NBER WORKING PAPER SERIES

REGULATING ARTIFICIAL INTELLIGENCE

Joao Guerreiro
Sergio Rebelo
Pedro Teles

Working Paper 31921
http://www.nber.org/papers/w31921

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2023, revised May 2026

We thank Rodrigo Adao, Mark Aguiar, Marco Bassetto, Diana Bonfim, Eduardo Dávila, Lukas Freund, Luis Garicano, Doh-Shin Jeon, Yassine Lefouili, Chen Lian, Alessandro Pavan, Pascual Restrepo, Sanjay Singh, Hélène Rey, and Eduard Talamàs for their comments, and Ramya Raghavan for excellent research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Minneapolis, the Federal Reserve System, or the National Bureau of Economic Research.

Regulating Artificial Intelligence
Joao Guerreiro, Sergio Rebelo, and Pedro Teles
NBER Working Paper No. 31921
November 2023, revised May 2026
JEL No. H21, O33

ABSTRACT

Advances in AI offer substantial benefits but also pose societal risks. We analyze optimal regulation under uncertainty about societal costs, differing expectations regarding risks, and opportunities to reduce uncertainty through beta testing. Pigouvian taxes fail to achieve the first-best outcome due to heterogeneous beliefs about risks and the regulator's inability to observe developers' expectations. We propose a two-stage optimal policy: first, deciding between immediate release or sandbox experimentation; second, using gathered information to determine whether to publicly release or withdraw the algorithm. This approach achieves the socially optimal outcome.

Joao Guerreiro
University of California, Los Angeles
and Federal Reserve Bank of Minneapolis
and also NBER
jguerreiro@econ.ucla.edu

Pedro Teles
Banco de Portugal
and Univ Catolica Portuguesa and
CEPR
pteles@ucp.pt

Sergio Rebelo
Northwestern University
Kellogg School of Management
Department of Finance
and CEPR
and also NBER
s-rebelo@northwestern.edu

1 Introduction

In 1950, Isaac Asimov published *I, Robot*, a collection of short stories about the dilemmas of a world where robots powered by artificial intelligence (AI) interact with humans. Recent advances in AI have brought these dilemmas from the realm of science fiction to the pages of newspapers and the halls of parliaments.

AI algorithms promise great benefits but also pose substantial risks. Some risks stem from the alignment problem (Wiener, 1960), where AI systems optimize narrowly defined objectives while neglecting broader human values. For example, social media algorithms may maximize user engagement at the cost of user well-being (Russell et al., 2015, Amodei et al., 2016). Other dangers arise from AI's potential to facilitate harmful activities, such as generating deepfakes for fraud, automating cyberattacks, manipulating users, and aiding the design of biological or chemical weapons.

There is substantial uncertainty about these risks. In a recent interview (Tyrangiel, 2025), Sam Altman, the CEO of OpenAI, says "I still expect that on cybersecurity and bio stuff we'll see serious, or potentially serious, short-term issues that need mitigation. Long term, as you think about a system that really just has incredible capability, there's risks that *are probably hard to precisely imagine and model*. But I can simultaneously think that these risks are real and also believe that the only way to appropriately address them is to ship product and learn."

There is also considerable disagreement about AI's societal risks, even among AI pioneers. Geoffrey Hinton resigned from Google to openly discuss the potential threats AI poses to humanity (Heaven, 2023). In contrast, Yann LeCun, Meta's Chief AI Scientist from 2013 to 2025, and Richard Sutton, a professor at the University of Alberta and, like Hinton and LeCun, a Turing Award recipient, have both dismissed these concerns as overblown (Hart, 2024, Scott, 2025).

Our analysis emphasizes these two distinct frictions. The first is uncertainty:

the societal costs of AI are unknown. The second is disagreement: developers and regulators may hold systematically different beliefs about these risks. These frictions are reflected in recent efforts by major insurers to exclude AI-related liabilities from corporate insurance policies (see, e.g., Harris and Criddle, 2025).

We classify societal costs into two categories. The first is negative externalities, such as fueling political polarization, facilitating fraud, disseminating false information, and jeopardizing financial stability. The second is "internalities," where AI systems manipulate or exploit cognitive biases, leading individuals to make choices that reduce their own welfare.

Beta testing and red-teaming can help identify AI risks before deployment. Beta testing exposes the AI algorithm to a limited group of users to assess societal costs. Red-teaming involves hiring experts to actively probe for vulnerabilities.[1] To simplify, we focus our analysis on beta testing.

In our model, beta testing emerges endogenously as a response to uncertainty. When uncertainty is high and deployment is reversible, beta testing is valuable because it generates information about whether large-scale deployment is socially desirable.

When we first presented this paper, the idea that firms would deliberately restrict deployment to learn about societal risks was often viewed as implausible. Recent developments suggest otherwise. In April 2026, Anthropic announced Project Glasswing, a restricted release of its Claude Mythos model designed to generate information about potential risks before wider deployment. This type of limited rollout is precisely the form of endogenous beta testing that emerges in our framework.

In our model, developers' equilibrium choices regarding beta testing and algorithm release are not socially optimal because developers do not fully internalize

[1] The term "red teaming" originated in Cold War military strategy, where red teams simulated adversarial attacks to test defenses. It has since been adopted in cybersecurity and AI safety to describe efforts to uncover vulnerabilities.

the external effects imposed on the broader population. Pigouvian taxes are commonly used to align private and social incentives. However, when developers and regulators disagree about risks and developers' beliefs are private information, taxes based on expected or realized harms fail to implement the socially optimal level of algorithm testing and release.

Our main result is that the optimal policy is a two-stage regulatory mechanism. In the first stage, the regulator either permits deployment or requires beta testing on a limited set of users. In the second stage, the regulator uses the information gathered in the first stage to decide whether broader deployment should be permitted or the algorithm should be withdrawn. This policy implements the social optimum without requiring the regulator to observe developers' beliefs about AI risks.

Even though our paper focuses on AI regulation, our analysis applies more broadly to industries characterized by substantial uncertainty and heterogeneous beliefs about external harms. The regulatory framework that emerges as optimal in our model has been adopted in two prominent domains that share these features.

The first domain is financial regulation. Since the UK Financial Conduct Authority introduced regulatory sandboxes in 2015, more than 50 regulators, including the U.S. Consumer Financial Protection Bureau, have adopted this approach (Cornelli et al., 2024). By allowing firms to test products with limited users in controlled environments, sandboxes reveal consumer protection, cybersecurity, regulatory, and systemic risks before broader market introduction.

The second domain is autonomous vehicle regulation. The U.S. National Highway Traffic Safety Administration (Burd, 2021) and Germany's Federal Ministry of Transport both operate sandbox programs that test autonomous vehicles in controlled environments. These programs generate data, reduce uncertainty, and guide deployment decisions.

Our paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces our benchmark model and 4 discusses optimal policy. In Section

5, we analyze scenarios where AI algorithms create internalities. In Section 6, we discuss other frictions which might be relevant for the design of AI regulation and how our results relate to current regulatory approaches in the U.S. and the European Union. Finally, we summarize our results and discuss additional considerations relevant to the design of AI regulation.

2 Related literature

Our paper relates to several strands of literature. Recent work analyzes AI's effects on political polarization, misinformation, surveillance, and privacy (Acemoglu, 2021; Beraja et al., 2023; Acemoglu et al., 2025; Argyle, 2025). Related work studies AI internalities, including weakened cognitive skill formation, distorted learning incentives, and the exploitation of cognitive biases (Puri and Veldkamp, 2026; Bursztyn et al., 2025; Acemoglu et al., 2026). Our contribution is to characterize optimal regulatory responses in environments with uncertainty and disagreement about these harms.

The second strand is an emerging economics literature on AI regulation. Acemoglu and Lensman (2024) study optimal AI adoption when disaster risk is learned exogenously over time, creating incentives to delay deployment. Gans (2024) considers endogenous learning through adoption and shows that learning about AI's social costs can justify faster diffusion when adoption is reversible. Both papers abstract from heterogeneous and unobservable beliefs held by AI developers and from the endogenous design of beta-testing and conditional-approval mechanisms, which are central to our analysis. Lian and Schaab (2026) study AI regulation under international competition, showing how strategic interactions across countries shape regulatory incentives.

Our paper also relates to work on the design of regulatory instruments and the choice between price- and quantity-based regulation. In work subsequent to ours, Gans (2025) study regulation across technological paths with uncertain externalities and show that unrestricted ex-ante Pigouvian taxation achieves the first best. By contrast, in our setting, which features heterogeneous and privately held beliefs, taxes based on expected or realized harms generally fail to implement the first-best allocation.

Weitzman (1974, 1978) compares price- and quantity-regulation under uncertainty when firms observe cost shocks unknown to the planner. He shows that quantity regulation is preferred when marginal costs are relatively flat, because price errors generate large, inefficient quantity responses. In our model, price instruments fail for a different reason: firms' beliefs about external harms are unobservable, so the planner can only set taxes based on its own expectations about firms' beliefs. In related work, Farhi and Gabaix (2020) show that quantity regulation can dominate Pigouvian taxation when agents differ in their attention to taxes. In our setting, heterogeneity arises from disagreement about risks rather than from behavioral biases, making optimal Pigouvian taxes informationally demanding. Our key contribution is to endogenize information generation through beta testing, which informs regulatory approval.

Our emphasis on beta testing, regulatory sandboxes, and conditional approval relates to a broader literature on experimentation, information asymmetries, and regulatory learning. Callander (2011) studies the value of experimentation, while the literature on clinical trials and conservative bandit algorithms analyzes how experimentation can be constrained to limit downside risk during learning (Thompson, 1933; Gittins, 1974; Wu et al., 2016; Kazerouni et al., 2017; Jagerman et al., 2020).

In the context of AI regulation, Koh and Sanguanmoo (2025) show that adaptive sandboxes with evolving quantity limits are robustly optimal when regulators are uncertain about firms' learning processes, while Zarra (2026) analyze AI regulatory sandboxes under the European Union AI Act as mechanisms for reducing information asymmetries and facilitating regulatory learning. Related work studies random audits and disclosure requirements in algorithmic regulation (Kleinberg et al., 2018; Rambachan et al., 2020), the effects of competition on regulators' ability to extract firms' private information about harms (Callander and Li, 2024), the costs of requiring transparent and interpretable algorithms (Blattner et al., 2021), and liability design under catastrophic risk (Chen and Hua, 2026). Lehr and Restrepo (2025) study socially minded AI firms that trade off access expansion against labor-market disruption, while Karger et al. (2026) document substantial disagreement among experts and the public about AI's economic effects, consistent with our emphasis on heterogeneous beliefs. We contribute to this literature by studying environments in which experimentation generates externalities and firms hold heterogeneous beliefs about social risks.

Another strand of literature examines the impact of AI on the labor market (e.g., Burstein, Morales, and Vogel, 2019, Martinez, 2021, Acemoglu and Restrepo, 2022, Guerreiro, Rebelo, and Teles, 2022, Costinot and Werning, 2023, Thuemmel, 2023, Ide and Talamàs, 2025, Firooz, Liu, and Wang, 2025, Freund and Mann, 2026, and Hampole et al., 2025), the role of data in AI (e.g., Jones and Tonetti, 2020, and Farboodi and Veldkamp, 2021), and potential existential AI risks (Jones, 2024). Our contribution relative to this literature is to characterize optimal policy responses to AI's externalities and internalities.

In Section 6, we further relate our results to the regulation literature, particularly work on regulation under limited liability and imperfect verifiability by Tirole (2010) and Kolstad and Ulen (1983).

Our paper makes three main contributions. First, we study how uncertainty about AI's internal and external effects, together with disagreement about their likelihood, shape optimal regulation. Second, we characterize the role of beta testing in mitigating downside risks. Third, we use the model to shed light on the contrasting regulatory approaches of the United States and the European Union.

3 Benchmark model

We consider a two-period model with a continuum of identical households and a single AI developer. We interpret the first period as the short run and the second as the long run.[2] In our model, using AI carries inherent risks, as it can create misalignments or facilitate activities that generate significant social costs. When the algorithm is deemed too risky for full initial release, the developer may choose to conduct beta testing by distributing the algorithm to a limited subset of the population. Based on the outcome of this beta testing, the developer can then decide whether to release the algorithm in the second period.

In period one, before uncertainty is realized, developers choose whether to beta test or fully release the algorithm and households decide whether to purchase the algorithm. In period two, decisions are made after beta-testing results are observed but before period-two uncertainty is realized.

We allow for disagreement between society and AI developers about the likelihood of societal risks. This disagreement can arise because developers may be overly optimistic, expecting negative external effects to be small.

We now discuss the household problem, the AI developer's problem, and the unregulated equilibrium. Then, we characterize the social optimum and compare it with the unregulated equilibrium.

3.1 Unregulated equilibrium

Household problem The economy has a continuum of households indexed by $j \in [0, N]$, where N denotes the total number of households in the population. Each household lives for two periods.

