Regulation and responsible innovation

Nicolas Figueroa, Carla Guadalupi, Jorge Lemus

The Journal of Law, Economics, and Organization, 2026, **00**, 1–27
https://doi.org/10.1093/jleo/ewag012
Article

OXFORD

Regulation and responsible innovation

Nicolas Figueroa[1], Carla Guadalupi[2], Jorge Lemus [3,*]

[1]Instituto de Economía, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Santiago, 8240000, Chile
[2]Facultad de Economía y Negocios, Universidad Andrés Bello, Av. Fernández Concha 700, Santiago, 7550000, Chile
[3]Department of Economics, University of Illinois, 1407 West Gregory Drive, Urbana, Illinois, 61801, United States

*Corresponding author: E-mail: jalemus@illinois.edu

ABSTRACT

Should product harms be investigated by firms via "Responsible Innovation" (RI) or by regulators? Firms can investigate potential harms early on, but only by diverting resources from innovation. Regulators, by contrast, often investigate harms after products are on the market. We characterize the efficient solution, which hinges on the firm's opportunity cost of RI—how sensitive innovation success is to resource diversion—and the regulator's cost of investigating harms. Our main insights are threefold. First, efficiency assigns primary responsibility to the lowest-cost party: firms when innovation success is insensitive to resource diversion, regulators when innovation success is highly sensitive. Second, efficient regulatory actions can be implemented with standard tools—fines, bans, and investigations—but this hinders firms' RI incentives. Third, competition reduces RI investments for each firm but can raise aggregate RI. Our results highlight the trade-off between efficient oversight and incentives for responsible innovation while cautioning against one-size-fits-all regulation (*JEL* L51, M14, O32, D82).

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 that responsibility may be at odds with fast-paced innovation. For instance, OpenAI disbanded a team that managed long-term risks. The team was promised

[1] Such technologies include artificial intelligence, social media, cryptocurrencies, biometrics, and self-driving cars. See, for example, Bliss et al. (2024) and Shah (2023).

20% of the company's computing resources, but it did not materialize. This contributed to 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 artificial intelligence (AI) risks. A former OpenAI safety team leader stated that "over the past years, safety culture and processes have taken a backseat to shiny products."[2] 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."[3]

These trends underscore the importance of the regulatory regime: It determines the trade-off firms face between innovating and learning about potential harms, and the regulator's incentive and ability to mitigate risks. Previous research has examined these issues separately: Some studies focus on how fixed regulations influence innovation decisions (see, e.g. Immordino et al. 2011), while others examine post-market learning but ignore the initial innovation trade-offs that firms face (see, e.g. Henry et al. 2022). We contribute by developing a framework for the strategic interaction between the firm and the regulator. The regulator acts only after the product is on the market. Anticipating the regulator's actions, the firm decides whether to divert some resources from innovation to harm investigation. When resources are limited, the cost of responsible innovation is a lower probability of success.

The regulator, who places greater weight on consumer surplus than on firm profits, makes decisions constrained by a regulatory regime only after the product enters the market.[4] If the harm is *known*, the regulator can *ban* the product or *allow* it and impose a fine. Fines are a transfer from the firm to the consumers, and they can be interpreted as a firm investment mandated by the regulator to mitigate harm, as a permit for product usage (similar to pollution permits), or as partial remedies (e.g. usage limits or warning labels). If the harm is *unknown*, the regulator can *investigate*, subject to a predetermined budget. If the investigation is inconclusive, he may be able to ban the product without proof of harm under the "precautionary principle."[5]

From an economic perspective, the precautionary principle can produce both inefficiencies and welfare gains. Over-application can block innovations that generate social value, as the ban on GMOs or the US ban on cyclamate sweeteners, where regulators removed products despite limited ex post evidence of harm in other jurisdictions. Conversely, precautionary bans can avert severe harm, as illustrated by the US ban of thalidomide and the Montreal Protocol's elimination of CFCs. Our analysis focuses primarily on products like the ones described by these examples, that is, products with intrinsic characteristics that cannot be easily modified once the harm is identified. Other examples include asbestos, lead-based paint, some pollutants, tobacco products, and social media platforms.

Our main insights are threefold. First, efficient delegation requires correcting the regulator's bias toward consumer surplus, which otherwise leads to excessive caution. This can be done by limiting investigative resources, imposing sufficiently large fines, and allowing precautionary bans when expected harm is very large. Second, regulatory regimes that induce efficient ex post regulation undermine firms' incentives to invest in RI. Put differently, encouraging RI requires a regime

[2] These issues are discussed by Knight (2024), Wiggers (2024), Metz (2023), and Rafferty (2024).

[3] Google launched Bard despite employees' concerns (Gustafson 2023). Meta disbanded its "Responsible AI" team and reallocated its members to product development (Picciotto 2023). More broadly, Ahmed et al. (2024) find that nearly 90% of AI companies conduct no research into responsible AI.

[4] Many agencies are explicitly mandated to protect consumers, rather than to maximize aggregate surplus. For instance, the FDA's primary mandate is to safeguard public health; the Consumer Financial Protection Bureau in the United States or national consumer watchdogs in Europe also prioritize consumer protection.

[5] 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. 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.

that incentivizes the regulator to make inefficient decisions. Third, when firms compete for the market, the aggregate level of RI can be insufficient or excessive.

To derive these insights, we first characterize the efficient allocation of responsibility for assessing harm between the regulator and the firm. The key trade-off is between the cost of post-entry investigation and the reduction in innovation success from diverting resources to RI. When post-entry investigation is relatively cheap, efficiency requires the regulator rather than the firm to learn the harm. Moreover, the harm investigation by the firm and the regulator transitions from substitutes to complements as the product's social value increases. Specifically, with low social value, the firm has the primary responsibility of assessing harm. As social value rises, the regulator assumes more responsibility, while the firm invests less in RI, given the higher opportunity cost of foregone innovation. Finally, when social value exceeds expected harm, both parties' incentives to learn diminish, since the product's benefits outweigh its expected harm.

