's computing resources but it did not materialize.[3] This follows the decision by OpenAI's board of directors to temporarily remove Sam Altman as CEO, partly due to worries that he was prioritizing profitability over addressing potential AI risks.[4] A former OpenAI safety team leader stated that "over the past years, safety culture and processes have taken a backseat to shiny products."[5] Similarly, other technology companies have disbanded ethics teams, prioritizing profitability over long-term risk management, in line with Silicon Valley's traditional mantra "move fast and break things."[6]

Given these trends, the regulatory framework governing emerging technologies is crucial. In this paper, we model the trade-off firms face between pursuing innovation and investing in learning about potential harms. We explicitly capture the reality that regulators often act after the product has already entered the market, and they do so subject to certain regulatory constraints, which influence their ability to mitigate risks. Our analysis identifies situations where regulation can achieve efficiency in equilibrium, and others where either the firm or the

[1] Such technologies include artificial intelligence, social media, cryptocurrencies, biometrics, and self-driving cars. See, e.g., https://time.com/6344160/a-year-in-time-ceo-interview-sam-altman/.

[2] https://www.wired.com/story/openai-superalignment-team-disbanded/

[3] https://techcrunch.com/2024/05/17/openai-created-a-team-to-control-superintelligent-ai-then-let-it-wither-source-says/

[4] https://www.nytimes.com/2023/11/18/technology/open-ai-sam-altman-what-happened.html

[5] https://www.theverge.com/2024/5/17/24159095/openai-jan-leike-superalignment-sam-altman-ai-safety

[6] Google has also had a controversial relationship with their ethics teams, see https://www.bloomberg.com/news/features/2023-04-19/google-bard-ai-chatbot-raises-ethical-concerns-from-employees. Meta recently disbanded its "Responsible AI" division, dedicated to regulating AI product safety, and reallocate its members to product development divisions, see https://www.cnbc.com/2023/11/18/facebook-parent-meta-breaks-up-its-responsible-ai-team.html. More broadly, Ahmed et al. (2024) document a lack of industry engagement in responsible AI research, with nearly 90% of AI firms with commercial patents conduct no research into responsible AI, and among AI research firms, only 11.2% engage in substantive responsible AI research.

regulator make inefficient choices. Ultimately, achieving the right balance between innovation and RI requires regulatory regimes tailored to specific product and market characteristics.

Our paper makes a distinct contribution by analyzing the dynamic interaction between firms' incentives for RI and regulatory oversight. Previous contributions to the literature have focused separately on how a fixed regulatory regime shapes innovation decisions (see, e.g., Immordino et al., 2011) or learning after the product is already in the market, ignoring the firm's innovation trade-off (see, e.g., Henry et al., 2022). Our model introduces a delegation problem where the regulatory actions occurring after the product is in the market influence the firm's pre-entry allocation of resources between innovation or RI. Thus, our unified framework captures both the firm's trade-off between innovation and RI, as well as the strategic decisions made by regulators. One of our key findings is that fines alone are generally insufficient to achieve efficiency because they distort both the regulator's ability to investigate harm and discourage the firm from investing in RI. Thus, more nuanced regulatory actions are necessary to effectively align both parties' incentives.

In our setting, a firm can allocate resources to investigate a product's potential harm before market entry at the expense of diverting them from innovation. Our analysis considers products that cannot be easily modified once the harm is identified, leaving regulators with the option to ban them or impose fines, which can also be interpreted as partial remedies (e.g., usage limits or warning labels) but may be insufficient to counteract all the harm.[7,8]

The regulator, who advocates for consumers, makes decisions only after the product enters the market.[9] He is constrained by a regulatory regime, which defines the penalties for known harm, the ability to ban products without proof of harm, and the budget for investigating the potential harm. If the harm is known, he can ban the product or allow it as long as the firm makes a transfer to consumers (alternatively, a costly investment to reduce the harm), which we simply call a fine. If the harm is unknown, the regulator can investigate to learn it. If the investigation is inconclusive, he may be able to ban the product without proof of harm under the "precautionary principle."[10]

[7] Examples of harmful products that are hard to modify include asbestos, lead-based paint, certain persistent organic pollutants, tobacco products, and some core features of social media platforms like Instagram (which some research links to teen depression).

[8] In our model, we can interpret fines as an investment made by the firm to mitigate the harm (with a particular mitigation technology), or as a permit for using the product (e.g., a pollution permit).

[9] Many agencies are explicitly mandated to protect consumers from excessive pricing or unsafe practices. For instance, the FDA's primary mandate is to safeguard public health, while the Consumer Financial Protection Bureau in the U.S. or national consumer watchdogs in Europe also prioritize consumer protection over maximizing aggregate surplus.

[10] For instance, the FDA can ban a device without proof of illness or injury based on risk and potential harm. See https://www.fda.gov/medical-devices/medical-device-safety/medical-device-bans.

Our main insights are twofold. We first show that the expected severity of harm plays a crucial role in determining whether efficient regulatory actions can be achieved through delegation. When the expected harm is either very small or very large, mild regulatory constraints are sufficient to implement efficient regulatory outcomes, as the misalignment caused by the regulator's focus on consumer surplus, while ignoring firm profits, is relatively modest. However, for moderate harm — where the harm is significant relative to consumer surplus but small compared to total surplus — stronger regulatory constraints are needed to achieve efficiency. Our second insight is that regulatory regimes that ensure efficiency once the product is on the market discourage firms from investing in RI. In other words, to incentivize RI, regulatory decisions must be inefficient. To derive these insights, we first characterize the efficient solution and then explore how this efficient outcome could be implemented through a combination of fines, probabilistic access to the precautionary principle, and the resources available to investigate potential harm. In other words, to encourage RI, the regulator's decisions must be inefficient. To derive these insights, we first characterize the efficient solution, and then study whether and how the efficient outcome could be implemented by a combination of fines, probabilistic access to the precautionary principle, and resources to investigate the harm.

Responsible innovation is efficient when the social value of learning about the harm outweighs the costs of potentially failing to innovate. For example, if the regulator's cost of investigating is low, it may be efficient to let the regulator handle the investigation without relying on RI. Otherwise, it may be efficient to have the firm engage in RI at the expense of reducing the probability of innovation success. Furthermore, we show that the efficient levels of RI and the regulator's investigation transition from substitutes to complements as the product's social value increases. Specifically, when the social value is low, it is efficient to rely primarily on RI to assess the harm. As the social value rises, the regulator is willing to investigate more, while the firm allocates fewer resources to RI. This is because the benefit of innovating increases as the product's social value grows. Once the product's social value exceeds its expected harm, the firm invests less in RI and regulator also investigates less as the social value continues to rise because the product's benefit outweigh its potential harm.

We then analyze whether it is possible to implement the regulatory and firm's efficient decisions in equilibrium. An insight is that to implement efficient regulatory actions, the regulatory regime should reduce the restrictions on the regulator and impose higher fines,

Principle 15 in the 1992 Rio Declaration states that whenever "there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation." https://www.un.org/en/conferences/environment/rio1992

as the expected harm increases.[11] For small expected harm, it is key to limit the regulator's investigative resources because he is overly cautious (investigates too much) relative to the efficient level. For moderate expected harm, a regulatory regime should impose higher fines but still limit investigative resources and forbid the regulator from banning the product without proof of harm. For large expected harm, fines must be maximal to incentivize an efficient investigation, and the regulator must be given enough resources to investigate and the power to ban the product without proof of harm.

We then show that firms do not invest in RI under regulatory regimes that implement efficient regulatory actions. To see why this occurs, consider two cases. First, when the expected harm is not large, the regulator allows the product when the harm is unknown. In this case, revealing the harm would only lead to fines, so the firm has no incentive to invest in RI. Second, when the expected harm is large, the regulator bans the product if the harm is unknown and imposes maximal fines if the harm is known. To avoid a ban, the firm might invest in RI to preemptively demonstrate harm, but it would earn zero profits from doing so, as the fines are maximal. As a result, the firm has no incentive to invest in RI. Therefore, to implement the efficient level of RI, it is necessary to distort the regulator's decisions. For instance, allowing the regulator to ban the product without proof of harm when the expected harm is small, or limiting fines when the expected harm is large can incentivize the firm to invest in RI.

It is worth emphasizing that, when the harm is known, the regulator internalizes firm profits through fines.[12] However, when the expected harm is large and the harm is unknown, the product is banned. Hence, fines are insufficient to align incentives because the regulator does not account for the loss in firm profits whenever his investigation is inconclusive. Conversely, if the expected harm is small, the prospect of collecting fines may motivate an excessive investigation. From the perspective of the firm, if fines are not too harsh, it may be worth it to invest in RI because revealing information can avoid unnecessary bans from an overly cautious regulator who bans the product without proof of harm.

Our results contribute to the open debate on how to regulate innovative products. Some scholars have argued that they should be subject to strict liability.[13] Others propose ex-ante

[11] Generally, the regulator focuses on consumers and ignores firm profits, making him more cautious than a planner considering overall welfare. This excess caution can result in an inefficiently large investigation or the overuse of the precautionary principle, leading to product bans without evidence of harm.

[12] In practice, fines are often limited due to constraints on pecuniary damages. These limitations prevent fines from being excessively large, even in cases involving significant harm (Weil, 2024).

[13] https://www.vox.com/future-perfect/2024/2/7/24062374/ai-openai-anthropic-deepmind-legal-liability-gabriel-weil

regulation.[14] Some even propose banning products when the harm is unknown.[15] Our results show that there is no one-size-fits-all solution, and regulatory regimes should be flexible and tailored to the specific market and product characteristics. In particular, a regulator should face different constraints, in terms of resources and discretionary power, depending on the social value of the product, its expected harm, and the wedge between consumer and total surplus.