Household j's momentary utility in period t, $v_{j,t}$, has a quasi-linear form:

$$v_{j,t} = y_t + [u - \mathbb{E}_t^s(i_t^2) - p_t] \times \mathcal{I}_{j,t} - \mathbb{E}_t^s(e_t^2), \tag{1}$$

[2]We omit time subscripts throughout the text whenever doing so does not compromise clarity.

where y_t is the exogenous income earned in period t, u is the utility of using the algorithm and p_t is the price of the algorithm in period t. The indicator function $\mathcal{I}_{j,t}$ takes the value one if household j buys the AI license and zero otherwise. The mass of AI users at time t is $\mu_t \equiv \int_0^N \mathcal{I}_{j,t} dj$. We now discuss the variables i_t and e_t.

Alignment and other problems In the introduction, we classify AI risks and misalignments into two types: internal and external. Internal risks and misalignments arise when AI algorithms manipulate households into making decisions that reduce their welfare. This effect, i_t, decreases momentary utility by i_t^2, which is measured in units of output.

External risks and misalignments occur when an AI algorithm affects a household indirectly through the use of the AI algorithm by other households. For example, AI-driven social media may polarize public opinion and distort election outcomes. This effect, e_t, reduces momentary utility by e_t^2, which is measured in units of output. This reduction is increasing in the number of users, μ_t.

Households can control internal risks and misalignments by choosing not to purchase the algorithm. In this section, we assume they account for the expected welfare reduction from the internal effect ($\mathbb{E}^s(i_t^2)$ in equation (1)) when making their purchase decision. In Section 5, we examine a scenario where behavioral biases lead households to overlook these internal effects when deciding whether to adopt the algorithm.

In contrast, households have no control over external misalignments and risks, as these depend on the adoption decisions made by other households.

Expectations of short- and long-run risks and misalignments We assume that the short-run impact of internal and external risks and misalignments on utility is equal to the long-run impact, ϕ_x^2, plus a mean-zero random variable, ξ_x for $x \in \{i, e\}$:

$$i_1^2 = \phi_i^2 + \xi_i, \qquad i_2^2 = \phi_i^2,$$
$$e_1^2 = (\phi_e^2 + \xi_e)\mu_1, \qquad e_2^2 = \phi_e^2\mu_2.$$

The random variables ϕ_x represent internal ($x = i$) and external ($x = e$) effects generated by the AI algorithm. Because welfare losses depend on ϕ_x^2, any deviation of ϕ_x from zero, whether positive or negative, is socially harmful, with larger realizations corresponding to more severe harms. The perceived distributions of ϕ_x can include very large realizations, corresponding to catastrophic events.

The random variables ξ_x capture the idea that the full consequences of AI usage may not be fully realized in the short run but emerge over the long run. The external effect in period t is proportional to μ_t, the number of households who use the technology in that period.

We allow developers to have different beliefs over the likelihood of misalignments than the rest of society. These differences in beliefs are assumed to be dogmatic. The superscript d denotes the developer's beliefs, $\mathbb{E}_t^d(\cdot)$, and the superscript s denotes societal beliefs, $\mathbb{E}_t^s(\cdot)$. For each $k = d, s$, we assume that the expected value of ϕ_x is zero for $x \in \{i, e\}$:

$$\mathbb{E}_1^k(\phi_x) = 0.$$

Let $\sigma_{k,x}^2$ denote uncertainty at time one about the algorithm's potential misalignment:

$$\sigma_{k,x}^2 = \mathbb{E}_1^k(\phi_x^2).$$

To assess internal and external risks and misalignments, the developer can release the algorithm to a sample of μ_1 users in period one. The outcomes from this partial release inform the developer's decision about a full-scale release in the second period.

We assume the probability of generating information from the initial release, denoted by $\pi(\mu_1)$, depends on the number of licenses μ_1,

$$\pi(\mu_1) = \begin{cases} (\mu_1/\kappa)^\alpha & \text{if } \mu_1 < \kappa, \\ 1 & \text{if } \mu_1 \geq \kappa. \end{cases}$$

Here, κ denotes the lowest number of participants required to obtain information with certainty. If $\kappa = N$, information is generated with probability one only when the algorithm is released to the whole population. The parameter $\alpha < 1$ determines the effectiveness of information generation. As $\alpha \to 0$, $\pi(\mu_1) \to 1$ if $\mu_1 > 0$ and $\pi(\mu_1) = 0$ if $\mu_1 = 0$. In this limiting case, testing on an infinitesimally small sample generates information with certainty.[3]

We define an indicator function $\mathcal{B}$ that takes the value one if information is generated and zero otherwise. If information is generated, a public signal about the algorithm's risks and misalignments becomes available. Rather than detailing the distributions of the random variables ξ_x, we model the effect of this information directly on posterior beliefs. In particular, the posterior beliefs about ϕ_x at the beginning of the second period become:

$$\mathbb{E}_2^k(\phi_x) = \hat{\phi}_{k,x}, \quad \text{VAR}_2^k(\phi_x) = \hat{\sigma}_{k,x}^2 < \sigma_{k,x}^2.$$

Beta testing reduces uncertainty but some residual uncertainty ($\hat{\sigma}_{k,x}^2 > 0$) remains, requiring decisions to be made under incomplete information. Posterior beliefs satisfy the consistency conditions:

$$\mathbb{E}_1^k(\hat{\phi}_{k,x}) = 0, \quad \mathbb{E}_1^k(\hat{\phi}_{k,x}^2 + \hat{\sigma}_{k,x}^2) = \sigma_{k,x}^2.$$

If no information is generated ($\mathcal{B} = 0$), initial priors remain unchanged for both the developer and society.

[3] To simplify the analysis, we abstract from fixed costs associated with beta testing. Introducing such costs would generate an inaction region in which some algorithms are neither released nor tested. This inaction region is different for the developer and the planner. For all other algorithms, however, both the unregulated equilibrium and the planner's problem would yield the same solutions described here.

Developer optimism Consistent with concerns that developers may underestimate societal harms, we assume that developers are more optimistic than society about the both the internal ($x = i$) and external ($x = e$) effects of the AI algorithm:

$$\mathbb{E}_t^d(\phi_x^2) \leq \mathbb{E}_t^s(\phi_x^2), \tag{2}$$

and

$$\mathbb{E}_1^d[\mathbb{E}_2^s(\phi_x^2)] \leq \mathbb{E}_1^s(\phi_x^2). \tag{3}$$

Household decisions Household j chooses whether to purchase an algorithm license in each period to maximize their expected lifetime utility, given by

$$\mathcal{U}_j = (1-\beta)v_{j,1} + \beta\mathbb{E}_1^s(v_{j,2}). \tag{4}$$

The household buys an AI license in period t if the expected private benefits, net of expected internal effects, exceed the algorithm's price. In period one, this condition is

$$u - \sigma_{s,i}^2 \geq p_1.$$

A similar condition applies in period two:

$$u - \mathbb{E}_2^s(\phi_i^2) \geq p_2.$$

The expected negative welfare consequences of internal misalignments ($\sigma_{s,i}^2$ in period one and $\mathbb{E}_2^s(\phi_i^2)$ in period two) reduce the price households are willing to pay for the algorithm in both periods.

The AI developer's problem There is a single AI developer who has designed an algorithm. In period one, the developer releases the algorithm to a fraction $\mu_1 \in [0, N]$ of the population. We assume that the decision to deploy the algorithm in period one can be reversed in period two. If this reversal occurs, the algorithm does not impact period two utility. The situation in which the release of the algorithm is irreversible is a special case of our model where $\beta = 0$.

The developer discounts the future at rate β. In each period, it chooses the price p_t and the number of licenses μ_t to maximize expected discounted utility, which is given by

$$\mathcal{V} = (1-\beta)\left[p_1\mu_1 - \chi\mathbb{E}_1^d(e_1^2)\right] + \beta\mathbb{E}_1^d\left[p_2\mu_2 - \chi\mathbb{E}_2^d(e_2^2)\right].$$

The parameter χ captures the extent to which developers also suffer from the externality. We consider two cases. When $\chi = 0$, developers are unaffected by the externality, while when $\chi = 1$, they experience the same disutility from the externality as an individual household does.

The developer's problem in period two At the beginning of period two, the developer decides whether to release the algorithm to the population, choosing the number of AI licenses to offer for sale (μ_2) and the price of each license (p_2). At the end of period two, uncertainty about internal and external misalignments is realized.

The developer's utility in the second period is,

$$\mathcal{V}_2 = \begin{cases} p_2\mu_2 - \chi\mathbb{E}_2^d(\phi_e^2)\mu_2, & \text{if } p_2 \leq u - \mathbb{E}_2^s(\phi_i^2) \text{ and } \mathcal{B} = 1, \\ p_2\mu_2 - \chi\sigma_{d,e}^2\mu_2, & \text{if } p_2 \leq u - \sigma_{s,i}^2 \text{ and } \mathcal{B} = 0, \\ 0, & \text{otherwise.} \end{cases}$$

where $p_2\mu_2$ is the developer's revenue when p_2 is sufficiently low for sales to be positive, that is, when p_2 is weakly below the household's utility net of the expected internality effect.

We assume that developers do not experience the algorithm's internal distortions directly, but are partially exposed to external harms through the parameter χ. The variables $\chi\mathbb{E}_2^d(\phi_e^2)\mu_2$ and $\chi\sigma_{d,e}^2\mu_2$ represent the expected externality borne by the developer with and without testing in period one, respectively.

The developer, acting as a monopolist, sets the price to extract the expected household surplus.[4] Whenever the developer markets the algorithm, it charges the highest price the household is willing to pay, that is, the utility of the algorithm net of the expected internality effect,

$$p_2 = \begin{cases} u - \mathbb{E}_2^s(\phi_i^2), & \text{if } \mathcal{B} = 1, \\ u - \sigma_{s,i}^2, & \text{if } \mathcal{B} = 0. \end{cases}$$

In period two, the developer releases the algorithm if the maximum price the household is willing to pay is greater than the reduction in the developer's utility caused by the externality associated with the algorithm, i.e., if $p_2 \geq \chi \mathbb{E}_2^d(\phi_e^2)$.

If new information is generated in period one ($\mathcal{B} = 1$) the posterior means of ϕ_x^2, for $x \in \{i, e\}$, are given by $\mathbb{E}_2^d(\phi_x^2) = \hat{\phi}_{d,x}^2 + \hat{\sigma}_{d,x}^2$. The developer releases the algorithm in the second period if the maximum price it can charge exceeds the reduction in utility caused by the external effect on the developer,

$$u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2) \geq \chi(\hat{\phi}_{d,e}^2 + \hat{\sigma}_{d,e}^2).$$

Otherwise, the algorithm is not released ($\mu_2 = 0$).

In the unregulated equilibrium, household expectations regarding internal risks and misalignments matter because they determine their willingness to pay for the algorithm, while the developer's expectations regarding internal risks are inconsequential.

Conversely, household expectations about external risks and misalignments have no impact. Instead, the developer's expectations about external effects matter because they determine the release decision.

If no information is generated in period one ($\mathcal{B} = 0$), the algorithm is released in period two to the whole population ($\mu_2 = N$) if

$$u - \sigma_{s,i}^2 \geq \chi \sigma_{d,e}^2,$$

[4] This pricing strategy generates no deadweight losses. It simply redistributes resources from the households to the monopolist.

and not released otherwise ($\mu_2 = 0$).

The optimized developer utility in period two is,

$$\mathcal{V}_2^* = \begin{cases} \max\{u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2) - \chi(\hat{\phi}_{d,e}^2 + \hat{\sigma}_{d,e}^2), 0\}N, & \text{if } \mathcal{B} = 1, \\ \max\{u - \sigma_{s,i}^2 - \chi\sigma_{d,e}^2, 0\}N, & \text{if } \mathcal{B} = 0. \end{cases}$$

The asterisk indicates that the value function is evaluated using the developer's optimal pricing and deployment strategy in period two.

To make the problem interesting, we assume that the distributions of ϕ_x are such that there is a strictly positive probability that both $u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2) > \chi(\hat{\phi}_{d,e}^2 + \hat{\sigma}_{d,e}^2)$, in which case the developer releases the algorithm, and $u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2) < \chi(\hat{\phi}_{d,e}^2 + \hat{\sigma}_{d,e}^2)$ in which case the algorithm is not released. This assumption means that the probability of the algorithm being implemented in period two is strictly positive but less than one.

The following proposition states that, from the perspective of period two, a higher value of μ_1 has a strictly positive value. The intuition for this result is that the developer is better off because it can make decisions based on the expected value of the information obtained from releasing the algorithm to μ_1 households.

Lemma 1 (Private benefits of releasing in period one)**.** *The developer's expected utility in the second period increases with the number of licenses sold in period* 1 *if* $\mu_1 < \kappa$.

$$\frac{d\mathbb{E}_1^d(\mathcal{V}_2^*)}{d\mu_1} > 0.$$

This lemma will be useful in characterizing the developer's optimal release policy in period one, to which we turn next.

The developer's problem in period one In period one, the developer chooses the number of licenses, μ_1, and the price per license, p_1. The developer's objective function is given by:

$$\mathcal{V} = (1-\beta)\left(\begin{cases} p_1\mu_1 - \chi\sigma_{d,e}^2\mu_1, & \text{if } p_1 \leq u - \sigma_{s,i}^2 \\ 0, & \text{if } p_1 > u - \sigma_{s,i}^2 \end{cases}\right) + \beta\mathbb{E}_1^d(\mathcal{V}_2^*).$$

The optimal price for the developer is the maximum price the household is willing to pay: $p = u - \sigma^2_{s,i}$.