We then explore whether the efficient outcome can be implemented through a combination of fines, precaution, and investigative capacity. Because the regulator focuses on consumers and ignores firm profits, he is more cautious than a planner considering overall welfare. This excess caution can lead to inefficiently large investigations or overuse of precaution to ban products without evidence of harm. We show that three commonly used levers (fines, precaution, and budget) are enough to implement efficient regulatory choices. For small expected harm, it is key to limit the regulator's investigative resources because his over-caution incentivizes him to investigate 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 products without proof of harm. For large expected harm, fines must be maximal to incentivize an efficient investigation, the regulator must be given enough resources to investigate, and the power to ban a 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 harm only triggers 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, if confined to these three regulatory levels, when a strictly positive level of RI is efficient, it is necessary to distort the regulator's decisions to incentivize RI by the firm. 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.

It is worth emphasizing that the regulatory regime determines who generates evidence about harm and when it is produced. Fines can make the regulator internalize firm profits.[6] But because fines require proof of harm, they cannot correct incentives after an inconclusive investigation. In that case, the regulator may lean on precautionary bans. Conversely, the prospect of collecting fines may induce the regulator to over-investigate. From the firm's perspective, if fines are not excessively large, pre-market RI can be profitable because evidence on harm can avoid unnecessary bans from an overly cautious regulator.

[6] 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).

We also study how competition affects RI. We show that when firms compete in a winner-takes-all contest for the market, each individual firm invests less in RI than a monopolist, since diverting resources toward RI weakens its competitive position. Nevertheless, aggregate RI may be higher under competition because multiple firms contribute, and the likelihood that at least one generates information about harm increases. Whether competition induces aggregate underinvestment or overinvestment depends on the role of chance in determining market leaders: if who becomes the leader is highly uncertain, firms worry less about losing market share from investing in RI, and aggregate RI can exceed the monopoly benchmark; if it is more predictable, competitive pressure reduces RI and shifts the burden of learning about harm to the regulator.

Our results contribute to the ongoing debate on how to regulate innovative products. While some scholars advocate for strict liability (Matthews 2024), others propose ex ante regulation (Kretschmer et al. 2023) or even outright bans when the harm is unknown (Yudkowsky 2023). Our results show that there is no one-size-fits-all solution. Generally, efficient regulation design requires calibrating fines, precautionary principle, and investigative budgets to the product and industry characteristics, and on whether it is more costly for firms or regulators to generate information about potential harms.

When innovation success is very sensitive to resource diversion and regulators can evaluate harm with relative ease—as in the aerospace industry, where innovation requires significant capital and specialized talent but harm evaluation is well established—regulators should primarily investigate potential harms. By contrast, when innovation success is insensitive to resource diversion but assessing harms is costly for regulators—as in digital platforms or algorithmic pricing—firms should undertake responsible innovation.[7] Allocation of responsibility becomes more ambiguous when both sides face high costs—as in gene editing, where firms rely on scarce molecular biologists and regulators lack expertise—or when both sides face low costs—as in consumer electronics, where engineering talent is plentiful and safety standards (e.g. electrical shocks, battery fires) are clear. In such cases, firms and regulators should typically share investigative responsibility, with regulatory regimes calibrated to industry characteristics to incentivize learning by both parties.

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) investigate dynamic learning by incorporating learning-by-doing into firm decisions and conclude that this may result in sub-optimal care under strict liability. Henry et al. (2022) examine the use of liability, withdrawal, and authorization when information is revealed gradually. Ottaviani and Wickelgren (2023) examine the trade-off between ex ante and ex post regulation, when it is costly to undo a firm's action. In these papers, products are already on the market, so firms

[7] During a congressional hearing, lawmakers were criticized for their limited understanding of Facebook, underscoring the broader challenge regulators face in keeping up with disruptive technologies (Stewart 2018).

do not strategically allocate resources between innovation and learning about potential harms. 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 and 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.[8]

During the innovation phase, the firm allocates its limited resources (normalized to 1) between product development and "Responsible Innovation" (henceforth RI). By investing a fraction $1-p$ of its resources in product development, the firm succeeds with probability $F(1-p)$, where F is a probability distribution on $[0,1]$ with continuous and positive density, f. That is, to achieve success, the amount invested in innovation must be higher than a random breakthrough threshold, X, with $X \sim F$. Thus, an investment of $1-p$ results in success with probability $\Pr(X<1-p)=F(1-p)$. The remaining fraction of resources, p, is allocated to RI, which publicly reveals the product's harm with probability p, and reveals nothing with probability $1-p$.[9]

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 can the regulator immediately act or investigate the product's potential harm before taking action.[10]

As mentioned in Section 1, our analysis focuses on innovative products that cannot be easily modified. 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 a harm-specific fine, $L(h)$. This fine can also be interpreted as a costly investment by the firm to mitigate consumer

[8] Supplementary Appendix C shows that our results extend to $h \sim G(\cdot|V)$ under mild additional assumptions.
[9] For example, Lewis and Sappington (1994) and Ottaviani and Sørensen (2006) use this "truth-or-noise" structure.
[10] For simplicity, we assume that both profits and harm are zero during the regulator's investigation. In Supplementary Appendix A, we allow for harms to be exogenously revealed after entry and before an investigation.