Generally, regulators should have more resources and discretion for products that are expected to be more harmful. However, efficient regulation can discourage the firm from investing in RI. Responsible innovation is valuable precisely when the regulator faces a high cost of learning about the harm. In particular, disruptive technologies tend to be hard to understand for regulators, which can make it costly to assess the harm for them.[16] These are situations where responsible innovation should be encouraged. Our results suggest that regulation in these cases should share some features of an "ex-ante" approval regime, that is, banning the product if harm remains unknown after a (perhaps limited) investigation by the regulator. Moreover, perhaps counterintuitively, if the firm produces proof of harm, the regulator should be "lenient" and not penalize the firm so harshly for revealing such findings.

2 Literature Review

The literature has examined how different regulatory regimes shape firms' incentives to gather information about potential harm absent innovation incentives. Shavell (1992) shows that strict liability motivates firms to acquire information and prevents harm when conclusive evidence can be obtained at a fixed cost. Building on this model, other papers have explored how results vary with the information acquisition technology. Friehe and Schulte (2017) show that inefficiencies arise when firms may not obtain definitive evidence, while Requate et al. (2023) find that strict liability may lead to greater efficiency when firms can improve their knowledge over time. Baumann and Friehe (2016) investigates dynamic learning by incorporating historical accident data into firm decisions (learning-by-doing) and conclude that relying solely on such data may result in sub-optimal care under strict liability. Henry et al. (2022) examine the use of liability, withdrawal, and authorization when information is

[14] https://cepr.org/publications/dp18517

[15] https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/

[16] During a congressional hearing, some lawmakers were criticized for their limited understanding of Facebook, emphasizing the urgent need for more informed social media oversight. https://www.vox.com/policy-and-politics/2018/4/10/17222062/mark-zuckerberg-testimony-graham-facebook-regulations

revealed gradually. Ottaviani and Wickelgren (2023) examines the trade-off between ex-ante and ex-post regulation, when it is costly to undo a firm's action. In these papers, the product has already been introduced, so firms do not face a strategic decision between investing in innovation and acquiring knowledge about potential harm. Moreover, we allow for strategic learning on both the firm's and the regulator's side.

Other articles examine how different regulatory regimes influence firms' incentives to innovate, when regulators do not actively engage in learning about product harm. Immordino et al. (2011) show that laissez-faire regulation is optimal for low-risk products, with escalating fines for dangerous ones. Strict authorization is optimal when harm is likely and fines are bounded. De Chiara and Manna (2022) study how the regulatory regime affects investment under public official corruption rather than considering the regulator's investigation or the firm's trade-off between innovation and information. Corruption deters investment under strict liability but may still permit innovation in other regimes. In these articles, it is assumed that a successful investment reveals the product's harm once it is on the market, with higher investment increasing both the likelihood of success and the chance of learning about harm. In contrast, our framework captures the tension between enhancing the chances of successful product development and learning about the harm. Moreover we consider a setting in which the regulator strategically learns about the potential harm.

Furthermore, our work complements recent studies on regulation and innovation. Acemoglu and Lensman (2024) and Gans (2024) discuss trade-offs between regulatory regimes in gradual technology adoption, finding that socially optimal adoption is often gradual due to potential harms. However, these models do not consider strategic learning by the firm or regulator, nor the trade-off between innovating quickly rather than generating information.

3 Model

We consider a dynamic game between a firm and a regulator. The firm attempts to develop a new product, which may cause social harm $h \sim G[0, \infty)$. A successful product generates a profit of π for the firm, and a net benefit of $V - h$ for the regulator.

During the innovation phase, the firm allocates its limited resources (normalized to 1) between product development and "Responsible Innovation" (henceforth RI).[17] By investing a fraction $1 - p$ of its resources in product development, the firm succeeds with probability $F(1 - p)$.

[17] See Footnote 3 for an illustration of the tradeoff that arises from limited resources in practice.

The function F is a probability distribution on $[0,1]$ with continuous and positive density, f. The remaining fraction p is allocated to RI, which publicly reveals the product's harm with probability p, and reveals nothing with probability $1-p$.[18]

The regulator has been delegated the task of monitoring and enforcing appropriate actions against this potentially harmful product. A central feature of our model is that the regulator is *unable* to investigate a product's harm before its market entry, capturing the so-called 'pacing problem.' Only after market entry, the regulator can act immediately or investigate the product's potential harm before acting.[19]

As mentioned in the Introduction, our analysis focuses on innovative products that cannot be easily modified (see Footnote 7). Therefore, whenever h is known, the regulator can either ban the product (action $b_h = 0$) or allow it (action $b_h = 1$) in exchange for harm-specific fine, $L(h)$. This fine can also be interpreted as a costly investment by the firm to mitigate consumer harm (e.g., warning labels).[20] If h remains unknown after the firm enters the market, the regulator may investigate (action $a_U = 1$) or try to ban the product outright (action $a_U = 0$). If the regulator investigates, he learns h with probability r at a cost of $c(r)$, where $c(\cdot)$ is increasing, convex, and $c(0) = 0$.[21] If the investigation is inconclusive, the regulator chooses whether to allow the product (action $b_U = 1$) or try to ban it (action $b_U = 0$). Because fines are generally imposed to penalize a violation of established safety standards or known risks, regulators cannot typically impose fines without evidence of harm.

The regulator's decisions are constrained by a *regulatory regime* $(L(\cdot), \phi, \bar{r})$. First, if there is proof of harm, the regulator imposes harm-specific fines, so the firm's payoff is $\pi - L(h)$ and the regulator's payoff is $V - h + L(h)$. Without proof of harm, the regulator can choose to ban the product either before his investigation or afterward but the product ban only succeeds with probability $\phi \in [0,1]$. Banning a product without evidence of harm reflects the use of the "precautionary principle," and ϕ captures the political and judicial hurdles that may prevent the regulator from applying it. Lastly, the regulator has limited resources to investigate the harm, constraining the probability of learning h to be less than or equal to $\bar{r}$.

Figure 1 illustrates the timing of the model and the regulator's actions.

[18] E.g., Lewis and Sappington (1994) and Ottaviani and Sørensen (2006) use this "truth-or-noise" structure.

[19] We assume that during the investigation, given limited consumer awareness, both the firm's profits and the extent of the harm are moderate. For simplicity, we assume that both profits and harm are zero during the regulator's investigation.

[20] See footnote 8. In this interpretation, each dollar the firm invests reduces consumer harm by one dollar.

[21] We also impose that $c'(0) = 0$ and $c'(1) \to \infty$ to focus on unique interior solutions.

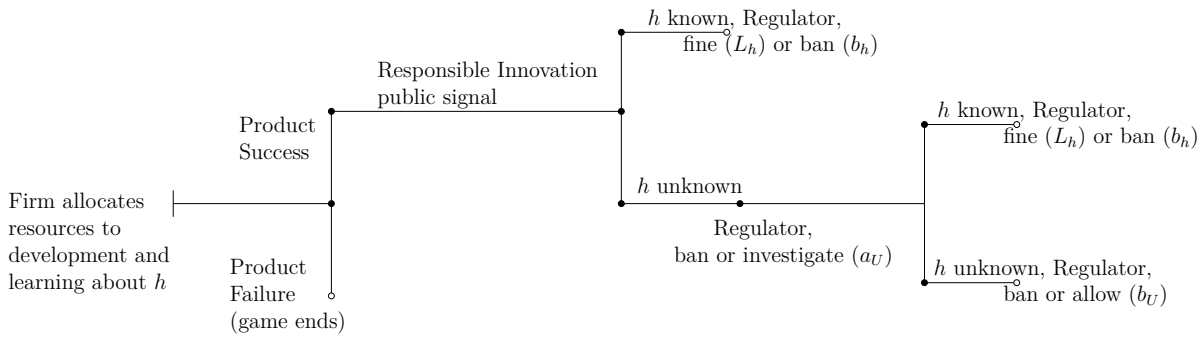

Figure 1: Timing of the model.

The firm's and the regulator's decisions are summarized by $(p, r, (b_h)_{h\geq 0}, b_U, a_U)$. The equilibrium decisions are characterized by using backward induction. We make the following technical assumption to streamline our analysis:

Assumption 1. *The function* $\frac{F(x)}{f(x)} + x$ *is strictly increasing in* x.[22]

4 Efficient Solution

We first analyze the efficient solution, where a planner makes both the firm's and the regulator's decisions. When the product is successfully developed and allowed in the market, the planner receives the total surplus, $S = \pi + V$, minus the harm, h.[23] We denote by $(p^*, r^*, (b_h^*)_{h\geq 0}, b_U^*, a_U^*)$ the efficient outcome.[24]

It is efficient to allow the product when the harm is smaller than the total surplus, and therefore, $b_h^* = 1(h \leq S)$. Without proof of harm, the efficient decision is $b_U^* = 1(E[h] \leq S)$.

Efficient Investigation. When RI fails to reveal h and the regulator investigates, his efficient probability of learning, r^*, is the solution to:

$$R^* \equiv \max_{r\in[0,1]} \; r\int (S-h)b_h^* dG(h) + (1-r)(S - E[h])b_U^* - c(r). \tag{1}$$

With probability r, the planner learns h and allows the product whenever $h \leq S$, generating a surplus of $S - h$. With probability $1 - r$, he does not learn h, allowing the product

[22] This assumption holds for distributions with decreasing reverse hazard rate (e.g., Weibull, Gamma, Pareto, and Lognormal) as well as distributions with decreasing hazard rate (see, e.g., Kayid et al., 2011).