If $u - \sigma^2_{s,i} \geq \chi\sigma^2_{d,e}$, it is optimal to release the algorithm to the entire population, $\mu_1 = N$. Instead, if $u - \sigma^2_{s,i} < \chi\sigma^2_{d,e}$, the expected benefit of releasing the algorithm in period one is negative. Still, it might be optimal to release the algorithm to part of the population to obtain information that can be used in period two (see Lemma 1). We call this type of experimentation *beta testing*. Since $\alpha < 1$, the developer's utility is increasing around $\mu_1 = 0$, so the optimal solution features positive beta testing: $\mu_1 > 0$. The intuition is that the informational benefits of testing rise sufficiently rapidly with μ_1 to outweigh the expected cost of testing, $(u - \sigma^2_{s,i} - \chi\sigma^2_{d,e})\mu_1$.

Proposition 1 summarizes the developer's optimal release policy. To describe this policy, it is useful to define the developer's information benefit-cost ratio, Λ^d:

$$\Lambda^d \equiv \frac{\beta}{1-\beta} \frac{\mathbb{E}^d_1\left[\max\left\{u - \mathbb{E}^s_2(\phi^2_i) - \chi\mathbb{E}^d_2(\phi^2_e), 0\right\}\right]}{\sigma^2_{s,i} + \chi\sigma^2_{d,e} - u}. \tag{5}$$

This ratio compares the expected future gains from information learned via beta testing, $\beta\mathbb{E}^d_1\left[\max\left\{u - \mathbb{E}^s_2(\phi^2_i) - \chi\mathbb{E}^d_2(\phi^2_e), 0\right\}\right]$, to the immediate cost to the developer of selling the AI algorithm to an additional person today. This cost is the external effect on the developer minus the sale price of the algorithm, $(1-\beta)[\chi\sigma^2_{d,e} - (u - \sigma^2_{s,i})]$, which is positive in the region where beta testing is relevant.

Proposition 1 (Uncertainty, beta testing, and algorithm release)**.** *In an unregulated equilibrium, the developer's release policy depends on the perceived external risk and the information benefit-cost ratio. The equilibrium has the following properties:*

1. *If uncertainty about external effects is low, $\chi\sigma^2_{d,e} \leq u - \sigma^2_{s,i}$, the developer foregoes beta testing and releases the AI algorithm to the entire population in the first period ($\mu_1 = N$).*

2. *If uncertainty about external effects is high, $\chi\sigma^2_{d,e} > u - \sigma^2_{s,i}$, then the developer beta*

tests the algorithm on a sample of size,

$$\mu_1 = \min\left\{\left[\alpha\Lambda^d \frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}}, 1\right\}\kappa. \tag{6}$$

The developer may opt to withdraw the product from the market even when the expected misalignment, $\hat{\phi}_{k,x}$, is relatively small in absolute value, provided that the residual uncertainty about the actual size of the misalignment, $\hat{\sigma}^2_{k,x}$, remains substantial.

3.2 The first-best solution (planner's problem)

Following Weyl (2007) and Sandroni and Squintani (2007), we evaluate welfare using the planner's beliefs.

In the first period the planner chooses the number of households that can use the algorithm. When this number is positive, the planner may obtain information about the algorithm's internal and external effects. In the second period, the planner decides whether to make the algorithm available and how many licenses to issue.

We define social welfare as the sum of household and developer utility, $\int_0^N \mathcal{U}_j dj + \mathcal{V}$. Because utility is quasi-linear, maximizing social welfare is equivalent to maximizing efficiency, and any desired distribution of utility can be implemented through lump-sum transfers. As a result, prices, which simply transfer resources between households and the developer, do not affect social welfare.

To compute the socially optimal allocations, we describe the solution to the second-period problem, contingent upon the choices made in the first period about μ_1.

The planner's problem in period two The expected social welfare in the second period, considering the available information, is given by:

$$\mathcal{W}_2 = \begin{cases} Ny_2 + \left[u - (\hat{\phi}^2_{s,i} + \hat{\sigma}^2_{s,i}) - (N+\chi)\left(\hat{\phi}^2_{s,e} + \hat{\sigma}^2_{s,e}\right)\right]\mu_2, & \text{if } \mathcal{B} = 1, \\ Ny_2 + \left[u - \sigma^2_{s,i} - (N+\chi)\,\sigma^2_{s,e}\right]\mu_2, & \text{if } \mathcal{B} = 0. \end{cases}$$

We now determine the optimal value of μ_2. If $\mathcal{B} = 1$, the posterior expectation is given by $\mathbb{E}_2^s(\phi_x^2) = \hat{\phi}_{s,x}^2 + \hat{\sigma}_{s,x}^2$. In this case, releasing the algorithm is optimal if

$$\frac{u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2)}{N + \chi} \geq \hat{\phi}_{s,e}^2 + \hat{\sigma}_{s,e}^2,$$

ensuring that the household's utility from using the algorithm, net of expected internal effects, exceeds the external effect. If this condition is not satisfied, then $\mu_2 = 0$.

If $\mathcal{B} = 0$, then $\mu_2 = N$ if

$$\frac{u - \sigma_{s,i}^2}{N + \chi} \geq \sigma_{s,e}^2,$$

and otherwise $\mu_2 = 0$.

In period two, the planner only releases AI algorithms when expected social benefits exceed expected social costs, including the external harms imposed on households and developers, $(N + \chi)\mathbb{E}_2^s(\phi_e^2)$. By contrast, the developer considers only its own expected exposure to the externality, $\chi\mathbb{E}_2^d(\phi_e^2)$. As a result, the developer may commercialize algorithms expected to be socially harmful.

The social welfare in period two is given by:

$$\mathcal{W}_2^* = Ny_2 + \begin{cases} \max\{u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2) - (N + \chi)(\hat{\phi}_{s,e}^2 + \hat{\sigma}_{s,e}^2), 0\}N, & \text{if } \mathcal{B} = 1 \\ \max\{u - \sigma_{s,i}^2 - (N + \chi)\sigma_{s,e}^2, 0\}N, & \text{if } \mathcal{B} = 0, \end{cases}$$

where the asterisk indicates that the value function has been maximized with respect to the choice of deployment in period two.

We assume that there is a strictly positive probability that $u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2) > (N + \chi)(\hat{\phi}_{s,e}^2 + \hat{\sigma}_{s,e}^2)$, in which case it is optimal to release the algorithm, and $u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2) < (N + \chi)(\hat{\phi}_{s,e}^2 + \hat{\sigma}_{s,e}^2)$, in which case it is not. This assumption means that the probability that the planner releases the algorithm in the second period, given the information obtained in the first period, is strictly positive but less than one.

The equivalent of Lemma 1 for the planner is as follows.

Lemma 2 (Social benefits of beta testing in period one)**.** *Expected social welfare in the second period increases with the size of the sample used for beta testing in the first period for $\mu_1 < \kappa$:*

$$\frac{d\mathbb{E}_1^s(\mathcal{W}_2^*)}{d\mu_1} > 0.$$

This lemma is useful in characterizing the planner's optimal release policy in period one, to which we turn next.

The planner's problem in period one Expected social welfare is given by

$$\mathcal{W} = (1-\beta)\left[Ny_1 + \left\{u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2\right\}\mu_1\right] + \beta\mathbb{E}_1^s(\mathcal{W}_2^*).$$

Absent the informational benefits of experimentation, it would be optimal to set $\mu_1 = 0$ if $u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2 < 0$. However, beta testing in the first period ($\mu_1 > 0$) creates value by generating information that the planner can use in the second period (see Lemma 2).

Proposition 2 summarizes the planner's optimal release policy. To describe this solution, it is useful to define the planner's information benefit-cost ratio, Λ^s:

$$\Lambda^s \equiv \frac{\beta}{1-\beta}\frac{\mathbb{E}_1^s\left[\max\left\{u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2), 0\right\}\right]}{\sigma_{s,i}^2 + (N+\chi)\sigma_{s,e}^2 - u}. \tag{7}$$

The planner's information benefit-cost ratio is lower than the developer's ($\Lambda^s < \Lambda^d$) for two reasons. First, the developer does not take into account external effects on the population. Second, because the developer is optimistic, the expected value of external damages are larger for the planner.

Proposition 2 (Uncertainty, beta testing, and algorithm release)**.** *In the first best, the number of user licenses offered in the first period depends on the perceived external risk, the effectiveness of beta testing, and the information benefit-cost ratio. The solution is as follows:*

1. *If uncertainty is low, $\sigma_{s,e}^2 \leq \frac{u-\sigma_{s,i}^2}{N+\chi}$, the planner always foregoes beta testing and releases the AI algorithm to the entire population in the first period ($\mu_1 = N$)*

2. *If uncertainty is high $\sigma^2_{s,e} > \frac{u-\sigma^2_{s,i}}{N+\chi}$, the planner beta tests the algorithm on a sample of size*

$$\mu_1 = \min\left\{\left[\alpha\Lambda^s\frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}}, 1\right\}\kappa. \tag{8}$$

Figure 1 depicts the optimal values of μ_1 (Panel A) and μ_2 (Panel B) chosen by the planner and the developer for various levels of uncertainty. In the first period, both the developer and the planner agree to release the algorithm to the entire population when uncertainty about the external effect is low ($\sigma^2_{s,e} \leq \frac{u-\sigma^2_{s,i}}{N+\chi}$). At higher levels of uncertainty, the planner is more cautious than the developer, releasing the algorithm to fewer users. The reason is that the planner takes into account the impact of the externalities generated by the algorithm on the entire population and on the developer.

In the second period, the developer and planner make the same release decisions when external effects are low or high. However, when external effects are in an intermediate range, they disagree: the developer opts to release the algorithm, whereas the planner chooses not to. This disparity arises because the developer internalizes only its own exposure (χ) to external harms, whereas the planner accounts for harms imposed on the entire population.

4 Regulating AI

Social welfare in the unregulated equilibrium falls short of the social optimum because developers overlook AI's external impact on households. Aligning private and social incentives through Pigouvian taxes is challenging because of uncertainty and disagreement over AI's external effects.

We consider two types of Pigouvian taxes: ex-post taxes, which hold developers liable for realized damages, and ex-ante taxes, which depend on expected external effects. Ex-post taxes implement the first best under homogeneous beliefs but fail under heterogeneous beliefs because optimistic developers underestimate expected liabilities. Ex-ante taxes can restore efficiency when beliefs are observable and contractible, but fail when developer beliefs are private information because developers have incentives to feign pessimism to reduce tax liabilities. We show that, under private information, the optimal policy combines mandatory beta testing with regulatory approval contingent on beta-test outcomes.

4.1 Pigouvian taxes

We first consider ex-post Pigouvian taxes in an economy with homogeneous beliefs. Throughout, we assume that any tax revenue is redistributed to households as lump-sum transfers.

4.1.1 Ex-post Pigouvian taxes

Suppose that regulators levy liability taxes on developers equal to the welfare cost of the realized external effects imposed on the households:

$$T_t = N \times e_t^2. \tag{9}$$

The developer understands that selling the AI to a larger population increases expected tax liabilities because realized external damages scale with the number of users: $\mathbb{E}_t^d(T_t) = N\mathbb{E}_t^d(\phi_e^2)\mu_t$. The welfare properties of these taxes are summarized by the following proposition.

Proposition 3 (Ex-Post Pigouvian Taxes under Homogeneous Beliefs)**.** *Suppose that developers and society have the same beliefs. Then, the ex-post Pigouvian taxes in equation* (9) *align private and social incentives. As a result, the developer's decisions regarding testing, deployment, and innovation are socially optimal.*

Proof. If beliefs are homogeneous, then $\mathbb{E}_t^d(\phi_e^2) = \mathbb{E}_t^s(\phi_e^2)$, and for notational convenience we drop the indices s and d. It is still optimal for the developer to set

$p_t = u - \mathbb{E}_t^s(\phi_i^2)$. Replacing this price and the expected taxes into the utility of the developer, we find that $\mathcal{V} = (1-\beta)\mathcal{V}_1 + \beta\mathbb{E}_1(\mathcal{V}_2)$, where

$$\mathcal{V}_1 = [u - \sigma_i^2 - (N+\chi)\sigma_e^2]\mu_1$$
$$\mathcal{V}_2 = \begin{cases} [u - (\hat{\phi}_i^2 + \hat{\sigma}_i^2) - (N+\chi)(\hat{\phi}_e^2 + \hat{\sigma}_e^2)]\mu_2, & \text{if } \mathcal{B} = 1, \\ [u - \sigma_i^2 - (N+\chi)\sigma_e^2]\mu_2, & \text{if } \mathcal{B} = 0. \end{cases}$$

It follows that $\mathcal{V}_t = \mathcal{W}_t - Ny_t$. Private and social incentives are aligned, so privately optimal decisions coincide with the social optimum. □

Heterogeneous beliefs When the developer and society hold different beliefs about the algorithm's potential risks, ex-post Pigouvian taxes fail to align the developer's incentives with those of society. This divergence in incentives arises because the developer, being more optimistic, assigns a lower probability to negative external effects and is therefore more inclined to release the algorithm than the planner.