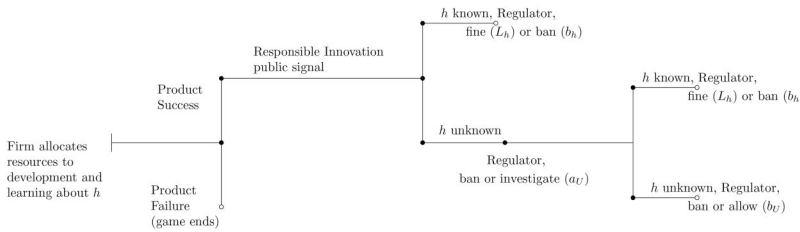

Figure 1. Timing of the model.

harm (e.g. warning labels).[11] 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 $c(0) = 0$.[12] 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.

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.

This assumption is a regularity condition that guarantees a unique interior solution and thus a unique equilibrium. The assumption is satisfied by standard distributions with decreasing reverse hazard rates, including Weibull, Gamma, Pareto, and Lognormal distributions (see, e.g. Kayid et al. 2011). Assumption 1 also holds for any distribution that is strictly concave on $[0, 1]$. Concavity captures diminishing marginal returns to R&D. That is, early "low-hanging-fruit" breakthroughs yield large increases in success probability, whereas additional investment generates marginal improvements. Empirical studies have documented such diminishing returns in firm-level R&D productivity (e.g. Bloom et al. 2020). Note that concavity is sufficient but not necessary for Assumption 1. For example, $F(x) = x^2$ satisfies the assumption despite exhibiting increasing marginal returns. Intuitively, the condition rules out densities that exhibit a local increase in probability mass near the upper end of the support. For instance, the density $f(x) = 6x^2 - 8x + 3$ violates the assumption for $x \in [0.9, 1]$.

[11] In this interpretation, each dollar the firm invests reduces consumer harm by one dollar.

[12] We also impose that $c'(0) = 0$ and $c'(1) \to \infty$ to focus on unique interior solutions. In Supplementary Appendix A, we allow for exogenous revelation of harm (e.g. via media reports, whistleblowers, or accidents) before any regulatory action. We show that we can reinterpret our baseline framework to accommodate this feature by adjusting the investigation cost, which remains convex and nondecreasing, so that our main insights continue to hold for small levels of exogenous revelation.

Technologies with "emergent capabilities," where breakthroughs appear somewhat suddenly with scaling, may violate this condition, potentially generating multiple equilibria.

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.[13] We denote by $(p^*, r^*, (b_h^*)_{h\geq 0}, b_U^*, a_U^*)$ the efficient outcome.[14]

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

4.1 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 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.

Standard results on stochastic orders show that if $\tilde{G}$ is a mean-preserving spread of G, then the efficient amount of learning is higher under $\tilde{G}$. In other words, there is more value in learning about potential harms when the distribution of harm has higher variance.

[13] The efficient solution is not constrained by $\bar{r}$ or ϕ. Fines $L(h)$ are irrelevant because they net out.
[14] This is a "second-best" solution because the planner does not set the firm's prices.

4.2 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 $h\leq S$, that is, $b_h^*=\mathbb{I}(h\leq S)$.*
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 $E[h]\leq S$, that is, $b_U^*=\mathbb{I}(E[h]\leq S)$.*
3) *It is inefficient to invest in RI, that is, $p^*=0$, 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 (which is R^*) is large, in which case it may be efficient to leave the burden of proof entirely to the regulator.[15] However, if acquiring information is costly for the regulator, it may be efficient to allocate resources to RI, provided that $f(1)$ is not too low, which means that marginally 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.

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

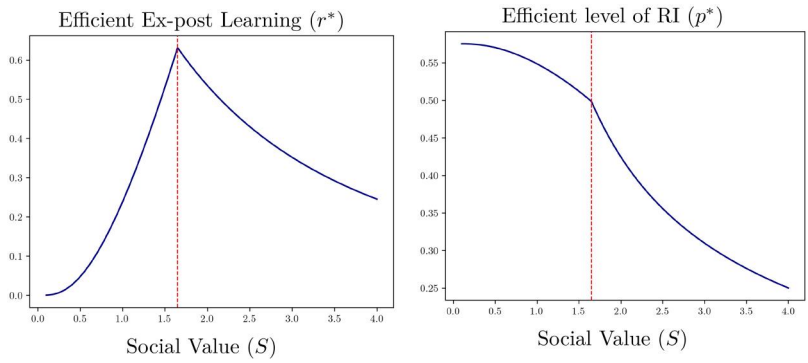

Figure 2. $r^*(S)$ and $p^*(S)$ when the harm distribution is $G = \text{lognormal}(0, 1)$, so that $E[h] = \exp(0.5) = 1.65$ (shown by the vertical dashed line); $F = \text{Beta}(2, 5)$, which satisfies Assumption 1. The regulator's cost of learning the harm with probability r is $c(r) = \frac{r^2}{2}$.

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.

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.

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.[16] 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.

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. We investigate whether it is possible to implement the efficient outcome, $(p^*, r^*, (b_h^*)_{h \geq 0}, b_U^*, a_U^*)$, as the

[16] In the United States, 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.

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.

5.1 Investigation in equilibrium

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

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 $\hat{a}_U = \mathbb{I}(\hat{R} \geq 0)$.

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 learns h from its investigation, which is given by the solution to*

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

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.[17] 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

[17] 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 $V - E[h] \leq S - E[h]$.

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.[18]

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 $h \leq V + L(h)$, that is, $\hat{b}_h = \mathbb{I}(h \leq V + L(h))$.*
2) *If RI does not reveal h:*
   a) *If $\hat{R} < 0$, the regulator invokes the precautionary principle, banning the product with probability ϕ. With probability $1-\phi$, the precautionary ban fails, and the regulator proceeds to conduct an investigation.*
   b) *If $\hat{R} \geq 0$, the regulator does not invoke the precautionary principle, that is, $\hat{a}_U = 1$, and investigates, learning h with probability ρ.*
3) *If the regulator's investigation does not reveal h, the product is allowed with probability $\phi\hat{b}_U + 1 - \phi$, where $\hat{b}_U = \mathbb{I}(E[h] \leq V)$.*

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 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.