[23] The efficient solution is not constrained by $\bar{r}$ or ϕ. Fines $L(h)$ are irrelevant because they net out.

[24] This is a "second-best" solution because the planner does not set the firm's prices.

and obtaining $S - E[h]$ whenever the expected surplus is positive. The regulator's efficient learning probability, r^*, and the decision to investigate the harm, a_U^*, are characterized in the next lemma.

Lemma 1. *If RI does not reveal h, it is efficient to further investigate the harm, $a_U^* = 1$, and spend resources to learn h with probability r^*, where*

$$c'(r^*) = \begin{cases} \int_S^{\infty} (h - S) dG(h) & \text{, if } S \geq E[h], \\ \int_0^{S} (S - h) dG(h) & \text{, if } S < E[h]. \end{cases}$$

When $S \geq E[h]$, the product is allowed in the market, unless new information reveals substantial harm (i.e., $h > S$). Therefore, learning about events at the "right tail" of the distribution of harm is valuable. The opposite happens when $S < E[h]$ because the product is banned unless new information reveals minimal harm, so learning about events at the "left tail" of the distribution of harm is valuable.

Efficient RI. We now determine the efficient fraction of the firm's resources to allocate to RI. The larger this fraction, the lower the likelihood of successful product development. The efficient investment in RI resolves this tradeoff by solving

$$\max_{p \in [0,1]} F(1-p) \left\{ p \int_0^S (S - h) dG(h) + (1-p) R^* \right\}. \tag{2}$$

With probability $F(1-p)$, product development is successful. Then, with probability p, RI reveals the product's harm, h, and the product is allowed in the market whenever $h \leq S$, generating a surplus of $S - h$. With probability $(1-p)$, h remains unknown and the planner receives the expected continuation payoff of R^*. Problem (2) can be rewritten as

$$\max_{p \in [0,1]} F(1-p)[A + Bp], \tag{3}$$

where $A \equiv R^*$ corresponds to the expected payoff when the firm does not generate evidence about h (the "uninformed entry payoff"), and $B \equiv \int_0^S (S - h) dG(h) - R^*$ reflects the change in expected payoff resulting from generating early evidence through RI (the "information premium"). The next lemma characterizes the solution to the general formulation in (3) as a function of the uninformed entry payoff, A, and the information premium, B.

Lemma 2. *Consider A, B such that $A, B \geq 0$. The function*

$$Z(p) \equiv \frac{F(1-p)}{f(1-p)} + 1 - p,$$

is strictly positive and, by Assumption 1, strictly decreasing. The unique solution of (3), p^*, is characterized by $Z(p^*)B = A + B$ when $B > f(1)A$, and $p^* = 0$ when $B \leq f(1)A$.

A positive level of RI is efficient whenever the marginal benefit of learning, B, is greater than the marginal cost of diverting resources from development, which is the marginal change in the probability of developing the product, $f(1)$, times the uninformed entry payoff, A.

We now characterize the efficient solution as a function of the product's social value and expected harm.

Proposition 1. *The efficient outcome, $(p^*, r^*, (b_h^*)_{h\geq 0}, b_U^*, a_U^*)$, is characterized as follows:*

1. *When h is known, it is efficient to allow the product if and only if $h \leq S$, i.e., $b_h^* = 1(h \leq S)$.*
2. *When RI does not reveal h, it is efficient for the regulator to investigate the harm. The regulator learns h with probability r^* (see Lemma 1). If the regulator's investigation does not reveal h, it is efficient to allow the product if and only if $E[h] \leq S$, i.e., $b_U^* = 1(E[h] \leq S)$.*
3. *It is inefficient to invest in RI if*

$$\int_0^S (S-h)dG(h) - R^* < f(1)R^*. \tag{4}$$

Otherwise, the efficient investment in RI is characterized by

$$Z(p^*) = \frac{\int_0^S (S-h)dG(h)}{\int_0^S (S-h)dG(h) - R^*}. \tag{5}$$

Allocating resources to RI is not necessarily efficient. It depends on four key factors: the product's surplus (S), the distribution of harm (G), the regulator's continuation payoff from acquiring information (captured by R^*), and the marginal reduction in the probability of success due to diverting resources to RI (captured by $f(1)$).

For instance, when the regulator's cost of acquiring information is low, the continuation payoff from investigating the harm is large, in which case it may be efficient to leave the

burden of proof entirely to the regulator.[25] However, if acquiring information is costly for the regulator, it may be efficient to allocate resources to RI, provided that diverting resources from innovation does not significantly reduce the probability of the firm's success.

Although it is not always efficient for the firm to invest in RI, it is always efficient for the regulator to pursue an investigation when the firm does not provide evidence of harm, because an unconstrained regulator can always ban the product without evidence of harm (i.e., R^* is positive).

To further characterize the efficient solution, we examine how RI and the regulator's investigative resources vary with the product's social value.

Proposition 2. *We have:*

1. *The closer the product's social value is to the expected harm, the more resources the regulator uses to investigate. Specifically, $r^*(S)$ increases with S when $S < E[h]$ and decreases when $S > E[h]$. Furthermore, $\lim_{S\to 0} r^*(S) = \lim_{S\to\infty} r^*(S) = 0$.*
2. *As the product's social value increases, the firm's optimal investment in RI decreases. Specifically, $p^*(S)$ decreases with S and, furthermore, $\lim_{S\to\infty} p^*(S) = 0$.*

Figure 2 provides a numerical illustration of the results in Proposition 2. The left panel shows a non-monotonic relationship between the regulator's investment in learning the harm and social value S. When $S < E[h]$, the regulator investigates more as S rises because a ban based on prior information is more detrimental when the product's social value is higher. Conversely, when $S > E[h]$, the regulator investigates less as S increases because the product's higher social value can offset potential harm.

The right panel of Figure 2 shows that the efficient investment in responsible innovation decreases with S, approaching zero as S becomes large. Together, these results show that investments in learning the harm by the firm and the regulator are substitutes when the product's social value is low relative to the expected harm and complements otherwise.

To understand why, note that when $S < E[h]$, RI is the primary vehicle for assessing potential harm because the regulator's investigation will be minimal. As S increases within this range, the regulator's investigation gradually replaces RI. Intuitively, when the social value is relatively low, it is efficient for the firm to invest in RI, reducing its probability of success, because the cost of not having the product in the market is relatively small.

[25] Suppose that $c(r) = 0$ for all r. Then, it is efficient to learn h with probability 1, i.e., $r^* = 1$, which means that $R^* > 0$ and the left-hand side of (4) is zero. Hence, (4) holds.

However, when $S > E[h]$, the firm and the regulator invest less to learn the harm as S increases, so the investments act as complements. Intuitively, when the product's social value is high, it is inefficient for the firm to lower the probability of market success by engaging in RI to save on regulatory investigation costs. Instead, both the firm and the regulator reduce their efforts, acknowledging that the high social value of the product outweighs concerns about potential harm.

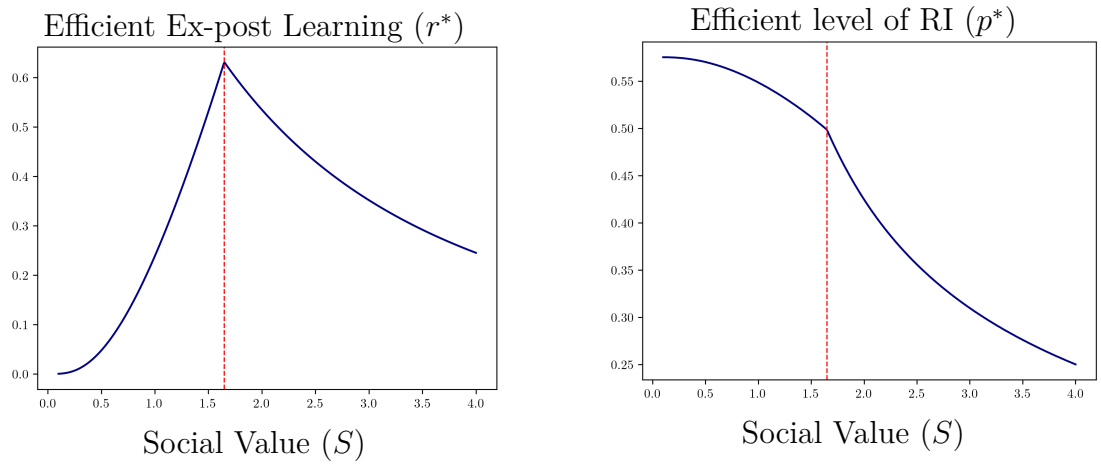

Figure 2: $r^*(S)$ and $p^*(S)$ when the harm distribution is $G = \text{lognormal}(0, 1)$, so that $E[h] = \exp(0.5) = 1.65$ (shown by the vertical dashed line); $F = \text{Beta}(2, 5)$, which satisfies Assumption 1; The regulator's cost of learning the harm with probability r is $c(r) = \frac{r^2}{2}$.

The efficient solution has implications for regulatory mechanisms used in practice. As noted before, efficient implementation never puts the burden of proof on the firm only (Lemma 1), that is, if the firm does not produce an informative signal about the product's harm, the product is allowed in the market while the regulator investigates the potential harm. An *ex-ante approval* regime where the firm has all the burden of proof is inefficient because it does not permit the regulator's investigation.[26] It is efficient to be cautious and ban the product if neither the firm nor the regulator uncovers evidence of harm only when the expected harm is large, that is, $E[h] > S$.

We now explore if and when the efficient solution can be implemented as a decentralized equilibrium in which the regulatory actions are delegated to a regulator who operates within a fixed regulatory framework that specifies fines, the ability to invoke the precautionary principle, and resources to investigate the harm.