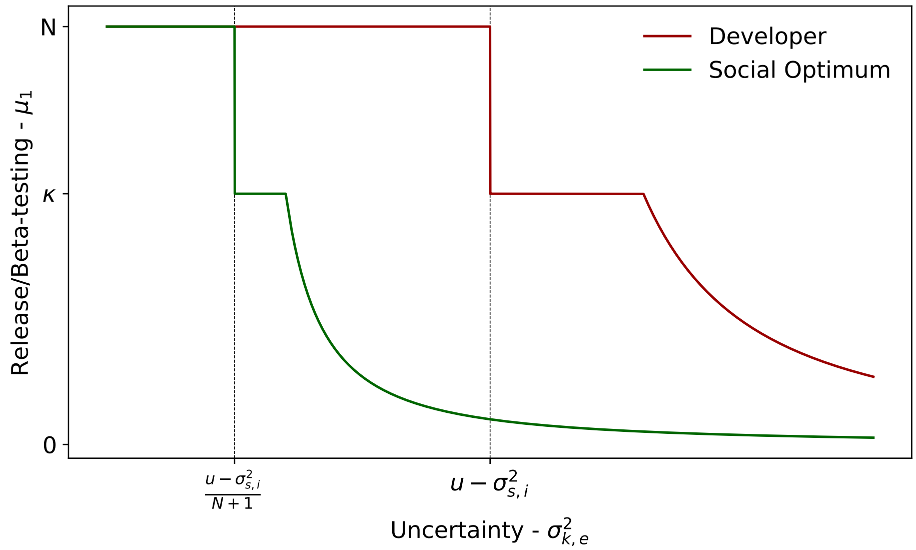
To see this result formally, note that in period t, the developer's expectation of their tax liability is lower than the planner's $\mathbb{E}_t^d(T_t) = N\mathbb{E}_t^d(\phi_e^2)\mu_t < N\mathbb{E}_t^s(\phi_e^2)\mu_t$. It follows that

$$\mathcal{V}_t = \mathcal{W}_t - Ny_t + (N+\chi)\left\{\mathbb{E}_t^s(\phi_e^2) - \mathbb{E}_t^d(\phi_e^2)\right\}\mu_t$$

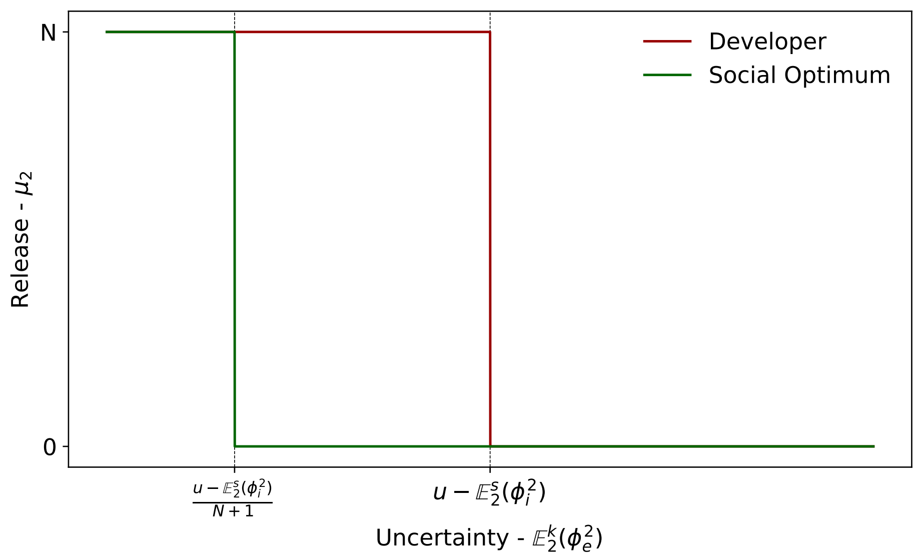
This expression shows that the developer is willing to release the algorithm in cases where the planner would not. In the extreme case where $\mathbb{E}_t^d(\phi_e^2) = 0$, the developer is always willing to release the algorithm, while the planner is more cautious.

Figure 1: Release decisions in the first and second periods

Panel A: Time 1



Panel B: Time 2



4.1.2 Ex-ante Pigouvian taxes

We now show that ex-ante Pigouvian taxes can achieve the efficient outcome when beliefs are heterogeneous, but only if beliefs are publicly known and contractible.

Suppose the regulator sets taxes at the beginning of each period based on the expected external damages caused by the algorithm:

$$T_t^{ex-ante} = \mathbb{E}_t^s(Ne_t^2) = N\mathbb{E}_t^s(\phi_e^2)\mu_t. \tag{10}$$

When beliefs are heterogeneous, the ex-ante taxes in equation (10) do not implement the social optimum because we need to correct for differences in beliefs. For instance, the taxes in period two need to be corrected as follows:

$$T_2^{ex-ante} = \begin{cases} N\mathbb{E}_2^s(\phi_e^2)\mu_2 + \chi\left[\mathbb{E}_2^s(\phi_e^2) - \mathbb{E}_2^d(\phi_e^2)\right]\mu_2 & \text{if } \mathcal{B} = 1, \\ N\sigma_{s,e}^2\mu_2 + \chi\left(\sigma_{s,e}^2 - \sigma_{d,e}^2\right)\mu_2 & \text{if } \mathcal{B} = 0. \end{cases} \tag{11}$$

The first term internalizes expected external harms on households based on the regulator's beliefs. The second term corrects for differences between the regulator's and the developer's expectations regarding the developer's own exposure to these external harms.

Note that the developer pays a higher tax when they are relatively optimistic (lower $\mathbb{E}_t^d(\phi_e^2)$), and a lower tax when they are relatively pessimistic (higher $\mathbb{E}_t^d(\phi_e^2)$).[5]

However, the analog of taxes (11) applied to period one do not implement the socially optimal level of μ_1. The reason is that the developer's information benefit-cost ratio is higher than that of the regulator,

$$\Lambda^d = \Lambda^s \times \frac{\mathbb{E}_1^d\left[\max\{u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2), 0\}\right]}{\mathbb{E}_1^s\left[\max\{u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2), 0\}\right]} > \Lambda^s. \tag{12}$$

[5] Laffont (1977) explores a similar result in a version of Weitzman (1974)'s model in which the firm and the planner have different expectations about fundamentals. When prices are used as incentives, they must correct for belief differences. As we discuss below, in our model, there is a dynamic element to the belief correction because information generated by beta testing or releasing at time one has value at time two.

To implement the first-best outcome, period-one taxes must correct not only for differences in expectations about the externality in that period but also for differences in expectations regarding the externality in period two. This dynamic correction is necessary because beta testing in period one generates information that affects release decisions in period two. The period-one taxes that achieve the first-best are given by:

$$
\begin{aligned}
T_1^{ex-ante} = {} & N\mathbb{E}_1^s(\phi_e^2)\mu_1 + \chi\left[\mathbb{E}_1^s(\phi_e^2) - \mathbb{E}_1^d(\phi_e^2)\right]\mu_1 \\
& + \frac{\beta}{1-\beta}\pi(\mu_1)N\Bigg(\mathbb{E}_1^d\left[\max\{u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2), 0\}\right] \\
& - \mathbb{E}_1^s\left[\max\{u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2), 0\}\right]\Bigg), \qquad (13)
\end{aligned}
$$

Proposition 4 (Ex-Ante Pigouvian Taxes under Heterogeneous Beliefs)**.** *Suppose the developer and society have different beliefs about the algorithm's potential external effects. If the regulator imposes the ex-ante Pigouvian taxes specified in equations* (11) *and* (13)*, the developer's decisions regarding release, withdrawal, and the sample size used in beta testing align with the socially optimal outcomes.*

Unfortunately, this policy is impractical because the developer's expectations required to construct the taxes (11) and (13) are generally unobservable. What is the optimal policy when the developer's beliefs are private information? If the regulator applies the taxes (11) and (13) to the developer's self-reported beliefs, then the developers have an incentive to pretend to be more pessimistic than they are to pay lower taxes.

4.2 Optimal policy when beliefs are private information

In this subsection, we study the optimal policy in a setting where developer beliefs are private information. Our main result is that, under these informational constraints, Pigouvian taxes fail to implement the social optimum. We show that the first best can be implemented by combining mandatory beta testing with regulatory approval.

The developer draws their belief type θ from a set $\Theta = [\underline{\theta}, \overline{\theta}]$, with cumulative density function F. We denote by $\underline{\theta}$ and $\overline{\theta}$ the most pessimistic and optimistic beliefs in the set Θ. We assume that the most optimistic developer assigns zero probability to non-zero external effects, so $\mathbb{E}_1^{\overline{\theta}}(\phi_e^2) = 0$. The most pessimistic developer has the same expectations as society $\mathbb{E}_1^{\underline{\theta}}(\phi_e^2) = \mathbb{E}_1^{s}(\phi_e^2)$. Under these conditions, developers are weakly more optimistic than society.

We consider two sets of policies without commitment: (i) a linear tax on the number of licenses sold and (ii) regulation that controls release and beta testing decisions. Timing is as follows. First, nature draws the developer's belief type θ, which is private information. Next, the regulator designs the optimal policy for period one. Given this policy, the developer decides their optimal price and release strategy. In the second period, all individuals observe the beta test results from period one, and beliefs are updated. The regulator then chooses the optimal policy for period two, after which the developer decides their optimal release strategy.

As is standard in the Pigouvian taxation literature, we model taxes as linear functions of the number of licenses sold, μ_t. Given that developer type is private information, taxes cannot depend directly on developers' beliefs. Consequently, a tax policy is defined as a set of state-contingent tax rates per unit sold $\boldsymbol{\tau} \equiv \{\tau_t\}$, for $t = 1, 2$. So total taxes are $T_t = \tau_t \mu_t$.

The second policy framework we consider consists of mandatory beta testing and regulatory approval. These policies set limits on the maximum number of licenses that can be sold in each period.

In period one, the regulator can either mandate a beta test involving up to $\overline{\mu}_1$ users or permit full release, in which case $\overline{\mu}_1 = N$. In period two, the regulator can approve full commercialization by setting $\overline{\mu}_2 = N$, require the developer to withdraw the algorithm by setting $\overline{\mu}_2 = 0$, or approve a number of users, $\overline{\mu}_2$, strictly between 0 and N.

A mandatory beta-testing and regulatory approval policy is a set of state-contingent restrictions on license sales, denoted by $\overline{\boldsymbol{\mu}} \equiv \{\overline{\mu}_t\}$ for $t = 1, 2$. These restrictions impose an upper limit on number of licenses that can be sold in each period such that $\mu_t \leq \overline{\mu}_t$.[6]

Optimal tax policy For any tax policy $\boldsymbol{\tau}$, the developer's beta testing and release strategies follow the same policy as before. Lemma 3 characterizes the developer's behavior for an arbitrary tax policy.

Lemma 3 (Optimal Developer Behavior Under Tax Policy)**.** *For any tax policy $\boldsymbol{\tau}$, the developer's release, beta-testing and withdrawal policies are as follows.*
At time $t = 2$:

1. *If $\chi \mathbb{E}_2^\theta(\phi_e^2) < u - \mathbb{E}_2^s(\phi_i^2) - \tau_2$, the developer releases the algorithm to the entire population, $\mu_2^\theta = N$.*

2. *If $\chi \mathbb{E}_2^\theta(\phi_e^2) \geq u - \mathbb{E}_2^s(\phi_i^2) - \tau_2$, the developer withdraws the algorithm, $\mu_2^\theta = 0$.*

At time $t = 1$:

1. *If $\chi \sigma_{\theta,e}^2 < u - \sigma_{s,i}^2 - \tau_1$, the developer foregoes beta testing and releases the algorithm to the entire population, $\mu_1^\theta = N$.*

2. *If $\chi \sigma_{\theta,e}^2 \geq u - \sigma_{s,i}^2 - \tau_1$, the developer beta tests the algorithm on*

$$\mu_1^\theta(\tau_1, \tau_2) = \min\left\{\left[\alpha \Lambda^\theta(\tau_1, \tau_2) \frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}}, 1\right\} \kappa, \tag{14}$$

where $\Lambda^\theta(\tau_1, \tau_2)$ is the developer's information benefit-to-cost ratio defined in equation (A.9).

[6]In the appendix, we discuss the case of non-linear taxes. We show that there is a non-linear, discontinuous tax schedule that replicates the outcomes obtained under the optimal mandatory beta-testing and regulatory approval policy. This tax schedule confiscates the developer's revenues when their choices deviate from the efficient allocations.

Taxes affect the release decision at time two and the sample size used for beta testing at time one.

Proposition 5 characterizes the optimal tax policy when beliefs are private information.