5.2 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 the harm is unknown are

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

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". Characterizing the firm's optimal investment in RI is direct using Lemma 2:

[18] 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 $\hat{a}_U\phi + 1 - \phi$.

Proposition 5. *The firm does not invest in RI, that is,* $\hat{p} = 0$*, 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. $\hat{A} = 0$) but a positive expected payoff if RI reveals the harm (i.e. $\hat{\pi} > 0$). This occurs, for example, when fines are not maximal, the regulator does not investigate, and the product is banned without proof of harm. These situations effectively replicate an ex ante approval regime: without proof of safety, the product cannot enter the market, and the firm must invest in RI to access any profits. In such cases, the firm's incentive to avoid the precautionary ban leads to overinvestment in RI from a social perspective. The firm's equilibrium investment is 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.[19] 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* $\hat{p} = p^*$.

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

Efficient regulatory decisions in Definition 1 comprise four components: (i) the decision to allow the product in the market when harm is known ($\hat{b}h$); (ii) the decision to allow the product in the market when, after investigation, harm remains unknown, accounting for the fact that banning the product without evidence of harm under the precautionary principle is effective with probability ϕ ($\phi\hat{b}U + (1-\phi)$); (iii) the decision to ban the product outright without prior regulatory investigation, accounting for the uncertainty inherent in applying the precautionary principle ($\phi\hat{a}_U + (1-\phi)$); and (iv) the regulator's investment in investigating potential harm (ρ). Proposition 6 characterizes regulatory regimes that implement efficient regulatory decisions.

Proposition 6. *The efficient regulatory decisions are implementable by the regulator under the following regulatory regimes for each level of expected harm:*

	Small, $E[h] \leq V$***:***	***Moderate,*** $E[h] \in (V, S)$***:***	***Large,*** $E[h] \geq S$***:***
Fines:	*Unrestricted,* $L(h) \in [0, \pi]$ *for all* $h \geq 0$	*Bounded,* $L(h) \geq h - V$ *for* $h \in [V, S]$	*Maximal,* $L(h) = \pi$ *for* $h \in [0, S]$
Precautionary:	*Irrelevant*	*Forbidden*	*Permitted*
Principle	$\phi \in [0, 1]$	$\phi = 0$	$\phi = 1$

In all cases, investigation resources must be limited to $\bar{r} = r^*$.

Proof. First, consider the case where the harm is known. For any regulatory regime $(L(\cdot), \phi, \bar{r})$, the regulator makes efficient decisions (i.e. $\hat{b}_h = b_h^*$) when the harm is known to be sufficiently small ($h < V$) or sufficiently large ($h > S$).[20] In these cases, the regulator and the planner's incentives are aligned, thus efficiency is guaranteed for any regulatory regime. If harm is known to be moderate, $h \in [V, S]$, the regulator needs a 'sizable' fine (specifically, $L(h) \geq h - V$) to internalize the loss in profit from banning the product.

Figure 3 illustrates that multiple fine schedules—any $L(h)$ in the shaded area—implement the efficient decision b_h^*.

Consider now the case of unknown harm. We begin by examining the decision to allow the product after the regulator's investigation is inconclusive. With small harm, $E[h] < V$, the planner and the regulator allow the product, making the precautionary principle an irrelevant tool for implementing the efficient decision b_U^*. That is, when $b_U^* = \hat{b}_U = 1$, any $\phi \in [0, 1]$

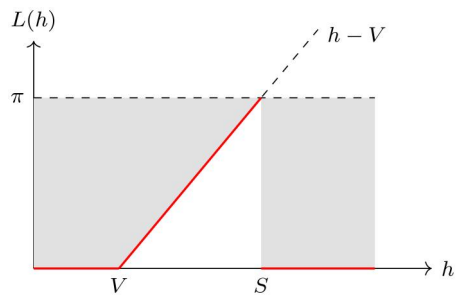

Figure 3. Any function $L(h)$ within the gray area implements $\hat{b}_h = b_h^*$. The solid line indicates the minimum fines that implement b_h^*.

[20] Here, limited liability matters: If $h > S = V + \pi$, then $h \geq L(h) + V$ because limited liability imposes $L(h) \leq \pi$.

satisfies $\phi\hat{b}_U + 1 - \phi = 1$. With moderate harm, $V \leq E[h] \leq S$, an uninformed regulator sees less upside from allowing the product because he ignores the loss in firm profits from banning it. Thus, the regulator tries to ban the product, whereas the efficient decision is to allow it. Therefore, it is necessary to "tie the regulator's hands," forbidding the use of the precautionary principle, that is, $\phi = 0$. Lastly, when $S < E[h]$, it is efficient to ban the product, and the regulator *tries to* do so. However, without evidence of harm, the regulator may fail to do so if the use of the precautionary principle is uncertain: if $\hat{b}_U = 0$, the product remains in the market with probability $1 - \phi$. Therefore, to guarantee that the product is banned, it is necessary to allow the use of the precautionary principle with certainty, that is, $\phi = 1$.

Next, let us consider the regulator's decision to investigate the harm if the firm enters the market without producing evidence of harm. Although the planner always investigates ($a_U^* = 1$), the regulator may prefer to try to ban the product outright and avoid the investigation ($\hat{a}_U = 0$). Proposition 3 shows that this can only happen for moderate harm or large harm when $\phi < 1$. But we imposed $\phi = 1$ with large harm and $\phi = 0$ with moderate harm to implement the efficient decisions *after* the regulator conducts the investigation. Thus, the investigation is always conducted by the regulator, that is, $a_U^* = 1$.[21]

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

$$c'(\hat{r}) = c'(r^*) + \underbrace{\int_0^S L(h)dG(h) + \pi(1 - G(S))}_{\geq 0}.$$

From the convexity of $c(\cdot)$, c' is increasing, so we obtain $r^* \leq \hat{r}$. Without restricting the regulator's budget, he investigates *too much* relative to the efficient level. This occurs because the regulator cares about receiving $L(h)$ when the product is allowed, and ignores the firm's profit loss 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 $\hat{r} \leq \bar{r} \equiv r^*$.