[26]In the U.S., for example, the EPA may investigate the potential harm of a new product (e.g., new chemicals, energy technologies, or materials). The FDA typically does not conduct clinical trials but reviews data submitted by firms, so its investigation corresponds to how thoroughly it evaluates the evidence provided by manufacturers.

5 RI and Regulation in Equilibrium

In practice, no central authority controls both the regulator's and the firm's decisions. Regulators primarily advocate for consumers, whereas firms maximize profits.[27] We investigate whether it is possible to implement the efficient outcome, $(p^*, r^*, (b_h^*)_{h\geq 0}, b_U^*, a_U^*)$, as the equilibrium of a dynamic game where the firm chooses RI to maximize profits and, after the product is in the market, the regulator takes action within the constraints imposed by some regulatory regime. To do so, we first compute the equilibrium $(\hat{p}, \hat{r}, (\hat{b}_h)_{h\geq 0}, \hat{b}_U, \hat{a}_U)$ for a fixed regulatory regime $(L(\cdot), \phi, \bar{r})$ by solving the game by backward induction. Then, we study whether any regulatory regime can implement the efficient outcome.

Investigation in Equilibrium. The regulator allows the product after uncovering the harm if and only if $V + L(h) \geq h$, so $\hat{b}_h = 1(h \leq V + L(h))$. When h is unknown after the regulator's investigation, the regulator allows the product if and only if $V \geq E[h]$, so $\hat{b}_U = 1(E[h] \leq V)$.

When RI does not reveal h and the regulator decides to investigate, the regulator chooses the extent of his investigation by solving

$$\hat{R} \equiv \max_{r\in[0,\bar{r}]} \; r\int (V + L(h) - h)\hat{b}_h dG(h) + (1-r)(\phi\hat{b}_U + 1 - \phi)(V - E[h]) - c(r). \tag{6}$$

With probability r the regulator learns h, and allows the product according to $\hat{b}_h$, receiving $(V + L(h) - h)\hat{b}_h$. With probability $(1 - r)$, the regulator's investigation is uninformative. In that case, if $E[h] \leq V$, the product is allowed ($\hat{b}_U = 1$), and if $E[h] > V$, the regulator invokes the precautionary principle (i.e., $\hat{b}_U = 0$), and the product is allowed with probability $1 - \phi$. Lastly, the regulator pays $c(r)$ to learn the harm with probability r, subject to $r \leq \bar{r}$.

If RI does not reveal h, the regulator investigates whenever $\hat{R} \geq 0$, so $\hat{a}_U = 1(\hat{R} \geq 0)$.

Proposition 3. *If RI does not reveal h, the regulator investigates the harm if and only if*

$$(1 - \phi)(E[h] - V) \leq \rho c'(\hat{r}) - c(\rho), \tag{7}$$

where $\rho \equiv \min\{\hat{r}, \bar{r}\}$, and $\hat{r}$ is the probability that the regulator learn h from its investigation, which is given by the solution to

$$c'(\hat{r}) = \int_0^\infty (V + L(h) - h)\hat{b}_h dG(h) - (\phi\hat{b}_U + 1 - \phi)(V - E[h]). \tag{8}$$

[27] See Footnote 9.

Proposition 3 shows that the regulator's payoff from an investigation $\hat{R}$ may be negative in contrast to the efficient solution, where R^* is positive. In equilibrium, the regulator may prefer to (inefficiently) ban the product without proof of harm, invoking the precautionary principle *before* starting an investigation. Under limited liability, the regulator receives a lower payoff than the planner for allowing the product in the market.[28] Thus, the regulator is less willing to allow the product, i.e., $\hat{b}_h \leq b_h^*$ and $\hat{b}_U \leq b_U^*$, and also $\hat{R} \leq R^*$, so $\hat{a}_U \leq a_U^* = 1$. In particular, this happens when the expected harm is high ($V < E[h]$) and the regulator faces hurdles to ban a product without proof of harm ($\phi < 1$).

Thus, the precautionary principle can be a double-edged sword. On the one hand, it can help stop potentially harmful products when their harm remains uncertain after an investigation. On the other hand, it allows the regulator to (inefficiently) remove potentially safe products early on, without the need to investigate before making this decision.[29]

Proposition 4. *The regulator's equilibrium strategy is characterized as follows:*

1. *If h is known, the regulator allows the product if and only if $h \leq V + L(h)$, i.e., $\hat{b}_h = 1(h \leq V + L(h))$.*

2. *If RI does not reveal h:*

   (a) *If $\hat{R} < 0$, the regulator invokes the precautionary principle, i.e., $\hat{a}_U = 0$, so it bans the product with probability ϕ, and investigates with probability $1 - \phi$.*

   (b) *If $\hat{R} \geq 0$, the regulator does not invoke the precautionary principle, i.e., $\hat{a}_U = 1$, and investigates, learning h with probability ρ.*

3. *If the regulator's investigation does not reveal h, the product is banned with probability $\phi\hat{b}_U + 1 - \phi$, where $\hat{b}_U = 1(E[h] \geq V)$.*

A comparison of the results in Propositions 1 and 4 highlights the misalignment between the regulator and the planner's preferences. Generally, the regulator is more cautious than the planner because his payoff is lower than the social surplus. It is noteworthy to mention that a regulatory regime imposing maximum penalties *may not* lead to efficient outcomes. Setting $L(h) = \pi$ aligns the regulator's and planner's preferences when h is known. Without proof

[28] When h is known and the product is allowed, it is immediate to see that the regulator's payoff is lower than the planner's payoff because $V + L(h) - h \leq S - h$. Similarly, when h is unknown and the product is allowed, the regulator gets $V - E[h] \leq S - E[h]$.

[29] If the regulator chooses to ban the product without investigating ($\hat{a}_U = 0$), the ban succeeds with probability ϕ. If the regulator fails to ban the product, sequential rationality forces him to investigate. Thus, the regulator investigates with probability $\hat{a}_U\phi + 1 - \phi$.

of harm, however, the regulator cannot impose a fine, which creates misalignment with the planner's when $E[h] \in (V, S)$. Moreover, in this case, the regulator may be unable to ban the product from the market if $E[h] > S$, whereas the planner can always ban it. Finally, a regulatory regime that sets maximum fines introduces other distortions, as discussed later.

RI in Equilibrium. We now consider the firm's problem of choosing the resources allocated to RI, $\hat{p}$. We denote the firm's expected profits when the harm is known by

$$\hat{\pi} \equiv \int (\pi - L(h))\hat{b}_h dG(h). \tag{9}$$

The firm's expected profits from entering the market when harm is unknown are

$$\hat{A} \equiv (\hat{a}_U \phi + 1 - \phi) \left[\rho\hat{\pi} + (1-\rho)\left(\phi\hat{b}_U + 1 - \phi\right)\pi\right]. \tag{10}$$

In this case, the regulator investigates the harm with probability $\phi\hat{a}_U + 1 - \phi$, learning h with probability ρ. Therefore, the firm solves

$$\max_{p\in[0,1]} F(1-p)[p\hat{\pi} + (1-p)\hat{A}]. \tag{11}$$

In the expression above, $\hat{A}$ is the "uninformed entry payoff" and $\hat{\pi} - \hat{A}$ is the "information premium". We use Lemma 2 to characterize the firm's optimal investment in RI.

Proposition 5. *The firm does not invest in RI, i.e., $\hat{p} = 0$, when*

$$\hat{\pi} - \hat{A} \leq f(1)\hat{A}, \tag{12}$$

Otherwise, the firm's optimal investment in RI, $\hat{p}$, satisfies

$$Z(\hat{p}) = \frac{\hat{\pi}}{\hat{\pi} - \hat{A}}.$$

The firm invests in RI as long as the marginal benefit of reallocating resources from innovation to learning h (captured by the term $\hat{\pi} - \hat{A}$), exceeds the marginal cost, which is the loss of the continuation payoff from not innovating, $f(1)\hat{A}$.

Consider cases where the firm expects to receive zero when RI does not reveal the harm (i.e., $\hat{A} = 0$). This can occur if, for example, fines are maximal, the expected harm is large, and the regulator has the power to ban the product without proof of harm. That is, $\hat{\pi} = 0$,

$\hat{b}_U = 0$, and $\phi = 1$. Such cases essentially correspond to an *ex-ante* approval regime because the firm receives an expected payoff of zero if RI does reveal h. In this situation, the firm's investment in RI is inefficiently large and implicitly determined by

$$\hat{p} = \frac{F(1-\hat{p})}{f(1-\hat{p})}.$$

When $\hat{A} > 0$, the firm's decision to invest in RI depends on the size of the fines. Specifically, the firm would never invest in RI if the information premium, $\hat{\pi} - \hat{A}$ is too low.[30] The firm invests in RI when the information premium is large relative to the reduction in the probability of innovation (see inequality 12).

6 Equilibrium and Efficiency

We now ask whether a combination of regulatory tools $(L(\cdot), \phi, \bar{r})$ can implement the efficient outcome.

Definition 1. *Consider the efficient outcome* $(p^*, r^*, (b_h^*)_{h\geq 0}, b_U^*, a_U^*)$.

(a) We say that the efficient regulatory decisions are implementable by the regulator *if there exists a regulatory regime,* $(L(\cdot), \phi, \bar{r})$*, such that:*

(i) $\hat{b}_h = b_h^*$*, for all* $h \geq 0$.

(ii) $\phi\hat{b}_U + 1 - \phi = b_U^*$,

(iii) $\phi\hat{a}_U + 1 - \phi = a_U^*$,

(iv) If the second stage is reached with positive probability, $\rho = r^*$.