Proposition 5 (Optimal Tax Policy When Beliefs Are Private Information)**.** *The optimal tax policy is as follows.*
At time $t = 2$:

1. *If uncertainty is small, $\mathbb{E}_2^s(\phi_e^2) \leq \frac{u - \mathbb{E}_2^s(\phi_i^2)}{N+\chi}$, the regulator sets the tax to zero $\tau_2 = 0$ and the developer releases the algorithm to the entire population, $\mu_2^\theta = N$.*
2. *If uncertainty is large, $\mathbb{E}_2^s(\phi_e^2) > \frac{u - \mathbb{E}_2^s(\phi_i^2)}{N+\chi}$, the regulator sets the tax $\tau_2 = p_2$ and the developer withdraws the algorithm, $\mu_2^\theta = 0$.*

At time $t = 1$:

1. *If uncertainty is small, $\sigma_{s,e}^2 \leq \frac{u - \sigma_{s,i}^2}{N+\chi}$, the regulator sets the tax to zero $\tau_1 = 0$ and the developer releases the algorithm to the entire population, $\mu_1^\theta = N$ for all $\theta \in \Theta$.*
2. *If uncertainty is large, $\sigma_{s,e}^2 > \frac{u - \sigma_{s,i}^2}{N+\chi}$, the regulator sets the tax so that*
$$\int_{\underline{\theta}}^{\bar{\theta}} \frac{d\mu_1^\theta / d\tau_1}{\int_{\underline{\theta}}^{\bar{\theta}} d\mu_1^\theta / d\tau_1 dF(\theta)} \pi'(\mu_1^\theta) \Lambda^s N dF(\theta) = 1, \tag{15}$$
and the developer sells μ_1^θ licenses as given in equation (14).

In the second period, the tax is set to zero if the algorithm is socially beneficial. If the algorithm is socially harmful, all revenue is taxed away, ensuring that developers of all types choose not to release the algorithm. This tax policy ensures that the release of the algorithm in period two is optimal.

In the first period, the planner chooses a tax rate that equates the expected value of beta testing across developer types with the welfare cost of conducting beta testing. Because this tax policy adjusts developer incentives based on average values across different belief types rather than individual beliefs, it generally falls short of achieving the efficient outcome.

Next, we describe the implementation of the efficient outcome using mandatory beta testing and regulatory approval of the algorithm, conditional on the results of the beta test.

Optimal regulation policy Lemma 4 characterizes the developer's behavior for an arbitrary policy regulating beta testing and release.

Lemma 4 (Optimal Developer Behavior Under Regulation Policy)**.** *For any regulation policy $\overline{\boldsymbol{\mu}}$, the developer's release, beta-testing and withdrawal policies are as follows.*
At time $t = 2$:

1. *If $\chi \mathbb{E}_2^\theta(\phi_e^2) < u - \mathbb{E}_2^s(\phi_i^2)$, the developer sells the maximum number of lisences allowed, $\mu_2^\theta = \overline{\mu}_2$.*
2. *If $\chi \mathbb{E}_2^\theta(\phi_e^2) \geq u - \mathbb{E}_2^s(\phi_i^2)$, the developer withdraws the algorithm, $\mu_2^\theta = 0$.*

At time $t = 1$:

1. *If $\chi \sigma_{\theta,e}^2 < u - \sigma_{s,i}^2$, the developer sells the maximum number of licenses allowed, $\mu_1^\theta = \overline{\mu}_1$.*
2. *If $\chi \sigma_{\theta,e}^2 \geq u - \sigma_{s,i}^2$, the developer beta tests the algorithm on*

$$\mu_1^\theta = \min\left\{ \min\left\{ \left[\alpha \Lambda^\theta \frac{N}{\kappa} \right]^{\frac{1}{1-\alpha}}, 1 \right\} \kappa, \overline{\mu}_1 \right\}, \tag{16}$$

where Λ^θ is the developer's information benefit-to-cost ratio given by equation (A.17).

Proposition 6 characterizes the optimal regulation policy when beliefs are private information.

Proposition 6 (Optimal regulation policy under private information)**.** *The optimal regulation policy is as follows. At time $t = 2$:*

1. *If uncertainty is low, $\mathbb{E}_2^s(\phi_e^2) \leq \frac{u - \mathbb{E}_2^s(\phi_i^2)}{N+\chi}$, the regulator sets a non-binding limit on the number of licenses $\overline{\mu}_2 = N$ and the developer releases the algorithm to the entire population, $\mu_2^\theta = N$ for all θ.*

2. *If uncertainty is high, $\mathbb{E}_2^s(\phi_e^2) > \frac{u - \mathbb{E}_2^s(\phi_i^2)}{N+\chi}$, the regulator mandates the withdrawal of the algorithm, setting $\overline{\mu}_2 = 0$ and so $\mu_2^\theta = 0$ for all θ.*

At time $t = 1$:

1. *If uncertainty is low, $\sigma_{s,e}^2 \leq \frac{u - \sigma_{s,i}^2}{N+\chi}$, the regulator sets a non-binding limit on the number of licenses $\overline{\mu}_1 = N$ and the developer releases the algorithm to the entire population, $\mu_1^\theta = N$ for all θ.*

2. *If uncertainty is high, $\sigma_{s,e}^2 > \frac{u - \sigma_{s,i}^2}{N+\chi}$, the regulator sets the following upper bound*

$$\overline{\mu}_1 = \min\left\{\left[\alpha \Lambda^s \frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}}, 1\right\}\kappa. \tag{17}$$

   The developer sells $\mu_1^\theta = \overline{\mu}_1$ licenses for all θ.

This regulation policy implements the socially optimal allocation in this economy, so it is superior to the best tax policy.[7] Overall, the optimal regulatory policy follows a simple threshold rule: intervention occurs only when uncertainty is sufficiently high. When uncertainty is low, the regulator permits full deployment. When uncertainty is high, the regulator limits beta testing in period one to balance informational gains against expected external harms and conditions period-two approval

[7] It is possible to formulate a discontinuous tax policy that replicates the effects of the optimal regulatory framework. This policy involves granting a subsidy to developers who select the socially optimal value of μ_1. The subsidy can be set at a sufficiently high level to ensure that all types of developers opt for the socially optimal μ_1.

on the information revealed through testing. This framework prevents premature large-scale deployment while preserving socially valuable experimentation.

It is useful to briefly consider the special case in which releasing the algorithm in period one cannot be reversed in period two. This setting corresponds to a special case of our model in which $\beta = 0$. In this case, the optimal policy is simply to allow the release of algorithms such that $u - \sigma_{s,i}^2 \geq (N + \chi)\sigma_{s,e}^2$ and forbid the release of all other algorithms.

We now briefly discuss the impact of limited liability on our results.

4.3 Limited liability

In practice, limited liability protects developers from bearing the full cost of large social damages. The model analyzed here abstracts from this consideration. In the Appendix, we analyze how this constraint affects our results, focusing, for simplicity, on the case of homogeneous beliefs. We model limited liability by assuming that taxes in period t cannot exceed the sales revenue generated in that period. Under this constraint, ex-post taxes do not achieve the first-best allocation in either period one or two because the developer does not take into account damages that exceed their limited liability. In contrast, ex-ante taxes succeed in implementing the first-best allocation in period two but fail to do so in period one. The regulatory policy described in Proposition 6 remains effective in that model.

5 A model with internalities

This section extends the benchmark model to incorporate internalities arising from behavioral biases. These biases lead households to make decisions that are not in their self-interest because of misinformation, self-control issues, cognitive biases, or time inconsistency problems, all of which can be exploited by AI algorithms.

For example, recent experimental evidence suggests that reliance on Large Language Models may generate cognitive internalities. Users are drawn to interfaces that minimize immediate effort and maximize short-run performance, while neglecting the long-run effects on memory formation, cognitive engagement, and autonomy induced by repeated reliance on the tool (Kosmyna et al., 2025).

5.1 Unregulated equilibrium

Household's problem In Section 3, we assume that households take the expected welfare reduction caused by internal effects, $\mathbb{E}_t^s(i_t^2)$, into account when deciding whether to use the algorithm. Here, we consider the case in which, due to behavioral biases, households disregard these internal effects when making their purchase decisions.

We formalize this idea by assuming that $\mathcal{U}_j$, defined in equation (4), is the household's "experienced utility," but that households base their choices on a different, misspecified, objective function that we refer to as the "decision utility."[8] Lifetime decision utility takes the form:

$$\mathcal{U}_j^b = (1-\beta)v_{j,1}^b + \beta \mathbb{E}_1^s(v_{j,2}^b),$$

where momentary decision utility is

$$v_{j,t}^b = y_t + (u - p_t) \times \mathcal{I}_{j,t} - \mathbb{E}_t^s(e_t^2). \tag{18}$$

The household decides whether to purchase the AI algorithm to maximize $\mathcal{U}_j^b$. As a result, households purchase the algorithm whenever $p_t \leq u$. Recall that without behavioral biases, the decision rule is to buy the algorithm when $p_t \leq u - \mathbb{E}_t(\phi_i^2)$.

We assume that the developer is immune to the algorithm's internal effects, because it does not use the algorithm or is more sophisticated than the households.[9]

[8] This terminology is common in behavioral price theory, see, e.g., Farhi and Gabaix (2020).

[9] Extending our analysis to the case where the algorithm's internal effects also affect the developer is straightforward. Such an extension would not significantly alter our findings.

What are the key differences between this model and our benchmark model? Because households ignore expected negative internal effects on utility, the developer can charge them a higher price: $p_t = u$ instead of $p_t = u - \mathbb{E}_t^s(\phi_i^2)$.

Internalities widen the gap between the unregulated equilibrium and the social optimum because households no longer discipline deployment through lower willingness to pay. In period one, the developer beta tests the algorithm when $\chi\sigma_{d,e}^2 > u$ and releases the algorithm otherwise. In contrast, the planner has a lower threshold for the level of uncertainty required for beta testing. It is socially optimal to beta test whenever $\sigma_{s,e}^2 > (u - \sigma_{s,i}^2)/(N + \chi)$.

In period two, the developer withdraws the algorithm only when $\chi(\hat{\phi}_{d,e}^2 + \hat{\sigma}_{d,e}^2) > u$. The planner uses a lower uncertainty threshold for withdrawal. It is socially optimal to withdraw the algorithm whenever $\hat{\phi}_{s,e}^2 + \hat{\sigma}_{s,e}^2 > (u - \hat{\phi}_{s,i}^2 - \hat{\sigma}_{s,i}^2)/(N + \chi)$.

The developer partially internalizes external harms through the parameter χ, since these harms also affect the developer. By contrast, the developer does not internalize internalities because it neither experiences them directly nor faces lower willingness to pay from biased households. As a result, internalities create a larger wedge between private and social incentives than externalities.

Figure 2 shows that internalities amplify excessive deployment in the unregulated equilibrium. Under moderate and high uncertainty, the developer chooses larger deployment levels than the planner because households fail to incorporate internal harms into their adoption decisions. As a result, the divergence between private and socially optimal release decisions is larger than in the benchmark model.

The qualitative policy implications remain unchanged in the presence of internalities. Pigouvian taxes continue to fail to implement the first-best allocation under heterogeneous beliefs. Mandatory beta testing combined with regulatory approval contingent on beta-test outcomes can implement the socially optimal allocation.

6 AI Regulation in Practice

Our analysis focuses on two frictions that we view as central to AI regulation: uncertainty about societal risks and systematic differences in beliefs about these risks between developers and regulators. When beliefs diverge and are private information, neither ex-ante nor ex-post Pigouvian taxes generally implement the social optimum. Ex-post taxes perform poorly because optimistic developers underestimate expected damages, while ex-ante taxes require regulators to condition policy on developers' privately held beliefs. In this environment, quantity regulation through mandatory beta testing and regulatory approval can implement the social optimum.

Of course, belief heterogeneity is not the only friction that matters in practice. One potentially important consideration is limited liability, which we analyze in Appendix B. When liability is capped, ex-post damages lose much of their deterrent power, since harms beyond the cap are not borne by the developer. Limited liability therefore strengthens the case for ex-ante interventions. Ex-ante instruments, such as mandatory insurance, ex-ante taxes, or testing and regulatory approval, are more effective and can come closer to implementing the social optimum. This point has been emphasized in the literature on liability and regulation under imperfect financial markets. Tirole (2010) shows that limited liability fundamentally alters the optimal design of Pigouvian taxation and can justify regulatory interventions that extend beyond standard ex-post liability.

Another potentially relevant friction is information asymmetry. Our analysis relates to the literature on regulating privately informed agents (e.g., Baron and Myerson, 1982, Laffont and Tirole, 1986, and Laffont and Tirole, 1993). In our setting, however, the central regulatory problem is not how to extract hidden information from an informed developer. Instead, it is how to design a mechanism that encourages both the regulator and the developer to learn about uncertain risks while limiting the harms that may arise during that learning process.

This shared uncertainty is what makes beta testing valuable in our framework. Beta testing is not used to screen developer types, as in the standard mechanism design tradition. Rather, it serves to generate new public information that neither the regulator nor the developer initially possesses. Heterogeneous priors across developers capture precisely this feature: differences in beliefs about risks do not arise from strategic misrepresentation, but from honest disagreement under uncertainty.