The opposite occurs when the expected harm is large. When $S < E[h]$, $b_U^* = 0$ and, therefore,

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

When the expected harm is large, without knowing h after the investigation, both the regulator and the planner ban the product. However, the regulator does not internalize the profit the firm will make if the investigation reveals $h \leq S$ and, as a consequence, the product is allowed in the market. Thus, the regulator learns *too little* compared to the efficient level. To provide proper investigation incentives, a regulatory regime must set $\int_0^S L(h)dG(h) = \pi$. Under limited liability, $L(h) \leq \pi$, so this condition implies setting the maximum possible fine, $L(h) = \pi$ for $h \leq S$ (almost everywhere). □

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.

[21] When $\hat{a}_U = 0$, $\phi\hat{a}_U + 1 - \phi = 1 - \phi$, which equals $a_U^* = 1$ when $\phi = 0$.

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, that is, $E[h] \leq S$, 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 the firm's profit.

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 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 fines, 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, and using fines that are not maximal, 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 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?

To answer this question, we define a *constrained-optimal regulatory regime* as the best outcome the planner can achieve when unable to directly control the firm's or the regulator's actions. Formally, the planner chooses a regulatory regime—fines, access to the precautionary principle, and investigative budget—to maximize overall welfare subject to the regulator's and the firm's equilibrium behavior. In this optimization problem, the planner trades off inducing responsible innovation (RI) against distorting the regulator's efficient decisions.

Table 1. Constrained-optimal regulatory regime and outcomes.

$E[h]$	μ	p^*	r^*	ϕ	L	$\hat{p}$	$\hat{r}(=\bar{r})$
1.22	−0.30	0.34	0.29	0.00	0.09	0.00	0.29
1.35	−0.20	0.37	0.36	0.00	0.10	0.00	0.36
1.49	−0.10	0.40	0.44	0.00	0.35	0.00	0.44
1.65	0.00	0.43	0.53	1.00	0.56	0.38	0.68
1.82	0.10	0.45	0.64	1.00	0.56	0.40	0.64
2.01	0.20	0.48	0.76	1.00	0.52	0.42	0.57
2.23	0.30	0.49	0.70	1.00	0.47	0.44	0.51
2.46	0.40	0.50	0.64	1.00	0.42	0.46	0.45

Notes: The table shows efficient outcomes, constrained-optimal regulatory regimes, indexed by ϕ and $L = \int_0^S L(h)dG(h))$, and the implied equilibrium values $\hat{p}$ and $\hat{r}$, for different values of expected harm, which come from different distributions of harm indexed by μ: $h \sim G = \text{Lognormal}(\mu, 1)$ Varying Expected Harm and μ.

Characterizing the precise form of these distortions requires simplifying assumptions and, in most cases, numerical simulation. In Supplementary Appendix B, we present these details.

Table 1 presents numerical solutions for the constrained-optimal regulatory regimes for different examples, when the product's social value is fixed at $S = 2$ and consumer surplus at $V = 1$. Each row corresponds to a different distribution of harm, $G \sim \text{LogNormal}(\mu, 1)$, which generate different expected harm levels $E[h]$. We focus on cases where $E[h] > V$ because Proposition 8 shows that for small harm, inducing RI is impossible, so the best outcome is to set the regulatory regime according to Proposition 6, implementing efficient decisions by the regulator.

Table 1 shows that for moderate expected harm (rows 1–3, with $E[h] \leq 1.49$), the unconstrained optimum would involve a positive level of responsible innovation ($p^* > 0$). As Proposition 8 shows, this could in principle be induced by distorting the regulator's decisions. However, this distortion is too costly: the constrained–optimal regime instead forgoes inducing RI and simply ensures that the regulator's decisions are efficient, so that $r^* = \hat{r}$ and $\hat{p} = 0$.

When expected harm exceeds 1.65 (rows 4 onward), it is suboptimal to rely on the regulator alone to investigate the potential harm, so inducing RI by distorting the regulator's actions becomes optimal. In row 4, this involves: (i) granting the regulator excessive resources ($\bar{r} > r^*$), and setting fines high enough to incentivize the regulator to investigate, and (ii) allowing the precautionary principle ($\phi = 1$), so the regulator bans the product if there is no evidence of harm. Since expected fines are lower than expected profits, the firm invests in RI to avoid inefficient bans.

For higher expected harm (rows 5–8), the distortion involves setting low fines to reduce the regulator's incentive to investigate, increasing the likelihood of unjustified bans. This incentivizes firms to invest in RI to avoid unjustified bans, even when the firm is penalized for revealing harm. That is, the benefit of avoiding bans is higher than the expected fines from revealing harm.

Thus, although the efficient solution implies positive RI at moderate harm ($E[h] < 1.49$), constrained optimality does not: only for relatively high harm-inducing RI become welfare-improving despite distortions.

7. COMPETITION

In this section, we consider the impact of competition on investment in RI. Recall that in our baseline model, to achieve success and enter the market, the firm must invest more than a random breakthrough threshold, X, with $X \sim F$.

Suppose now that there are two firms. Firm $i \in \{1,2\}$ invests p_i in RI, leaving $1-p_i$ resources for innovation. We assume that competition to win the market takes the form of a winner-takes-all contest, in which firm i competes with a "score" $s_i = 1 - p_i + \varepsilon_i$, where ε_i are independent, have a mean zero. Firm i wins the market if $s_i \geq s_j$, which is the same as

$$\varepsilon_j - \varepsilon_i \leq p_j - p_i.$$

We assume that $\varepsilon_j - \varepsilon_i \sim D$, where D is a continuous distribution with density d.