(b) We say that the efficient level of RI is implementable by the firm *if there exists a regulatory regime,* $(L(\cdot), \phi, \bar{r})$*, such that* $\hat{p} = p^*$.

Proposition 6 shows whether and how the efficient regulatory decisions are implementable: First, with *small harm* $(E[h] < V)$, the regulator and the planner's incentives are aligned, thus efficiency is guaranteed for any regulatory regime.

Second, with *moderate harm* $(V \leq E[h] \leq S)$, an uninformed regulator is more cautious than an uninformed planner when allowing the product because he underestimates its social value.

[30] To see this, note that inequality (12) does not hold if $\hat{\pi} - \hat{A}$ is low.

To align incentives, a regulatory regime must impose increasingly higher fines, forbidding the precautionary principle, and limiting the regulatory resources to investigate the harm.

Third, with *large harm* ($S < E[h]$), a regulatory regime must allow the precautionary principle with certainty, otherwise after conducting his investigation, an uninformed regulator may be unable to ban the product. Moreover, the regulatory regime must prescribe maximal penalties to make the regulator fully internalize the firm's profits, which motivates the regulator to conduct an efficient investigation; otherwise, the regulator investigates too little.

Proposition 6. *The efficient regulatory decisions are* implementable by the regulator *under the following regulatory regimes:*

1. *Fines: Are irrelevant for small harm, larger than $h - V$ for moderate harm, and must be maximal for large harm.*[31]
2. *Precautionary Principle: Irrelevant for small harm, must be forbidden for moderate expected harm, and must be allowed with certainty for large harm.*[32]
3. *Investigation Resources: Must be limited to $\bar{r} = r^*$.*

For a given regulatory regime $(L(\cdot), \phi, \bar{r})$, the regulator makes efficient decisions (i.e., $\hat{b}_h = b_h^*$) when the harm is known and sufficiently small ($h < V$) or large ($h > S$).[33] Inefficient decisions can occur only when $h \in [V, S]$. In this case, a 'sizable' fine makes the regulator internalize a sufficient amount of the firm's profits, leading to an efficient decision. Figure 3 illustrates these conditions, showing that multiple fine schedules can implement b_h^*.

[31] Specifically, when $E[h] > S$, $L(h) = \pi$ for $h \in [0, S]$ and $L(h) = [0, \pi]$ for $h > S$; When $E[h] \leq S$, $L(h) \geq h - V$ for $h \in [V, S]$ and $L(h) = [0, \pi]$ for $h \notin [V, S]$.

[32] Specifically, $\phi \in [0, 1]$ when $E[h] \leq V$ and $\phi = 0$ when $V < E[h] < S$ and $\phi = 1$ when $S < E[h]$.

[33] Note that if $h > S$, then $h \geq L(h) + V$ because limited liability imposes $L(h) \leq \pi$, and the optimal regulator decision is efficient. Similarly, if $h < V$, then $h \leq L(h) + V$ because $L(h) \geq 0$.

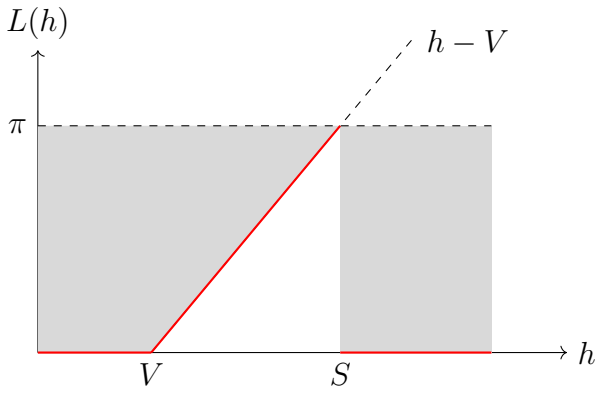

Figure 3: Any function $L(h)$ within the gray area implements $\hat{b}_h = b_h^*$. The red line indicates the minimum fines that implement b_h^*.

Consider now the case of unknown harm. The planner and the regulator allow the product when $E[h] < V$, making the precautionary principle an irrelevant tool for implementing the efficient decision b_U^*.[34] When $S < E[h]$, the planner bans the product and the regulator *tries to* ban it too. However, the regulator may be unable to do so because the use of the precautionary principle can be uncertain.[35] Therefore, to guarantee that the product is banned, it is necessary to allow the use of the precautionary principle with certainty, i.e., $\phi = 1$. Lastly, when $V < E[h] < S$, the regulator tries to ban the product, whereas the planner allows it. Therefore, it is necessary to "tie the hands" of the regulator, forbidding the use of the precautionary principle, i.e., $\phi = 0$.

Next, if RI does not reveal h, the planner always investigates ($a_U^* = 1$), whereas the regulator may prefer to immediately ban the product and avoid the investigation (Proposition 3). By forbidding the regulator from banning the product without first conducting an investigation, i.e., setting $\phi = 0$, it is possible to implement the efficient decision $a_U^* = 1$.[36]

Next, we compare the regulator's learning probability, ρ, with the efficient one, r^*, for any regulatory regime that implements b_U^* and b_h^*. Suppose for simplicity that the regulator is endowed with enough resources so that $\min\{\hat{r}, \bar{r}\} = \hat{r}$. If $E[h] \leq S$ then $b_U^* = 1$ and $\hat{r}$ satisfies

$$c'(\hat{r}) = c'(r^*) + \int_0^S L(h)dG(h) + \pi(1 - G(S)).$$

From the convexity of $c(\cdot)$, we obtain $r^* \leq \hat{r}$.[37] Therefore, if the regulator is endowed with

[34] To see that any $\phi \in [0,1]$ implements b_U^* when $b_U^* = \hat{b}_U = 1$, note that for any $\phi \in [0,1]$, $\phi\hat{b}_U + 1 - \phi = 1$.
[35] When $\hat{b}_U = 0$, the product remains in the market with probability $1 - \phi$.
[36] That is, when $\hat{a}_U$=0, $\phi\hat{a}_U + 1 - \phi = 1 - \phi$, which equals $a_U^* = 1$ when $\phi = 0$.
[37] Note that $L(h) \geq 0$ and $\pi(1 - G(S)) \geq 0$, so $c'(\hat{r}) \geq c'(r^*)$, which implies $r^* \leq \hat{r}$ because c' is increasing.

a large amount of resources, he investigates *too much* in equilibrium relative to the efficient level. This occurs because the regulator cares about receiving $L(h)$ when the product is allowed, and ignores the loss in profits from learning $h > S$ and (efficiently) banning the product, which is captured by the term $\pi(1-G(S))$. Therefore, to implement r^*, a regulatory regime must limit the regulator's investigative resources to $\hat{r} \leq \bar{r} \equiv r^*$.

Lastly, if $S < E[h]$, then $b_U^* = 0$ and we obtain

$$c'(\hat{r}) = c'(r^*) - \int_0^S (\pi - L(h))dG(h).$$

Setting $L(h) = \pi$ for $h \leq S$ implements the first best level of learning as long as the regulator is given enough resources, i.e., $r^* \leq \bar{r}$. The reason is that without learning h the regulator and the planner ban the product. However, the regulator does not internalize the expected benefit for the firm from learning that $h \leq S$, which is $\int_0^S (\pi - L(h))dG(h)$. Thus, the regulator learns too little compared to the efficient level. To provide the proper investigation incentives, a regulatory regime must set the maximum possible fine, $L(h) = \pi$.

Next, we ask whether the firm can implement the efficient level of RI, for any regulatory regime that induces the regulator to implement the efficient regulatory actions. It turns out that whenever the regulator's actions are efficient, the firm does not invest in RI.

Proposition 7. *Consider a regulatory regime such that efficient regulatory decisions are implementable by the regulator. Then, the firm does not invest in RI.*

With small harm, i.e., $E[h] \leq S$, the regulator *does not* ban the product when uncertain about h. Consequently, the firm benefits from not providing proof of harm, so it does not invest in RI, shifting the burden of proof to the regulator.

However, when $E[h] > S$, the efficient regulatory action is to remove the product from the market when the harm is unknown. Thus, in principle, the firm could have incentives to invest in RI to avoid getting its product banned if the regulator does not learn h. However, to ensure the regulator's investigation is efficient, fines must be maximal when $h \in [0, S]$, in which case the firm does not benefit from investigating the harm. In either case, the product is banned or allowed after the regulator extracts all firm's profit.[38]

This result shows the inherent tension between regulatory efficiency and RI. Any regulatory regime that induces the regulator to make efficient decisions deters the firm from investing in RI. Importantly, in contrast to previous papers in the literature, large fines are ineffective at

[38] If indifferent, suppose the firm pays a tiny cost from investing in RI.

aligning the firm's and the planner's incentives. In fact, in our setting, fines may *reduce* RI, since there is no benefit of revealing h to an efficient regulator. The following result follows from Proposition 7.

Corollary 1. *If it is efficient to place the burden of proof on the regulator, i.e. $p^* = 0$, then the equilibrium is efficient under the regulatory regime in Proposition 6.*

Suppose that $p^* > 0$. Given that we cannot implement both the firm's and the regulator's efficient decisions, we investigate whether there exists a regulatory regime that distorts the regulator's decisions in order to implement the firm's optimal investment in RI.

Proposition 8. *When the expected harm is large ($E[h] > V$) it is possible to incentivize the firm to efficiently invest in RI by (inefficiently) limiting the regulator's resources and allowing the precautionary principle. When the harm is small ($E[h] \leq V$), no regulatory regime induces the firm to invest in RI.*

When the expected harm is large, the regulator adopts a cautious approach and bans the product without proof of harm. By limiting the regulator's resources below the efficient level, the firm has an incentive to demonstrate the harm in order to avoid the likely ban. In contrast, when the expected harm is small, the regulator allows the product without proof of harm and penalizes the firm only when there is evidence of the harm, which the firm has no incentives to provide.