Another potentially important friction concerns the measurement and attribution of social harms. Ex-post damages may be difficult to quantify and link to a specific model or developer, weakening the effectiveness of regulatory regimes that rely on ex-post enforcement. These informational constraints reduce deterrence and tilt the balance toward ex-ante interventions, including precautionary requirements imposed before deployment (see Kolstad and Ulen, 1983).

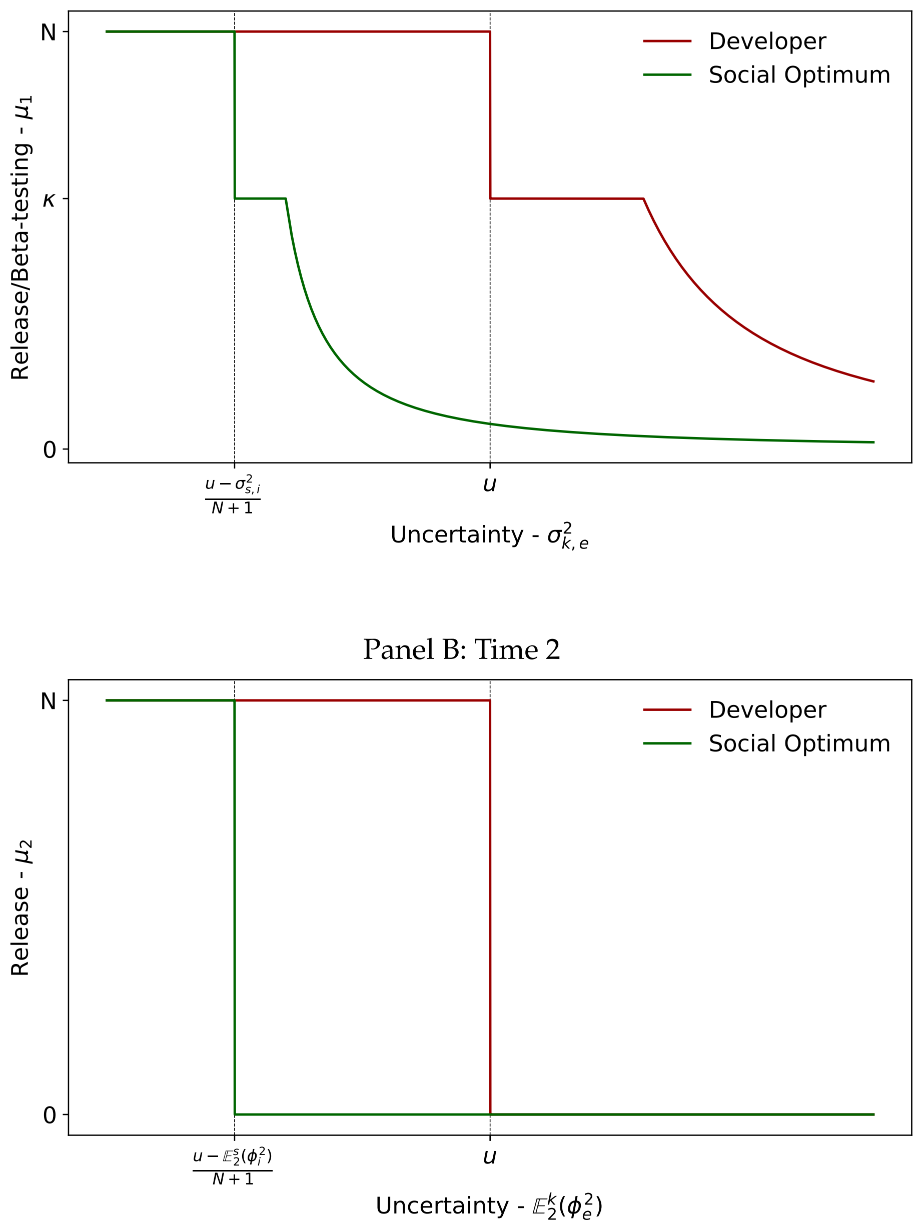
The optimal regulatory framework depends on which frictions dominate. When developers have substantially superior informationand disagreement with regulators is limited, enforcement of ex-post damage liability is effective, price-based instruments such as ex-post Pigouvian taxes perform well. By contrast, when belief disagreement is large, beliefs are private information, and liability or verifiability constraints are severe, beta testing and regulatory approval become more attractive.

This distinction helps interpret the divergent regulatory paths currently being pursued in practice. Europe has moved toward quantity regulation, adopting a risk-based framework that includes outright bans on certain high-risk applications, testing through regulatory sandboxes prior to full deployment (see articles 57 and 60 of the European Union Artificial Intelligence Act, European Union, 2024).

By contrast, the United States has leaned toward an ex-post, liability-based approach, relying on existing tort law, consumer protection, and sector-specific enforcement after damages materialize. This regulatory stance is consistent with an environment in which developers are presumed to have superior information about risks, making ex-post Pigouvian taxes a relatively effective tool.

Figure 2: Release decisions in the first and second periods

Panel A: Time 1



7 Conclusion

In this paper, we study optimal regulation in environments characterized by two salient features of artificial intelligence: substantial uncertainty about potential social harms and persistent disagreement between developers and regulators about their likelihood and magnitude. We show that these features fundamentally shape the relative performance of regulatory instruments.

Our main result is that optimal regulation in this environment takes the form of a two-stage process. In the first stage, regulators determine whether an algorithm should undergo limited beta testing or receive broader deployment approval. Information generated during this experimental phase then informs the second-stage decision on whether deployment should expand or be withdrawn. When belief disagreement is significant and beliefs are private information, this combination of mandatory experimentation and regulatory approval can strictly dominate both ex-ante and ex-post Pigouvian taxation.

While our analysis emphasizes belief heterogeneity, real-world regulation is shaped by multiple frictions, including information asymmetries, limited liability, and constraints on verifiability. These frictions help explain cross-jurisdictional differences in regulatory approaches.

Europe's stronger emphasis on banning high-risk applications and regulatory sandboxes closely mirrors the quantity-based regulation that is optimal in our model when belief disagreement is central. By contrast, the greater reliance in the United States on ex-post liability enforced through existing legal frameworks is more consistent with environments in which developers are presumed to have superior information about risks.

AI regulation faces broader challenges that are common to other industries. Regulatory capture, where dominant firms shape policies to serve their interests (Stigler, 1971 and Peltzman, 1976), is a significant risk. Additionally, measuring externalities and internalities is complex, and high compliance costs could stifle innovation.

While our analysis focuses on the regulatory framework in a single country, international cooperation is essential when AI systems generate cross-border externalities (see Choi, Jeon, and Menicucci, 2026 and Lian and Schaab, 2026 for recent analyses).

Isaac Asimov wrote, "The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom." Continued work on AI regulation is essential to build the wisdom required to harness the benefits of this new technology while managing its risks.

References

ACEMOGLU, D. (2021): "Harms of AI," in *The Oxford Handbook of AI Governance*, Oxford University Press.

ACEMOGLU, D., D. KONG, AND A. OZDAGLAR (2026): "AI, Human Cognition and Knowledge Collapse," Working Paper 34910, National Bureau of Economic Research.

ACEMOGLU, D. AND T. LENSMAN (2024): "Regulating Transformative Technologies," *American Economic Review: Insights*, 6, 359–376.

ACEMOGLU, D., A. OZDAGLAR, AND J. SIDERIUS (2025): "AI and Social Media: A Political Economy Perspective," Working Paper 33892, National Bureau of Economic Research.

ACEMOGLU, D. AND P. RESTREPO (2022): "Tasks, Automation, and the Rise in U.S. Wage Inequality," *Econometrica*, 90, 1973–2016.

AMODEI, D., C. OLAH, J. STEINHARDT, P. CHRISTIANO, J. SCHULMAN, AND D. MANÉ (2016): "Concrete Problems in AI Safety," *arXiv preprint arXiv:1606.06565*.

ARGYLE, L. P. (2025): "Political persuasion by artificial intelligence," *Science*, 390, 983–984.

BARON, D. P. AND R. B. MYERSON (1982): "Regulating a Monopolist with Unknown Costs," *Econometrica*, 50, 911–930.

BERAJA, M., A. KAO, D. Y. YANG, AND N. YUCHTMAN (2023): "Exporting the Surveillance State via Trade in AI," Working paper, Brookings Center on Regulation and Markets.

BLATTNER, L., S. NELSON, AND J. SPIESS (2021): "Unpacking the Black Box: Regulating Algorithmic Decisions," *arXiv preprint arXiv:2110.03443*.

BURD, J. T. (2021): "Regulatory Sandboxes for Safety Assurance of Autonomous Vehicles," *U. Pa. JL & Pub. Aff.*, 7, 194.

BURSTEIN, A., E. MORALES, AND J. VOGEL (2019): "Changes in Between-Group Inequality: Computers, Occupations, and International Trade," *American Economic Journal: Macroeconomics*, 11, 348–400.

BURSZTYN, L., A. IMAS, R. JIMÉNEZ-DURÁN, A. LEONARD, AND C. ROTH (2025): "Social Dynamics of AI Adoption," Working Paper 34488, National Bureau of Economic Research, also circulated as University of Chicago Becker Friedman Institute Working Paper No. 2025-144.

CALLANDER, S. (2011): "Searching and Learning by Trial and Error," *American Economic Review*, 101, 2277–2308.

CALLANDER, S. AND H. LI (2024): "Regulating an Innovative Industry," Working paper.

CHEN, Y. AND X. HUA (2026): "Product Safety in the Age of AI: Autonomy, R&D, and Liability," *The Economic Journal*.

CHOI, J. P., D.-S. JEON, AND D. MENICUCCI (2026): "AI Safety and Competition," Manuscript.

CORNELLI, G., S. DOERR, L. GAMBACORTA, AND O. MERROUCHE (2024): "Regulatory Sandboxes and Fintech Funding: Evidence from the UK," *Review of Finance*, 28, 203–233.

COSTINOT, A. AND I. WERNING (2023): "Robots, Trade, and Luddism: A Sufficient Statistic Approach to Optimal Technology Regulation," *The Review of Economic Studies*, 90, 2261–2291.

EUROPEAN UNION (2024): "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, oJ L, 2024/1689, 12 July 2024.

FARBOODI, M. AND L. VELDKAMP (2021): "A Model of the Data Economy," Tech. rep., National Bureau of Economic Research.

FARHI, E. AND X. GABAIX (2020): "Optimal Taxation with Behavioral Agents," *American Economic Review*, 110, 298–336.

FIROOZ, H., Z. LIU, AND Y. WANG (2025): "Automation and the Rise of Superstar Firms," *Journal of Monetary Economics*, 153, 103048.

Freund, L. B. and L. F. Mann (2026): "Job Transformation, Specialization, and the Labor Market Effects of AI," Tech. rep., Working Paper, march, 2026.

Gans, J. S. (2024): "How Learning About Harms Impacts the Optimal Rate of Artificial Intelligence Adoption," *Economic Policy*, 40, 199–219.

——— (2025): "Regulating the Direction of Innovation," *Journal of Public Economics*, 246, 105375.

Gittins, J. (1974): "A Dynamic Allocation Index for the Sequential Design of Experiments," *Progress in statistics*, 241–266.

Guerreiro, J., S. Rebelo, and P. Teles (2022): "Should Robots Be Taxed?" *The Review of Economic Studies*, 89, 279–311.

Hampole, M., D. Papanikolaou, L. D. W. Schmidt, and B. Seegmiller (2025): "Artificial Intelligence and the Labor Market," *Working Paper*, working paper.

Harris, L. and C. Criddle (2025): "Insurers Retreat from AI Cover as Risk of Multibillion-Dollar Claims Mounts," *Financial Times*, accessed: 2025-11-29.

Hart, R. (2024): "AI Models Like ChatGPT Won't Reach Human Intelligence, Meta's AI Chief Says," *Forbes*.

Heaven, W. D. (2023): "Geoffrey Hinton Tells Us Why He's Now Scared of the Tech He Helped Build," *MIT Technology Review*.

Ide, E. and E. Talamàs (2025): "Artificial Intelligence in the Knowledge Economy," *Journal of Political Economy*, 133, 3762–3800.

Jagerman, R., D. M. Roijers, D. Hennes, and M. de Rijke (2020): "Safe Exploration for Optimizing Contextual Bandits," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1613–1616.

Jones, C. I. (2024): "The AI Dilemma: Growth Versus Existential Risk," *American Economic Review: Insights*, 6, 575–590.

Jones, C. I. and C. Tonetti (2020): "Nonrivalry and the Economics of Data," *American Economic Review*, 110, 2819–2858.

Karger, E., O. Kuusela, J. Abaluck, K. A. Bryan, B. Halperin, T. R. Jones, C. Murphy, P. Trammell, M. Reynolds, D. Mayland, R. Viswanathan, A. Mittal, R. Ceppas de Castro, J. Rosenberg, and P. E. Tetlock (2026): "Forecasting the Economic Effects of AI," Working Paper 35046, National Bureau of Economic Research.

KAZEROUNI, A., M. GHAVAMZADEH, Y. ABBASI-YADKORI, AND B. VAN ROY (2017): "Conservative Contextual Linear Bandits," in *Advances in Neural Information Processing Systems*, vol. 30.