Thus, two events need to occur for a firm to enter and win the market: invest sufficient resources to obtain a breakthrough, and win the contest for the market. Taking firm j's investment in RI as given, the probability that firm i achieves a breakthrough and wins the market is

$$\tilde{F}(p_i, p_j) = F(1-p_i)[1 - F(1-p_j) + F(1-p_j)D(p_j - p_i)].$$

In this expression, firm i achieves a breakthrough with probability $F(1-p_i)$. Without competition, this was all that firm i required to enter the market. Now, firm i also needs to win the market. With probability $1-F(1-p_j)$, firm j does not achieve the breakthrough and, therefore, firm i wins the market with probability 1. With probability $F(1-p_j)$, firm j achieves the breakthrough so firm i wins the market if $s_i \geq s_j$, which occurs with probability $D(p_j - p_i)$.

The function $\tilde{F}(p_i, p_j)$ is increasing in p_j and decreasing in p_i, with

$$\begin{aligned}\frac{\partial \tilde{F}(p_i, p_j)}{\partial p_i} &= -f(1-p_i)[1 - F(1-p_j) + F(1-p_j)D(p_j - p_i)] - F(1-p_i)d(p_j - p_i) \\ &\equiv -\tilde{f}(p_i, p_j).\end{aligned}$$

If firm i enters the market, the public signal can come from the RI of either firm, even if it is unsuccessful at innovation. Thus, a public signal is generated as long as at least one firm succeeds, which occurs with probability

$$\tilde{p}(p_i, p_j) = 1 - (1-p_i)(1-p_j).$$

We have that $\tilde{p}(p_i, p_j)$ increases both in p_i and p_j, with

$$\frac{\partial \tilde{p}(p_i, p_j)}{\partial p_i} = (1-p_j).$$

The regulator's actions and regulatory framework determine the firm's continuation payoff if it enters the market and product harm has been revealed $(\hat{\pi})$ of if it has not been revealed $(\hat{A})$. Let us denote $A = \hat{A}$ and denote the information premium by $B = \hat{\pi} - \hat{A}$. Taking p_j as given, the problem of firm i now becomes:

Table 2. Equilibrium values for different σ.

σ	p_c^*	$2p_c^* - (p_c^*)^2$	p^*
0.50	0.000	0.000	0.475
0.75	0.095	0.182	0.475
1.00	0.181	0.329	0.475
1.25	0.235	0.415	0.475
1.50	0.269	0.466	0.475
1.75	0.292	0.499	0.475
2.00	0.308	0.521	0.475

$$\max_{p_i \in [0,1]} \tilde{F}(p_i, p_j)[A + \tilde{p}(p_i, p_j)B]. \tag{13}$$

Proposition 9. *Each firm competing for the market invest less in RI than a monopolist. However, the aggregate RI investment can be (inefficiently) larger under competition.*

The proposition shows that competitive effects hamper *individual* incentives to invest in RI. However, the fact that multiple firms are investing in RI, albeit each one invests less than a single firm facing no competition, can lead to a higher *aggregate* investment in RI. It can be, in fact, excessive compared to the efficient level. Characterizing a closed-form solution of when this is the case is cumbersome, but we provide numerical examples illustrating that competition can either increase or decrease total RI investment. Intuitively, this depends on whether random factors play a large role in determining which firm dominates the market, which in our model corresponds specifically to the variance of ε_i. When randomness plays a prominent role in determining market outcomes, firms are less concerned about losing market share due to resource diversion toward RI investments.

Table 2 shows a numerical example with $A = 1$ and $B = 1.5$. In the example, the distribution of the random breakthrough threshold, F, is Beta(2,5), and the distribution of the difference of the shocks to win the market, D, is Normal$(0, \sigma^2)$. The table presents the individual investment in RI in a symmetric equilibrium, p_c^*, the probability of obtaining an informative signal from one of the firms, $2p_c^* - (p_c^*)^2$, and the investment in RI without competition, p^*, for different values of σ. The table shows that as σ increases (i.e. the contest for the market becomes noisier), the equilibrium investment in RI by each firm (p_c^*) goes up, although it is lower than what a firm that faces no competition would invest (p^*). However, the probability that at least one of the firms produces an informative signal about harm $(2p_c^* - (p_c^*)^2)$ is larger, whenever $\sigma \geq 1.75$.

Competition makes the breakthrough more likely since each firm invests more in innovation (and less in RI) than a firm without facing competition, and also there are multiple firms rather than a single one. Thus, in cases where σ is large, competition increases welfare from two channels: making breakthroughs more likely and generating a public signal about harm more frequently. However, when σ is low, randomness does not play a significant role in determining the market winner, regulators must play a central role by investigating potential harms because firms themselves will compete for the market and disregard RI.

8. 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 complexity and intractability.

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. One could incorporate lobbying efforts or political influence by redefining the regulator's preference as $V + x - h + L(h)$, where x is a transfer from the firm to the regulator, aligning the planner and the regulator's incentives.

We assume that information revelation is binary. A key reason for this simpler structure is that, in our model, fines or liability can be imposed only when actual harm is known. If information arrives as noisy signals, that assumption must be replaced by a legal standard of proof: penalties are triggered only when the signal clears a threshold of certainty (e.g. "beyond a reasonable doubt"), not merely on the basis of expected harm. More generally, information acquisition is gradual and probabilistic; firms and regulators receive partial signals and update beliefs. Allowing dynamic, partially informative signals would enrich the model, but the core logic persists: a Bayesian regulator would use threshold rules—investigate or impose fines when posterior evidence exceeds the standard—so our binary framework captures this policy margin in reduced form.