Therefore, if the efficient solution involves RI ($p^* > 0$), there are regulatory regimes that can implement either the efficient regulatory decisions or the firm's efficient investment in RI but not both at the same time. Given these results, it is natural to ask: Which regulatory regimes best balance distortions on both the firm and regulator's decisions?

In Appendix B, we introduce the notion of constrained-optimal regulatory regimes, which are regulatory regimes that maximize the planner's payoff balancing out the distortions to both the regulator and the firm.

As an illustration, we numerically find the constrained-optimal regulatory regimes for a parametric example. Our results show that regulatory distortions manifest in various forms. When consumer surplus is large relative to firms' profits, the regulator's and planner's preferences tend to align. In such cases, the optimal regulatory framework permits a "heavy-handed" approach, allowing the regulator to use the precautionary principle with certainty. However, as profits grow relative to consumer surplus, the regulator's and the planner's pref-

erences become increasingly misaligned because the regulator's consumer surplus bias grows. As a consequence, the optimal regulatory regime forbids the precautionary principle.

Figure 4 compares the equilibrium values of $r^*(S)$ and $p^*(S)$ (based on the same parametric assumptions as in Figure 2) with the equilibrium values $\hat{r}(S)$ and $\hat{p}(S)$ under the constrained-optimal regulatory regime, which sets $\phi = 1$ for $S \leq 2.2$ and $\phi = 0$ for $S > 2.2$. The figure shows that as long as $\phi = 1$, the firm invests in RI. The reason is that the regulator bans the product if h is unknown after his investigation. This makes it too risky for the firm to leave the burden of investigation to the regulator, which motivates it to invest in RI. Moreover, fines are lower than what is required to achieve efficient learning by the regulator. However, the constrained-optimal regulatory regime provides insufficient incentives to invest in RI because that would imply a large deviation from the efficient regulatory decisions. When S is sufficiently large and $V > E[h]$, the constrained-optimal regulatory regime perfectly aligns the regulator's and planner's incentives (on the left panel of Figure 2, the two curves overlap for $S > 2.2$.). However, as implied by Proposition 7, the firm does not invest in RI (on the right panel of Figure 2, $p^* = 0$ for $S > 2.2$.).

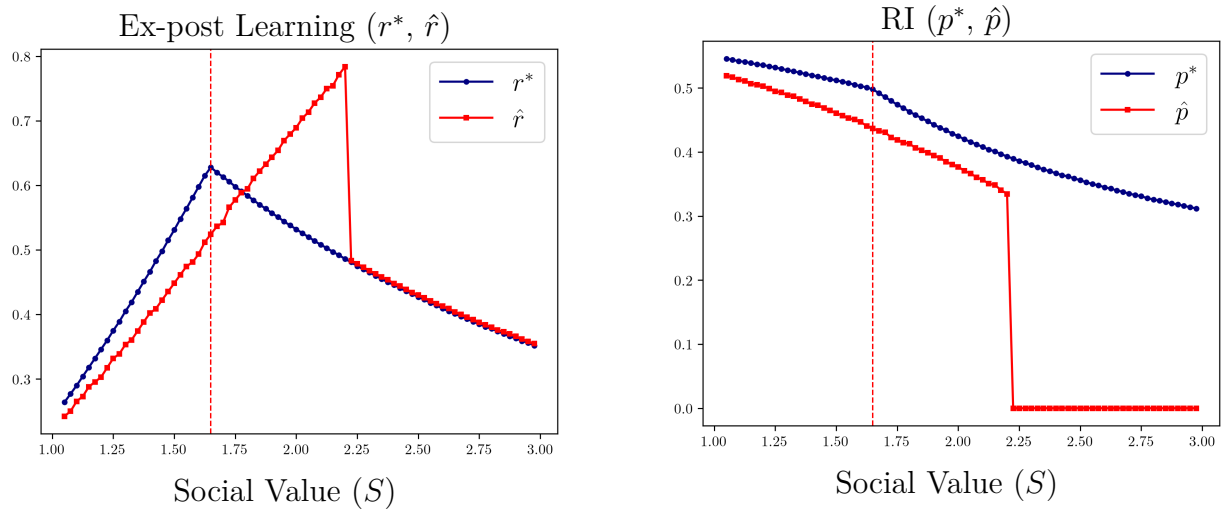

Figure 4: Comparison of $r^*(S)$ and $\hat{r}(S)$ (left panel) and $p^*(S)$ and $\hat{p}(S)$ (right panel) for the constrained-optimal regulatory regime for the harm distribution is $G = \text{lognormal}(0, 1)$, so that $E[h] = \exp(0.5)$ (shown by the vertical dashed line); $F = \text{Beta}(2, 5)$, which satisfies Assumption 1; The regulator's cost of learning the harm with probability r is $c(r) = \frac{r^2}{2}$. Consumer surplus is fixed at $V = 1$.

An important takeaway from our analysis is that harsh fines are not always desirable. In fact, it is important to assess the relative values of consumer surplus and profits. When most of the social surplus comes from consumer surplus, the optimal regulatory regime sets low fines,

limits the regulator's resources, and allows the use of the precautionary principle. As profits increase, both the size of the fines and the regulator's investigative budget increase (even beyond the efficient level). Consequently, the firm adjusts its RI investments in anticipation of regulatory actions, investing in RI because it fears the regulator may invoke the precautionary principle too liberally and ban the product. As profits are the primary component of social surplus, then the regulatory regime focuses on making sure that the regulator takes efficient actions, leaving the burden of proof to the regulator.

Discussion and Extensions

Our model provides a tractable framework to analyze the trade-off firms face between pursuing innovation and investing in learning about potential harm, in the presence of a regulator who can subsequently also learn about a product's potential harm and impose sanctions. Several extensions could incorporate additional features, some of which would generate similar insights, while others can provide new ones at the cost of higher complexity and untractability.

Our setting could be extended in different directions. For example, political and institutional pressures can influence regulators' decisions, which we partially capture through the uncertain application of the precautionary principle, limited liability, and budget constraints. An extension of our model could incorporate lobbying efforts or political influence by, for example, redifining the regulator's preference as $V + x - h + L(h)$, where x could help align the planner and the regulator's incentives, e.g., when x is a transfer from the firm to the regulator.

We assume that information revelation is binary. In reality, information acquisition is often gradual and probabilistic; firms and regulators operate under varying degrees of uncertainty, receiving partial signals and continuously updating their beliefs. Incorporating dynamic information arrival or acquisition, or partially informative signals could enrich the model. Even in a more involved setting, our main insights would likely remain unchanged. For instance, if RI yields informative but imperfect signals, a Bayesian regulator might employ threshold strategies, taking regulatory action when evidence of harm exceeds certain levels, which our binary framework already captures.

Our framework assumes that information is public. An interesting extension would allow the firm to strategically disclose information. First, if the firm does not disclose information, the regulator becomes more cautious, therefore more likely to learn or invoke the precautionary

principle. From the perspective of the firm, on the one hand the regulator's more conservative strategy would lower the incentives to learn. On the other hand, it provides the firm incentives to learn and disclose information in order to avoid the precautionary principle. Given the ambiguous overall effect, this trade-off can lead to complicated equilibrium dynamics.

Another possible extension would involve incorporating more nuances on payoffs after the regulator allows the product without proof of harm. The harm could be revealed ex post with some exogenous probability at no cost. Also, without proof of harm the regulator could allow the product and impose a fine. Although deriving the precise implications for the regulatory regime is more complex, the driving economic forces would be similar to those in the benchmark model.

Lastly, a more general framework could allow for profits and consumer surplus to be a function of the harm, i.e., $\pi(h)$, $V(h)$, and learning can be a function of both investments and harm (e.g., larger harm could be easier or harder to discover). For example, if higher perceived harm reduces both profits and consumer surplus, firms might have stronger incentives to invest in RI to mitigate negative market perceptions. Regulators might also adjust their decisions, as the social value of allowing the product diminishes when accounting for reduced consumer surplus.

7 Conclusion

We offer a novel framework that explores the strategic interaction between regulatory oversight and responsible innovation. Our model captures the trade-off a firm faces between investing in innovation and investigating potential harms, while the regulator acts after the product is in the market, often constrained by available resources and regulatory tools. We identify conditions under which efficient enforcement can be delegated to a regulator advocating for consumers.

Our analysis shows that the effectiveness of delegation hinges on the expected severity of harm. When expected harm is either very low or very high, regulatory regimes that grant the regulator greater discretion yield efficient regulatory outcomes. In contrast, moderate expected harm requires more restrictive measures. When it is optimal for the firm to focus on innovation — leaving the investigation of potential harms to the regulator — appropriate regulatory constraints achieve efficiency. Otherwise, regulatory regimes ensuring efficient ex-post regulation undermine the firm's incentive to invest in responsible innovation. This

implies that a welfare-maximizing regulatory regime must sometimes deliberately distort ex-post decisions to encourage early investigation of potential harms by the firm.

Our results suggest that a one-size-fits-all approach to regulation is insufficient. Efficient regulation calls for increased regulatory empowerment as expected harm rises, while firm incentives for responsible innovation must be preserved through calibrated interventions. Ultimately, the optimal regulatory regime is context-dependent, requiring careful calibration of fines, precautionary measures, and investigative resources to achieve the desired outcomes for both firms and regulators.

We contribute to the literature on regulatory design for innovation by emphasizing the strategic interplay between firms and regulators, showing that optimal regulatory regimes must balance economic efficiency with incentives for responsible innovation. Future work could extend our model by incorporating dynamic information revelation and institutional factors to further elucidate the balance between innovation, responsibility, and regulation.