KLEINBERG, J., J. LUDWIG, S. MULLAINATHAN, AND C. R. SUNSTEIN (2018): "Discrimination in the Age of Algorithms," *Journal of Legal Analysis*, 10, 113–174.

KOH, A. AND S. SANGUANMOO (2025): "Robust Technology Regulation," Revised November 2025.

KOLSTAD, C. D. AND T. S. ULEN (1983): "Ex Post Liability for Harm vs. Ex Ante Safety Regulation: Substitutes or Complements?" *American Economic Review Papers and Proceedings*, 73, 277–280.

KOSMYNA, N., E. HAUPTMANN, Y. T. YUAN, J. SITU, X.-H. LIAO, A. V. BERESNITZKY, I. BRAUNSTEIN, AND P. MAES (2025): "Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Tasks," *arXiv preprint*, preprint, under review.

LAFFONT, J. J. (1977): "More on Prices vs. Quantities," *The Review of Economic Studies*, 44, 177–182.

LAFFONT, J.-J. AND J. TIROLE (1986): "Using Cost Observation to Regulate Firms," *Journal of Political Economy*, 94, 614–641.

——— (1993): *A Theory of Incentives in Procurement and Regulation*, MIT Press.

LEHR, N. H. AND P. RESTREPO (2025): "The Price of Intelligence: How Should Socially-minded Firms Price and Deploy AI?" Working Paper 34424, National Bureau of Economic Research.

LIAN, C. AND A. SCHAAB (2026): "Regulating AI in the Grip of International Competition," Unpublished manuscript.

MARTINEZ, J. (2021): "Putty-Clay Automation," *DP16022*.

PELTZMAN, S. (1976): "Toward a More General Theory of Regulation," *Journal of Law and Economics*, 19, 211–240.

PURI, I. AND L. VELDKAMP (2026): "Artificial intelligence and cognitive inequality," *Journal of Monetary Economics*, 157, 103884.

RAMBACHAN, A., J. KLEINBERG, S. MULLAINATHAN, AND J. LUDWIG (2020): "An Economic Approach to Regulating Algorithms," NBER Working Paper 27111, National Bureau of Economic Research, revised January 2021.

RUSSELL, S., D. DEWEY, AND M. TEGMARK (2015): "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine*, 36, 105–114.

SANDRONI, A. AND F. SQUINTANI (2007): "Overconfidence, Insurance, and Paternalism," *American Economic Review*, 97, 1994–2004.

SCOTT, J. (2025): "New Turing Award Winner Richard Sutton Calls Doomers "Out of Line," Talks Path to Human-Like AI," *BetaKit*, accessed: 2025-03-09.

STIGLER, G. J. (1971): "The Theory of Economic Regulation," *The Bell Journal of Economics and Management Science*, 2, 3–21.

THOMPSON, W. R. (1933): "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples," *Biometrika*, 25, 285–294.

THUEMMEL, U. (2023): "Optimal Taxation of Robots," *Journal of the European Economic Association*, 21, 1154–1190.

TIROLE, J. (2010): "From Pigou to Extended Liability: On the Optimal Taxation of Externalities Under Imperfect Financial Markets," *The Review of Economic Studies*, 77, 697–729.

TYRANGIEL, J. (2025): "Sam Altman on ChatGPT's First Two Years, Elon Musk and AI Under Trump," *Bloomberg Businessweek*, accessed: 2025-02-09.

WEITZMAN, M. L. (1974): "Prices vs. Quantities," *The Review of Economic Studies*, 41, 477–491.

——— (1978): "Optimal Rewards for Economic Regulation," *The American Economic Review*, 68, 683–691.

WEYL, E. G. (2007): "Is Arbitrage Socially Beneficial?" *Working Paper*.

WIENER, N. (1960): "Some Moral and Technical Consequences of Automation," *Science*, 131, 1355–58.

WU, Y., R. SHARIFF, T. LATTIMORE, AND C. SZEPESVÁRI (2016): "Conservative Bandits," in *Advances in Neural Information Processing Systems*, vol. 29.

ZARRA, A. (2026): "Experimentalism beyond ex ante regulation: A law and economics perspective on AI regulatory sandboxes," *Cambridge Forum on AI: Law and Governance*, 2, 1–23.

Online Appendix
Regulating Artificial Intelligence

A Proofs

A.1 Proof of Lemma 1

First, note that

$$\begin{aligned}\mathbb{E}_1^d[\mathcal{V}_2^*] =& \pi(\mu_1)\mathbb{E}_1^d\left[\max\{u-(\hat{\phi}_{s,i}^2+\hat{\sigma}_{s,i}^2)-\chi(\hat{\phi}_{d,e}^2+\hat{\sigma}_{d,e}^2),0\}N\right] \\ &+(1-\pi(\mu_1))\max\{u-\sigma_{s,i}^2-\chi\sigma_{d,e}^2,0\}N\end{aligned}$$

So, assuming $\mu_1<\kappa$,

$$\begin{aligned}\frac{d\mathbb{E}_1^d[\mathcal{V}_2^*]}{d\mu_1} &= \pi'(\mu_1)\left\{\mathbb{E}_1^d\left[\max\{u-(\hat{\phi}_{s,i}^2+\hat{\sigma}_{s,i}^2)-\chi(\hat{\phi}_{d,e}^2+\hat{\sigma}_{d,e}^2),0\}N\right]-\max\{u-\sigma_{s,i}^2-\chi\sigma_{d,e}^2,0\}N\right\} \\ &> \pi'(\mu_1)\left\{\max\{u-\mathbb{E}_1^d\left[(\hat{\phi}_{s,i}^2+\hat{\sigma}_{s,i}^2)\right]-\chi\mathbb{E}_1^d\left[\hat{\phi}_{d,e}^2+\hat{\sigma}_{d,e}^2\right],0\}N-\max\{u-\sigma_{s,i}^2-\chi\sigma_{d,e}^2,0\}N\right\} \\ &\geq \pi'(\mu_1)\left\{\max\{u-\sigma_{s,i}^2-\chi\sigma_{d,e}^2,0\}N-\max\{u-\sigma_{s,i}^2-\chi\sigma_{d,e}^2,0\}N\right\}=0.\end{aligned}$$

The inequality holds because the expected value of the maxima is higher than the maximum of the expected value. The inequality is strict because the probability that the algorithm is implemented in period two, given the information obtained in period one, is strictly positive but less than one.

A.2 Proof of Proposition 1

The developer chooses μ_1 to maximize

$$\mathcal{V}\equiv(1-\beta)(u-\sigma_{s,i}^2-\chi\sigma_{d,e}^2)\mu_1+\beta\mathbb{E}_1^d(\mathcal{V}_2^*). \tag{A.1}$$

From Lemma 1, we know that $\mathbb{E}_1^d(\mathcal{V}_2^*)$ is increasing in μ_1. So, if $\chi\sigma_{d,e}^2\leq u-\sigma_{s,i}^2$, then the developer chooses $\mu_1=N$ since $\mathcal{V}$ is always increasing in μ_1.

Suppose instead that $\chi\sigma^2_{d,e} > u - \sigma^2_{s,i}$. In this case, if $\mathcal{B} = 0$, the developer does not release the algorithm at time two. Since $\alpha \in (0,1)$, then $\pi'(\mu_1) \to \infty$ as $\mu_1 \to 0$. It follows that the optimal $\mu_1 > 0$.

$$\frac{d\mathbb{E}^d_1(\mathcal{V}^*_2)}{d\mu_1} = \pi'(\mu_1)\mathbb{E}^d_1\left[\max\{u - (\hat{\phi}^2_{s,i} + \hat{\sigma}^2_{s,i}) - \chi(\hat{\phi}^2_{d,e} + \hat{\sigma}^2_{d,e}), 0\}N\right], \tag{A.2}$$

so, the first order condition is given by

$$(1-\beta)\{u - \sigma^2_{s,i} - \chi\sigma^2_{d,e}\} + \beta\pi'(\mu_1)\mathbb{E}^d_1\left[\max\{u - (\hat{\phi}^2_{s,i} + \hat{\sigma}^2_{s,i}) - \chi(\hat{\phi}^2_{d,e} + \hat{\sigma}^2_{d,e}), 0\}N\right] = 0$$

$$\Leftrightarrow \pi'(\mu_1)\Lambda^d N = 1 \Leftrightarrow \alpha\Lambda^d\frac{N}{\kappa} = \left(\frac{\mu_1}{\kappa}\right)^{1-\alpha} \Leftrightarrow \mu_1 = \left[\alpha\Lambda^d\frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}}\kappa,$$

or $\mu_1 = \kappa$ if $\left[\alpha\Lambda^d\frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}} > 1$.

A.3 Proof of Lemma 2

As before, assuming $\mu_1 < \kappa$,

$$\begin{aligned}
\frac{d\mathbb{E}^s_1(d\mathcal{W}^*_2)}{d\mu_1} &= \pi'(\mu_1)\left\{\mathbb{E}^s_1\left(\max\left\{u - \mathbb{E}^s_2(\phi^2_i) - (N+\chi)\mathbb{E}^s_2(\phi^2_e), 0\right\}N\right)\right. \\
&\quad \left. - \max\left\{u - \sigma^2_{s,i} - (N+\chi)\sigma^2_{s,e}, 0\right\}N\right\} \\
&> \pi'(\mu_1)\left\{\max\left\{u - \mathbb{E}^s_1(\phi^2_i) - (N+\chi)\mathbb{E}^s_1(\phi^2_e), 0\right\}N\right. \\
&\quad \left. - \max\left\{u - \sigma^2_{s,i} - (N+\chi)\sigma^2_{s,e}, 0\right\}N\right\} = 0.
\end{aligned}$$

A.4 Proof of Proposition 2

The planner chooses μ_1 to maximize

$$\mathcal{W} \equiv (1-\beta)[u - \sigma^2_{s,i} - (N+\chi)\sigma^2_{s,e}]\mu_1 + \beta\mathbb{E}^s_1(\mathcal{W}^*_2). \tag{A.3}$$

From Lemma 2, we know that $\mathbb{E}_1^s(\mathcal{W}_2^*)$ is increasing in μ_1. So, if $\sigma_{s,e}^2 \leq \frac{u-\sigma_{s,i}^2}{N+\chi}$, then the planner chooses $\mu_1 = N$ since $\mathcal{W}$ is always increasing in μ_1.

Suppose instead that $\sigma_{s,e}^2 > \frac{u-\sigma_{s,i}^2}{N+\chi}$. In this case, if $\mathcal{B} = 0$, then the planner does not release the algorithm at time two. Since $\pi'(\mu_1) \to \infty$ as $\mu_1 \to 0$, then $\mu_1 > 0$.

$$\frac{d\mathbb{E}_1^s[\mathcal{W}_2^*]}{d\mu_1} = \pi'(\mu_1)\mathbb{E}_1^s\left[\max\{u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2) - (N+\chi)(\hat{\phi}_{s,e}^2 + \hat{\sigma}_{s,e}^2), 0\}N\right]. \quad \text{(A.4)}$$

So, the first order condition is given by

$$(1-\beta)\{u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2\} + \beta\pi'(\mu_1)\mathbb{E}_1^s\left[\max\{u - (\hat{\phi}_{s,i}^2 + \hat{\sigma}_{s,i}^2) - (N+\chi)(\hat{\phi}_{s,e}^2 + \hat{\sigma}_{s,e}^2), 0\}N\right] = 0$$

$$\Leftrightarrow \pi'(\mu_1)\Lambda^s N = 1 \Leftrightarrow \alpha\Lambda^s\frac{N}{\kappa} = \left(\frac{\mu_1}{\kappa}\right)^{1-\alpha} \Leftrightarrow \mu_1 = \left[\alpha\Lambda^s\frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}}\kappa,$$

or $\mu_1 = \kappa$ if $\left[\alpha\Lambda^s\frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}} > 1$.

A.5 Ex-ante Pigouvian Taxes with Heterogeneous Beliefs and Proof of Proposition 4

Assume that the regulator enforces the ex-ante taxes specified in equation (11). Consider the problem of period two. Substituting the optimal license price, the developer's utility is given by

$$\begin{cases} [u - \mathbb{E}_2^s(\phi_i^2) - \chi\mathbb{E}_2^d(\phi_e^2)]\mu_2 - T_2^{ex-ante} = [u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2)]\mu_2, & \text{if } \mathcal{B} = 1 \\ [u - \sigma_{s,i}^2 - \chi\sigma_{d,e}^2]\mu_2 - T_2^{ex-ante} = [u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2]\mu_2, & \text{if } \mathcal{B} = 0. \end{cases}$$

It follows that private and social incentives are always aligned in period two.