Our framework assumes public information revelation. An interesting extension would allow firms to strategically conceal findings. Intuitively, if firms can withhold information, regulators become more cautious and more likely to investigate or invoke the precautionary principle, which can reduce the firms' incentives to learn or increase them to avoid precautionary restrictions. Such an extension transforms the model into a signaling game under asymmetric information, featuring technical complications that make equilibrium existence and monotonicity of best responses difficult to establish without further assumptions.

Lastly, a more general framework could allow for profits and consumer surplus to be a function of the harm, that is, $\pi(h)$, $V(h)$, 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.

9. CONCLUSION

This article studies who should bear the burden of discovering product harms—firms through RI before market entry or regulators through ex post investigation—and how regulatory instruments can be designed to align private and social incentives.

We first characterize the efficient benchmark, highlighting a trade-off between the reduction in the probability of successful innovation (which is an opportunity cost for the firm) when resources are diverted to RI and the regulator's cost of investigating harm. When post-entry investigation is relatively inexpensive, efficiency places the burden of proof on the regulator. We also show that RI and regulator investigation can be complements or substitutes depending on the size of the social surplus relative to the expected harm.

However, when decisions are decentralized, efficiency may not be possible. Even with fines, precautionary bans, and investigative budgets, a key tension emerges: whenever the regulator is induced to act efficiently, the firm optimally does not invest in RI. Conversely, regimes that elicit optimal RI generally do so by distorting the regulator's actions away from efficiency.

Recognizing this impossibility, we study constrained-optimal regimes. When expected harm is small or moderate, there is either no feasible way to induce RI or doing so is too costly. Only when harm is sufficiently large does it become welfare-improving to induce

positive RI by granting precautionary authority while limiting fines and, in some cases, investigative resources. Competition adds nuances to these conclusions. In winner-takes-all environments, firms individually under-invest in RI relative to monopolists, yet aggregate RI can exceed monopoly levels when market outcomes are sufficiently uncertain.

The analysis yields immediate policy implications. The design of regulatory regimes should consider two dimensions: (1) the opportunity cost of firm-side RI and (2) the regulator's marginal cost of post-entry learning. Aerospace-like environments call for the regulator to investigate harms; digital platforms call for firms to invest in RI, setting limits on fines and on the regulator's investigative budget. Industries such as gene editing, where the cost of investigating harm is high for the firm and the regulator, suggest shared responsibility with carefully calibrated instruments.

APPENDIX 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

$$[\frac{F(1-p)}{f(1-p)} + (1-p)]B - (A+B) \iff Z(p)B - (A+B).$$

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 $b_h^* = \mathbb{I}(h \leq S)$, (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

$$\int_0^S (S-h)dG(h) - R^* < f(1)R^*. \tag{A1}$$

Therefore, by Lemma 2, we obtain the result. □

Proof of Proposition 2

Proof. We divide the proof into 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 the derivative with respect to S of the RHS of (5), we obtain

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

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 the derivative of (5):

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

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 is positive $(f(1)(S-E[h])b_U^*)$. □

Proof of Proposition 3

Proof. Direct from taking the first-order condition in (6) (see also footnote 12).

Second, if the regulator can invoke the precautionary principle with certainty, that is, $\phi = 1$, 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 4

Proof. The results follow directly from the regulator's optimal decision rule and Proposition 3. □

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:

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

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

$$\frac{\hat{\pi}}{\hat{A}} = \frac{\int_0^S (S-h)dG(h)}{R^*}. \tag{A2}$$

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 (A2) 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 the 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 holds. 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 (A2) 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 (A2) 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 $E[h] < V$. □

Proof of Proposition 9

Proof. Taking the derivative with respect to p_i, we get

$$-\tilde{f}(p_i, p_j)[A + \tilde{p}(p_i, p_j)B] + \tilde{F}(p_i, p_j)B(1 - p_j)$$

Since $\tilde{f}(p_i, p_j) > 0$, the sign of this expression is the same as

$$-[A + \tilde{p}(p_i, p_j)B] + \frac{\tilde{F}(p_i, p_j)}{\tilde{f}(p_i, p_j)}B(1 - p_j)$$

Using the definition of $\tilde{p}(p_i, p_j)$, the expression becomes

$$-A - p_i(1 - p_j)B - p_jB + \frac{\tilde{F}(p_i, p_j)}{\tilde{f}(p_i, p_j)}B(1 - p_j)$$

Now, adding and subtracting B, we can write

$$-A - p_i(1 - p_j)B + B(1 - p_j) - B + \frac{\tilde{F}(p_i, p_j)}{\tilde{f}(p_i, p_j)}B(1 - p_j)$$

Factorizing terms, we get:

$$[\frac{\tilde{F}(p_i,p_j)}{\tilde{f}(p_i,p_j)}+1-p_i]B(1-p_j)-(A+B).$$

Unpacking the terms $\tilde{F}(p_i,p_j)$ and $\tilde{f}(p_i,p_j)$ for their definitions, the expression above is equivalent to:

$$[\frac{F(1-p_i)}{f(1-p_i)+Q(p_i,p_j)}+1-p_i]B(1-p_j)-(A+B), \tag{A3}$$

where

$$Q(p_i,p_j)=\frac{F(1-p_i)d(p_j-p_i)}{[1-F(1-p_j)+F(1-p_j)D(p_j-p_i)]}.$$

Since $Q(p_i,p_j)>0$, the term in the brackets in (A3) is positive, and $1-p_j\leq 1$,

$$[\frac{F(1-p_i)}{f(1-p_i)+Q(p_i,p_j)}+1-p_i]B(1-p_j)-(A+B)<[\frac{F(1-p_i)}{f(1-p_i)}+1-p_i]B-(A+B).$$

Note that the right-hand side of the inequality corresponds to the first-order condition for the firm problem without competition. Therefore, firm i's best response to p_j, denoted p_i^*, must be strictly lower than the firm solution without competition, p^*, by the assumption that

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

is monotonically decreasing. □

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of Law, Economics, & Organization* online.