8 References

Acemoglu, Daron and Todd Lensman (2024) "Regulating transformative technologies," *American Economic Review: Insights*, Vol. 6, pp. 359–376.

Ahmed, Nur, Amit Das, Kirsten Martin, and Kawshik Banerjee (2024) "The Narrow Depth and Breadth of Corporate Responsible AI Research," *arXiv preprint arXiv:2405.12193*.

Baumann, Florian and Tim Friehe (2016) "Learning-by-doing in torts: Liability and information about accident technology," *Economics Letters*, Vol. 138, pp. 1–4.

De Chiara, Alessandro and Ester Manna (2022) "Corruption, regulation, and investment incentives," *European Economic Review*, Vol. 142, p. 104009.

Friehe, Tim and Elisabeth Schulte (2017) "Uncertain product risk, information acquisition, and product liability," *Economics Letters*, Vol. 159, pp. 92–95.

Gans, Joshua S (2024) "Regulating the Direction of Innovation,"Technical report, National Bureau of Economic Research.

Henry, Emeric, Marco Loseto, and Marco Ottaviani (2022) "Regulation with experimentation: Ex ante approval, ex post withdrawal, and liability," *Management Science*, Vol. 68, pp. 5330–5347.

Immordino, Giovanni, Marco Pagano, and Michele Polo (2011) "Incentives to innovate and social harm: Laissez-faire, authorization or penalties?" *Journal of Public Economics*, Vol.

95, pp. 864–876.

Kayid, M, H Al-Nahawati, and IA Ahmad (2011) "Testing behavior of the reversed hazard rate," *Applied mathematical modelling*, Vol. 35, pp. 2508–2515.

Lewis, Tracy R and David EM Sappington (1994) "Supplying information to facilitate price discrimination," *International Economic Review*, pp. 309–327.

Ottaviani, Marco and Peter Norman Sørensen (2006) "Reputational cheap talk," *The Rand journal of economics*, Vol. 37, pp. 155–175.

Ottaviani, Marco and Abraham L Wickelgren (2023) "Approval regulation and learning, with application to timing of merger control," *The Journal of Law, Economics, and Organization*, p. ewac025.

Requate, Till, Tim Friehe, and Aditi Sengupta (2023) "Liability and the incentive to improve information about risk when injurers may be judgment-proof," *International Review of Law and Economics*, Vol. 76, p. 106168.

Shavell, Steven (1992) "Liability and the incentive to obtain information about risk," *The Journal of Legal Studies*, Vol. 21, pp. 259–270.

Weil, Gabriel (2024) "Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence," *Available at SSRN 4694006*.

A Proofs

Proof of Lemma 1

Proof. The characterization of r^* is direct from the first-order conditions. The continuation payoff R^* is weakly positive since $r = 0$ yields a positive payoff. □

Proof of Lemma 2

Proof. The derivative of the objective function in this problem is

$$\left[\frac{F(1-p)}{f(1-p)} + (1-p)\right] B - (A+B) \Leftrightarrow G(p)B - (A+B).$$

First, if $Z(0)B \leq (A+B)$, there are two cases. If $B < 0$, $Z(p)B < 0 \leq A+B$ for all p, so $p^* = 0$. If $B > 0$, $Z(p)B < Z(0)B \leq A+B$, so $p^* = 0$. Second, if $Z(0)B > A+B$, then since $Z(1) = 0$, by the intermediate value theorem, there must exist a solution to $Z(p^*)B = A+B$, and since $Z(\cdot)$ is strictly decreasing the solution is unique. □

Proof of Proposition 1

Proof. 1) is direct from optimal decision $b_h^* = 1(h \leq S)$, 2) is direct from Lemma 1. 3) Noting that the information premium ("B") equals $\int_0^S (S-h)dG(h) - R^*$, and the uninformed entry payoff ("A") is R^*, condition $B < f(1)A$ can be rewritten as

$$\int_0^S (S-h)dG(h) - R^* < f(1)R^*. \tag{13}$$

Therefore by Lemma 2 we obtain the result. □

Proof of Proposition 2

Proof. We divide the proof in several steps:

1. We first show that r^* decreases when $E[h] \leq S$. We have

$$c'(r^*) = \int_S^\infty (h - S) dG(h).$$

The right-hand side is decreasing in S when $G(S) < 1$ because its derivative is $-(1 - G(S))$. Moreover, $\int_S^\infty (h - S) dG(h) \leq \int_S^\infty h = E[h] - \int_0^S h dG(h) \xrightarrow{S\to\infty} 0$, we get $r^*(S) \xrightarrow{S\to\infty} 0$.

2. We now show that r^* increases when $S < E[h]$. We have

$$c'(r^*) = \int_0^S (S - h) dG(h).$$

The right-hand side is increasing in S when $G(S) > 0$ because its derivative is $G(S)$. Moreover, $\int_0^S (S - h) dG(h) \leq SG(S) \xrightarrow{S\to 0} 0$, so $r^*(S) \xrightarrow{S\to 0} 0$.

3. Next, we show that p^* decreases with S.

(1) Let $S > E[h]$. The denominator of the RHS of (5) can be written as $(1-r^*)c'(r^*)+c(r^*)$. Taking derivative with respect of S of the RHS of (5) we obtain

$$\frac{G(S)[(1 - r^*)c'(r^*) + c(r^*)] - \left(\int_0^S (S - h) dG(h)\right)(1 - r^*)c''(r^*)\frac{dr^*}{dS}}{((1 - r^*)c'(r^*) + c(r^*))^2}.$$

Since $\frac{dr^*}{dS} < 0$ for $S > E[h]$, we get $\frac{d}{dS}[Z(p^*)] > 0$. Since $Z' < 0$, we get $Z'(p^*)\frac{dp^*}{dS} > 0$, which implies $\frac{dp^*}{dS} < 0$.

(2) Let $S < E[h]$. Then, $R^* \equiv \max_{r\in[0,1]}\ r \int_0^S (S - h) dG(h) - c(r)$, so $\frac{dR^*}{dS} = r^* G(S)$. Taking derivative of (5):

$$\frac{d}{dS}\left(\frac{\int_0^S (S - h) dG(h)}{\int_0^S (S - h) dG(h) - R^*}\right) = \frac{G(S)\left(\int_0^S (S - h) dG(h) - R^*\right) - (1 - r^*)G(S) \int_0^S (S - h) dG(h)}{\left(\int_0^S (S - h) dG(h) - R^*\right)^2}$$

which simplifies to

$$\frac{d}{dS}\left(\frac{\int_0^S (S - h) dG(h)}{\int_0^S (S - h) dG(h) - R^*}\right) = \frac{G(S)\left(r^* \int_0^S (S - h) dG(h) - R^*\right)}{\left(\int_0^S (S - h) dG(h) - R^*\right)^2}$$

Using the definition of R^*, the numerator equals $c(r^*)$, which is positive.

Therefore, $\frac{d}{dS}[Z(p^*)] > 0$. Since $Z' < 0$, we get $Z'(p^*)\frac{dp^*}{dS} > 0$, which implies $\frac{dp^*}{dS} < 0$.

4. Finally, $p^* = 0$ when $S \to \infty$. When $S \to \infty$, we have that $r^* = 0$, and $R^* =$

$(S - E[h])b_U^*$. In this case inequality (4) holds since the left-hand side is zero and the right-hand side positive ($f(1)(S - E[h])b_U^*$).

□

Proof of Proposition 3

Proof. Direct from taking first-order condition in (6) (see also footnote 21).

Second, if the regulator can invoke the precautionary principle with certainty, i.e., $\phi = 1$, then the regulator's continuation value $\hat{R}$ is positive by optimality. It is also positive when $\phi < 1$ and $E[h] \leq V$, since by optimality $b_U^* = 1$. When $\phi < 1$ and $V < E[h]$, we have that $b_U^* = 0$. Therefore, the continuation value is positive if and only if

$$(1 - \phi)(E[h] - V) \leq \rho c'(\rho) - c(\rho),$$

where $\rho = \min\{\hat{r}, \bar{r}\}$. □

Proof of Proposition 7

Proof. First, if the regulator's actions are efficient, we have $\hat{A} = r^*\hat{\pi} + (1 - r^*)b_U^*\pi$. Then, the firm problem becomes

$$\max_{p \in [0,1]} F(1 - p)\{\hat{A} + p(\hat{\pi} - b_U^*\pi)(1 - r^*)\}.$$

Here we have two cases. First, if $b_U^* = 1$, since $\hat{\pi} \leq \pi$, the objective function is decreasing in p, so the optimal solution is $\hat{p} = 0$. Second, if $b_U^* = 0$, then to implement efficient learning by the regulator it is necessary to set fines such that $\hat{\pi} = 0$. In that case, the objective function is zero for all $\hat{p}$, so any $\hat{p}$ delivers a payoff of zero. □

Proof of Proposition 8

Proof. We will show that, when $E[h] > V$, it is possible to incentivize the firm to efficiently invest in RI by limiting the regulator's resources. Whenever $p^* > 0$, to obtain $\hat{p} = p^*$, we

need two conditions:

$$\frac{\hat{\pi}}{\hat{A}} > 1 + f(1) \text{ and } \frac{\hat{\pi}}{\hat{A}} = \frac{\int_0^S (S-h)dG(h)}{R^*}.$$

Imposing that $p^* > 0$ implies that $\frac{\int_0^S (S-h)dG(h)}{R^*} > 1 + f(1)$. Therefore, a necessary and sufficient condition for $\hat{p} = p^*$ is

$$\frac{\hat{\pi}}{\hat{A}} = \frac{\int_0^S (S-h)dG(h)}{R^*}. \tag{14}$$

We have $\hat{b}_U = 0$ because $E[h] > V$. Moreover, when $\phi = 1$, $\hat{A} = \rho\hat{\pi}$. Therefore, the left-hand side of the equality (14) equals $1/\rho$, so we need

$$\rho = \frac{R^*}{\int_0^S (S-h)dG(h)}.$$

Given that $\rho = \min\{\bar{r}, \hat{r}\}$, a necessary and sufficient condition for this equality to hold is

$$\hat{r} = [c']^{-1}\left(\int (V + L(h) - h)\hat{b}_h dG(h)\right) \geq \frac{R^*}{\int_0^S (S-h)dG(h)}.$$

It is easy to see that $\hat{r}$ is largest with maximal fines, that is, $L(h) = \pi$. In that case, the inequality above becomes:

$$[c']^{-1}\left(\int_0^S (S-h)dG(h)\right) \geq \frac{R^*}{\int_0^S (S-h)dG(h)}$$

Using the expression of r^* in the efficient solution, the inequality above is equivalent to

$$r^* \geq \frac{r^* \int_0^S (S-h)dG(h) - c(r^*)}{\int_0^S (S-h)dG(h)}$$

which always holds with strict inequality when $r^* > 0$. Thus, whenever $E[h] > V$, $L(h) = \pi$, and $\phi = 1$, we have $\hat{r} > \frac{R^*}{\int_0^S (S-h)dG(h)}$. Therefore, we can choose $L(h) < \pi$ and the inequality still hold. It is important that $L(h) < \pi$ to avoid indifference by the firm. Imposing $\bar{r} = \frac{R^*}{\int_0^S (S-h)dG(h)}$ we achieve $\hat{p} = p^*$.

Next, suppose that the expected harm is small, so $E[h] \leq V$. In this case, we have $\hat{b}_U = 1$. whenever $\hat{a}_U = 1$, equality (14) does not hold. To see this, note that it would imply $\hat{A} = \rho\hat{\pi} + (1-\rho)\pi$. Then, $\frac{\hat{\pi}}{\hat{A}} = \frac{\hat{\pi}}{\rho\hat{\pi}+(1-\rho)\pi}$ which is a number less than one because the denominator is larger than the numerator. Since the right-hand side of equation (14) is larger than $1+f(1)$,

it is impossible to implement p^* when $\hat{a}_U = 1$. Finally, note that from Proposition 3, $\hat{a}_U = 1$ because $E[h] < V$. □

B Constrained-Optimal Regulatory Regimes

Generally, regulatory regimes may distort the firm's RI choice and the regulator's decisions. In this section, we introduce the notion of constrained-optimal regulatory regimes, which are regulatory regimes that maximize the planner's payoff, when the planner cannot control the decisions of the firm and the regulator.

In addition to our baseline model, we introduce an additional parameter, K, that captures the firm's setup cost or opportunity cost, and it is use to avoid equilibria multiplicity from indifference. This parameter triggers an additional decision by the firm: to innovate or use its resources elsewhere. The firm decides to innovate if and only if

$$F(1-\hat{p})[\hat{p}\hat{\pi} + (1-\hat{p})\hat{A}] \geq K, \tag{15}$$

where $\hat{p}$ is the equilibrium RI, $\hat{\pi}$ and $\hat{A}$ are defined in the previous section (equations 9 and 10), all of which depend on the regulatory regime. Moreover, we assume that it is efficient for the firm to innovate.

Formally, given V, π, a regulator's cost function $c(\cdot)$, a distribution of harm $G(\cdot)$, a function (distribution) $F(\cdot)$, and K, a *constrained-optimal regulatory regime* $(L^O(\cdot), \phi^O, \bar{r}^O)$ solves the following problem:

$$\max_{\{L(\cdot),\phi,\bar{r}\}} F(1-\hat{p}) \left\{ \hat{p} \int (S-h)\hat{b}_h dG(h) + (1-\hat{p})(\hat{a}_U\phi + 1 - \phi)R_P \right\}$$

subject to $\phi \in [0,1]$, $L(h) \in [0,\pi]$, $\bar{r} \in [0,\infty)$, and (15), where

$$R_P = \hat{r} \int (S-h)\hat{b}_h dG(h) + (1-\hat{r})(\hat{b}_U\phi + 1 - \phi)(S - E[h]) - c(\hat{r}),$$

$(\hat{r}, (\hat{b}_h)_{h\geq 0}, \hat{b}_U, \hat{a}_U)$ is the regulator's strategy (Proposition 4), and $\hat{p}$ is the firm's investment in RI (Proposition 5).

Finding the constrained-optimal regulatory regime requires solving a complex optimization problem with multiple constraints. In some cases, it is possible to characterize the solution analytically.

Proposition 9. *If K is sufficiently small, $K \leq (1-\hat{r})\pi$, and harm is small, $E[h] \leq V$, then for any regulatory framework $(L(\cdot), \phi, \bar{r})$, the firm does not invest in RI. Therefore, a constrained-optimal regulatory regime is described in Proposition 6.*

Proof. Consider the *constrained-optimal regulatory regimes*, $(L^O(\cdot), \phi^O, \bar{r}^O)$ solves the following problem:

$$\max_{\{L(\cdot),\phi,\bar{r}\}} F(1-\hat{p}) \left\{ \hat{p} \int (S-h)\hat{b}_h dG(h) + (1-\hat{p})(\hat{a}_U \phi + 1 - \phi) R_P \right\}$$

subject to: $\phi \in [0,1]$, $L(h) \in [0,\pi]$, $\bar{r} \in [0,\infty)$. In addition,

$$R_P = \hat{r} \int (S-h)\hat{b}_h dG(h) + (1-\hat{r})(\hat{b}_U \phi + 1 - \phi)(S - E[h]) - c(\hat{r}),$$

and

$$F(1-\hat{p})[\hat{p}\hat{\pi} + (1-\hat{p})\hat{A}] \geq K.$$

where $(\hat{r}, (\hat{b}_h)_{h\geq 0}, \hat{b}_U, \hat{a}_U)$ is the regulator's strategy (Proposition 4) and $\hat{p}$ is the firm's investment in RI (Proposition 5).

When $V \geq E[h]$, $\hat{b}_U = 1$, and $\tilde{R} \geq 0$, so $\hat{a}_U = 1$. Therefore, ϕ is irrelevant in the RI firm's problem, which can be written as

$$\max_{\widetilde{p}\in[0,1]} F(1-\widetilde{p}) \left\{ \hat{\pi} + (1-\widetilde{p})(1-r^*)(\pi - \hat{\pi}) \right\}.$$

Because $\pi \geq \hat{\pi}$, the objective function is decreasing, so the solution is $\hat{p} = 0$. The firm receives an expected entry payoff of $r^*\hat{\pi} + (1-r^*)\pi - K$. As long as this payoff is positive (e.g., when $K \leq (1-r^*)\pi$) the planner can achieve it by implementing the regulator's optimal actions. If not, the planner may need to limit the regulator's resources. □

The proposition shows that when the expected harm is sufficiently small, there is no need to distort the efficient regulatory decisions. This can happen for two different reasons. It is either the case that it is efficient to place the burden of proof on the regulator ($p^* = 0$), or a positive level of RI is efficient but inducing RI in equilibrium would require distorting the regulator's actions too much. Therefore, it is optimal to give up on trying to induce RI.

However, if efficiency demands a positive investment in RI, then the planner's constrained-optimal regulatory framework will *distort* the regulator's efficient choices. This means that the regulator would invest too much (or too little) to learn ex-post, or would fine the firm

too little (or too much), or it would abuse the precautionary principle.

We simplified the problem of finding the constrained-optimal regulatory regimes by assuming that fines satisfy a single-crossing condition. This allows us to characterize the optimal solution using two additional parameters rather than a function.

We focus on the case of large harm, $V < E[h]$, so $\hat{b}_U = 0$. The choice of $L(h)$ is relevant for the implementation of $\hat{b}_h$ only when $h \in [V, S]$. In particular, $\hat{b}_h = 1$ for $h < V$ and $\hat{b}_h = 0$ for $h > S$, for any function $L(h)$. To simplify our analysis, We make the following simplification:

Assumption 2. *$V + L(h) - h$ satisfies single crossing for $h \in [V, S]$.*

Under this assumption, there exists $\bar{h} = \max\{h \in [V, S] : V + L(h) \geq h\}$. Thus, $\hat{b}_h = 1$ if and only if $h \leq \bar{h}$.

Let us define

$$L \equiv \int_0^{\bar{h}} L(h) dG(h).$$

The assumption $\hat{b}_h = 1$ if and only if $h \leq \bar{h}$ imposes that $L(h) \geq h - V$ for $h \in [0, \bar{h}]$. Therefore, $L(h)$ is bounded below by $\int_V^{\bar{h}} (h - V) dG(h)$ and above by $\pi G(\bar{h})$.

Moreover, we obtain:

$$\int (V + L(h) - h) \hat{b}_h dG(h) = \int_0^{\bar{h}} (V - h) dG(h) + L$$

and

$$\int (\pi - L(h)) \hat{b}_h dG(h) = \pi G(\bar{h}) - L.$$

These results allow us to search for a solution in which two parameters, $\bar{h} \in [V, S]$ and $L \in [\int_V^{\bar{h}} (h - V) dG(h), \pi G(\bar{h})]$, completely characterize the function $L(\cdot)$.

Finding the optimal-constrained regulatory regime for the parameters (V, π, G, F, K, $c(\cdot)$) is a standard optimization problem. However, finding a closed-form solution to this problem is intractable. For this reason, we provide a numerical solution to this problem in the main text.