FUNDING

None declared.

Conflict of interest statement. None declared.

ACKNOWLEDGMENTS

For comments and suggestions, we thank participants and discussants at IIOC (2024), ALEA (2024), JEI (2024), Universidad Andrés Bello, and TOI Chile (2024). Nicolás Figueroa gratefully acknowledges support from the Larraín Bascuñán Chair in Economics and Artificial Intelligence.

REFERENCES

Acemoglu, Daron, and Todd Lensman. 2024. "Regulating Transformative Technologies," 6 *American Economic Review: Insights* 359–76.

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

Baumann, Florian, and Tim Friehe. 2016. "Learning-by-Doing in Torts: Liability and Information about Accident Technology," 138 *Economics Letters* 1–4.

Bliss, Nadya, Kevin Butler, David Danks, Ufuk Topcu, and Matthew Turk. 2024. "Addressing the Unforeseen Harms of Technology CCC Whitepaper," *arXiv preprint: 2408.06431*.

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb. 2020. "Are Ideas Getting Harder to Find?" 110 *American Economic Review* 1104–44.

De Chiara, Alessandro, and Ester Manna. 2022. "Corruption, Regulation, and Investment Incentives," 142 *European Economic Review* 104009.

Friehe, Tim, and Elisabeth Schulte. 2017. "Uncertain Product Risk, Information Acquisition, and Product Liability," 159 *Economics Letters* 92–95.

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

Gustafson, Aaron. 2023. "Google Bard AI Chatbot Raises Ethical Concerns from Employees," *Bloomberg*. https://www.bloomberg.com/news/features/2023-04-19/google-bard-ai-chatbot-raises-ethical-concerns-from-employees

Henry, Emeric, Marco Loseto, and Marco Ottaviani. 2022. "Regulation with Experimentation: Ex Ante Approval, Ex Post Withdrawal, and Liability," 68 *Management Science* 5330–47.

Immordino, Giovanni, Marco Pagano, and Michele Polo. 2011. "Incentives to Innovate and Social Harm: Laissez-Faire, Authorization or Penalties?" 95 *Journal of Public Economics* 864–76.

Kayid, M., H. Al-Nahawati, and I. A. Ahmad. 2011. "Testing Behavior of the Reversed Hazard Rate," 35 *Applied Mathematical Modelling* 2508–15.

Knight, Will. 2024. "OpenAI's Long-Term AI Risk Team Has Disbanded," *Wired*. https://www.wired.com/story/openai-superalignment-team-disbanded/

Kretschmer, Martin, Tobias Kretschmer, Alexander Peukert, and Christian Peukert. 2023. "The Risks of Risk-based AI Regulation: Taking Liability Seriously," *arXiv preprint: 2311.14684*.

Lewis, Tracy R. and David E. M. Sappington. 1994. "Supplying Information to Facilitate Price Discrimination," 35 *International Economic Review* 309–27.

Matthews, Dylan. 2024. "Can the Courts Save Us from Dangerous AI? A Law Professor Proposes an Old-Fashioned Remedy for Very New Problems: Legal Liability," *Vox*. https://www.vox.com/future-perfect/2024/2/7/24062374/ai-openai-anthropic-deepmind-legal-liability-gabriel-weil

Metz, Cade. 2023. "What Happened When OpenAI's Sam Altman Was Suddenly Removed as CEO," *The New York Times*. https://www.nytimes.com/2023/11/18/technology/open-ai-sam-altman-what-happened.html

Ottaviani, Marco and Peter Norman Sørensen. 2006. "Reputational Cheap Talk," 37 *Rand Journal of Economics* 155–75.

Ottaviani, Marco and Abraham L. Wickelgren. 2024. "Approval Regulation and Learning, with Application to Timing of Merger Control," 40 *Journal of Law, Economics, and Organization* 597–624.

Picciotto, Rebecca. 2023. "Facebook-parent Meta breaks up its Responsible AI team," *CNBC*. https://www.cnbc.com/2023/11/18/facebook-parent-meta-breaks-up-its-responsible-ai-team.html

Rafferty, Conor. 2024. "OpenAI Researcher Resigns, Claiming Safety Has Taken 'A Backseat to Shiny Products'," *The Verge*. https://www.theverge.com/2024/5/17/24159095/openai-jan-leike-superalignment-sam-altman-ai-safety

Requate, Till, Tim Friehe, and Aditi Sengupta. 2023. "Liability and the Incentive to Improve Information about Risk When Injurers May be Judgment-Proof," 76 *International Review of Law and Economics* 106168.

Shah, Simmone. 2023. "Sam Altman on OpenAI, Future Risks and Rewards, and Artificial General Intelligence," *Time*. https://time.com/6344160/a-year-in-time-ceo-interview-sam-altman/

Shavell, Steven. 1992. "Liability and the Incentive to Obtain Information about Risk," 21 *Journal of Legal Studies* 259–70.

Stewart, Emily. 2018. "Lawmakers Seem Confused about What Facebook Does—and How to Fix It," *Vox*. https://www.vox.com/policy-and-politics/2018/4/10/17218982/mark-zuckerberg-testimony-congress-facebook

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

Wiggers, Kyle. 2024. "OpenAI Created a Team to Control 'Superintelligent' AI—Then Let It Wither, Source Says," *TechCrunch*. https://techcrunch.com/2024/05/18/openai-created-a-team-to-control-superintelligent-ai-then-let-it-wither-source-says/

Yudkowsky, Eliezer. 2023. "Pausing AI Developments Isn't Enough. We Need to Shut It All Down," *Time magazine*.

The Journal of Law, Economics, and Organization, 2026, 00, 1–27
https://doi.org/10.1093/jleo/ewag012
Article