Turning to the problem of the first period, with the taxes given by equation (11), the problem becomes:

$$\max_{\mu_1}\left\{(1-\beta)\left\{u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2\right\}\mu_1 + \beta\mathbb{E}_1^d[\mathcal{W}_2^*]\right\}. \quad \text{(A.5)}$$

We immediately see that if the risk is not very large, $\sigma^2_{s,e} \leq \frac{u-\sigma^2_{s,i}}{N+\chi}$, then the developer chooses $\mu_1 = N$. If the risk is large, $\sigma^2_{s,e} > \frac{u-\sigma^2_{s,i}}{N+\chi}$, then the developer's choice satisfies

$$\pi'(\mu_1) N \Lambda^s \frac{\mathbb{E}^d_1[\max\{u - \mathbb{E}^s_2(\phi^2_i) - (N+\chi)\mathbb{E}^s_2(\phi^2_e), 0\}]}{\mathbb{E}^s_1[\max\{u - \mathbb{E}^s_2(\phi^2_i) - (N+\chi)\mathbb{E}^s_2(\phi^2_e), 0\}]} = 1$$

$$\Leftrightarrow \mu_1 = \left[\alpha \Lambda^s \frac{\mathbb{E}^d_1[\max\{u - \mathbb{E}^s_2(\phi^2_i) - (N+\chi)\mathbb{E}^s_2(\phi^2_e), 0\}]}{\mathbb{E}^s_1[\max\{u - \mathbb{E}^s_2(\phi^2_i) - (N+\chi)\mathbb{E}^s_2(\phi^2_e), 0\}]} \frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}} \kappa.$$

So, in general, the developer prefers to beta test on a larger population sample than what would be socially optimal.

Suppose that the tax in period one is given by equation (13). Then, again, if the uncertainty is small enough, $\sigma^2_{s,e} \leq \frac{u-\sigma^2_{s,i}}{N+\chi}$, then the developer chooses $\mu_1 = N$. Instead, if the risk is large, $\sigma^2_{s,e} > \frac{u-\sigma^2_{s,i}}{N+\chi}$, the developer solves the same problem as the planner:

$$\max_{\mu_1} \left\{ (1-\beta) \left\{ u - \sigma^2_{s,i} - (N+\chi)\sigma^2_{s,e} \right\} \mu_1 + \beta \mathbb{E}^s_1(\mathcal{W}^*_2) \right\}. \tag{A.6}$$

So private and social incentives are aligned.

A.6 Proof of Lemma 3

Given the tax policy, the developer's problem at time two is:

$$\mathcal{V}_2 = \begin{cases} (p_2 - \tau_2)\mu_2 - \chi \mathbb{E}^\theta_2(\phi^2_e)\mu_2, & \text{if } p_2 \leq u - \mathbb{E}^s_2(\phi^2_i) \text{ and } \mathcal{B} = 1, \\ (p_2 - \tau_2)\mu_2 - \chi \sigma^2_{\theta,e}\mu_2, & \text{if } p_2 \leq u - \sigma^2_{s,i} \text{ and } \mathcal{B} = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the optimal price is

$$p_2 = \begin{cases} u - \mathbb{E}^s_2(\phi^2_i), & \text{if } \mathcal{B} = 1, \\ u - \sigma^2_{s,i}, & \text{if } \mathcal{B} = 0. \end{cases}$$

If $\mathcal{B} = 1$, the developer releases the algorithm to the entire population if $u - \mathbb{E}^s_2(\phi^2_i) - \tau_2 > \chi \mathbb{E}^\theta_2(\phi^2_e)$ and sets $\mu_2 = 0$ otherwise. If $\mathcal{B} = 0$, the developer releases the algorithm to the entire population if $u - \sigma^2_{s,i} - \tau_2 > \chi \sigma^2_{\theta,e}$ and sets $\mu_2 = 0$ otherwise.

At time one, it is optimal to set $p_1 = u - \sigma_{s,i}^2$. The developer's problem at time one (replacing p_1) is given by

$$\max_{\mu_1}(1-\beta)\{u - \sigma_{s,i}^2 - \chi\sigma_{\theta,e}^2 - \tau_1\}\mu_1 + \beta\mathbb{E}_1^{\theta}[\mathcal{V}_{2,\theta}^*], \tag{A.7}$$

where $\mathbb{E}_1^{\theta}(\mathcal{V}_{2,\theta}^*)$ is increasing in μ_1 as in Lemma 1.

So, if $\chi\sigma_{\theta,e}^2 \le u - \sigma_{s,i}^2 - \tau_1$, then the developer relases the algorithm to the entire population $\mu_1 = N$.

If $\chi\sigma_{\theta,e}^2 > u - \sigma_{s,i}^2 - \tau_1$, it is optimal to beta test the algorithm. In this case, the optimal μ_1 solves

$$\max_{\mu_1} -\mu_1 + \pi(\mu_1)\frac{\beta}{1-\beta}\frac{\mathbb{E}_1^{\theta}[\max\{u - \mathbb{E}_2^s(\phi_i^2) - \chi\mathbb{E}_2^{\theta}(\phi_e^2) - \tau_2, 0\}] - \max\{u - \sigma_{s,i}^2 - \chi\sigma_{\theta,e}^2 - \tau_2, 0\}}{\sigma_{s,i}^2 + \chi\sigma_{\theta,e}^2 + \tau_1 - u}N. \tag{A.8}$$

Let

$$\Lambda^{\theta}(\tau_1,\tau_2) \equiv \frac{\beta}{1-\beta}\frac{\mathbb{E}_1^{\theta}[\max\{u - \mathbb{E}_2^s(\phi_i^2) - \chi\mathbb{E}_2^{\theta}(\phi_e^2) - \tau_2, 0\}] - \max\{u - \sigma_{s,i}^2 - \chi\sigma_{\theta,e}^2 - \tau_2, 0\}}{\sigma_{s,i}^2 + \chi\sigma_{\theta,e}^2 + \tau_1 - u}. \tag{A.9}$$

The first order condition is,

$$\pi'(\mu_1)\Lambda^{\theta}(\tau_1,\tau_2)N = 1 \Leftrightarrow \mu_1 = \left[\alpha\Lambda^{\theta}(\tau_1,\tau_2)\frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}}\kappa.$$

A.7 Proof of Proposition 5

We solve for the optimal tax policy without commitment. At time two, the regulator chooses τ_2 to maximize

$$\begin{cases} \{u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2)\}\mathbb{E}_2^s(\mu_2^{\theta}), & \text{if } \mathcal{B} = 1, \\ \{u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2\}\mathbb{E}_2^s(\mu_2^{\theta}), & \text{if } \mathcal{B} = 0, \end{cases}$$

where $\mathbb{E}_2^s(\mu_2^{\theta})$ denotes the regulator's expectations of the developer's choice of μ_2 given the regulator's beliefs over the developer's type at time two. These beliefs are influenced by the developer's decisions observed at time one.

If given the regulator's beliefs the algorithm should be released, then it is optimal to set $\tau_2 = 0$, since under this tax $\mu_2^\theta = N$ for all θ:

$$\begin{cases} 0 \le u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2) < u - \mathbb{E}_2^s(\phi_i^2) - \chi\mathbb{E}_2^s(\phi_e^2) \le u - \mathbb{E}_2^s(\phi_i^2) - \chi\mathbb{E}_2^\theta(\phi_e^2), & \forall\theta \in \Theta \\ 0 \le u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2 < u - \sigma_{s,i}^2 - \chi\sigma_{s,e}^2 \le u - \sigma_{s,i}^2 - \chi\sigma_{\theta,e}^2, & \forall\theta \in \Theta. \end{cases}$$

If the regulator's beliefs are such that the algorithm should be withdrawn, then by setting $\tau_2 = p_2$ the regulator ensures that all developer types withdraw the algorithm.[10] Welfare at time two coincides with the social optimum.

At time one, the regulator chooses τ_1 to maximize

$$(1-\beta)\{u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2\} \int_{\underline{\theta}}^{\overline{\theta}} \mu_1^\theta dF(\theta) + \beta\mathbb{E}_1^s(\mathcal{W}_2^*). \tag{A.10}$$

First, suppose that uncertainty about the externality is small $\sigma_{s,e}^2 \le \frac{u-\sigma_{s,i}^2}{N+\chi}$. In this case, it is efficient to release the algorithm to the entire population. So, the regulator sets $\tau_1 = 0$ and $\mu_1^\theta = N$ for all θ, since

$$0 \le u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2 < u - \sigma_{s,i}^2 - \chi\sigma_{s,e}^2 \le u - \sigma_{s,i}^2 - \chi\sigma_{\theta,e}^2, \quad \forall\theta \in \Theta.$$

Suppose that uncertainty is large $\sigma_{s,e}^2 > \frac{u-\sigma_{s,i}^2}{N+\chi}$. Then, the regulator chooses τ_1 to maximize

$$\begin{aligned}(1-\beta)\{u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2\} \int_{\underline{\theta}}^{\overline{\theta}} \mu_1^\theta dF(\theta) + \beta \int_{\underline{\theta}}^{\overline{\theta}} \pi(\mu_1^\theta) dF(\theta) \\ \times \mathbb{E}_1^s[\max\{u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2), 0\}]N.\end{aligned} \tag{A.11}$$

Equivalently, τ_1 solves

$$\max_{\tau_1} \left(- \int_{\underline{\theta}}^{\overline{\theta}} \mu_1^\theta dF(\theta) + \int_{\underline{\theta}}^{\overline{\theta}} \pi(\mu_1^\theta) dF(\theta) \Lambda^s N \right). \tag{A.12}$$

[10] We assume that, in indifference, developers decide in favor of the planner.

The first order condition with respect to τ_1 is given by

$$\int_{\underline{\theta}}^{\overline{\theta}} \frac{d\mu_1^\theta/d\tau_1}{\int_{\underline{\theta}}^{\overline{\theta}} d\mu_1^\theta/d\tau_1 dF(\theta)} \pi'(\mu_1^\theta)\Lambda^s N dF(\theta) = 1, \tag{A.13}$$

where μ_1^θ is given by

$$\mu_1^\theta = \min\left\{ \left[\alpha\Lambda^\theta \frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}}, 1 \right\}\kappa \tag{A.14}$$

and

$$\Lambda^\theta \equiv \frac{\beta}{1-\beta} \frac{\mathbb{E}_1^\theta\left[\{u - \mathbb{E}_2^s(\phi_i^2) - \chi\mathbb{E}_2^\theta(\phi_e^2)\}\mathbb{1}_{\{u-\mathbb{E}_2^s(\phi_i^2)-(N+\chi)\mathbb{E}_2^s(\phi_e^2)\geq 0\}}\right]}{\sigma_{s,i}^2 + \chi\sigma_{\theta,e}^2 + \tau_1 - u}.$$

Since Λ^θ is increasing in θ, the developer's optimal beta test sample size is also increasing in θ.

A.8 Proof of Lemma 4

For any regulation policy, the problem of the developer at time two is given by:

$$\mathcal{V}_2 = \begin{cases} p_2\mu_2 - \chi\mathbb{E}_2^\theta(\phi_e^2)\mu_2, & \text{if } p_2 \leq u - \mathbb{E}_2^s(\phi_i^2) \text{ and } \mathcal{B} = 1, \\ p_2\mu_2 - \chi\sigma_{\theta,e}^2\mu_2, & \text{if } p_2 \leq u - \sigma_{s,i}^2 \text{ and } \mathcal{B} = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the optimal price is

$$p_2 = \begin{cases} u - \mathbb{E}_2^s(\phi_i^2), & \text{if } \mathcal{B} = 1, \\ u - \sigma_{s,i}^2, & \text{if } \mathcal{B} = 0. \end{cases}$$

So, if $\mathcal{B} = 1$, the developer releases the algorithm to the maximum $\mu_2 = \overline{\mu}_2$ if $u - \mathbb{E}_2^s(\phi_i^2) > \chi\mathbb{E}_2^\theta(\phi_e^2)$ and sets $\mu_2 = 0$ otherwise. Analogously, if $\mathcal{B} = 0$, the developer releases the algorithm to to the maximum $\mu_2 = \overline{\mu}_2$ if $u - \sigma_{s,i}^2 > \chi\sigma_{\theta,e}^2$ and sets $\mu_2 = 0$ otherwise.

At time one, it is optimal to set $p_1 = u - \sigma^2_{s,i}$. The developer's problem at time one (replacing p_1) is given by

$$\max_{\mu_1} (1-\beta)\{u - \sigma^2_{s,i} - \chi\sigma^2_{\theta,e}\}\mu_1 + \beta\mathbb{E}^{\theta}_1[\mathcal{V}^*_{2,\theta}]. \tag{A.15}$$

Where $\mathbb{E}^{\theta}_1(\mathcal{V}^*_{2,\theta})$ is increasing in μ_1 as in Lemma 1.

If $\chi\sigma^2_{\theta,e} \leq u - \sigma^2_{s,i}$, then the developer releases the algorithm to the maximum number of people $\mu_1 = \overline{\mu}_1$.

If $\chi\sigma^2_{\theta,e} > u - \sigma^2_{s,i}$, it is optimal to beta test the algorithm. In this case, the optimal μ_1 solves

$$\max_{\mu_1} -\mu_1 + \pi(\mu_1)\frac{\beta}{1-\beta}\frac{\mathbb{E}^{\theta}_1[\max\{u - \mathbb{E}^s_2(\phi^2_i) - \chi\mathbb{E}^{\theta}_2(\phi^2_e), 0\}\overline{\mu}_2] - \mathbb{E}^{\theta}_1[\max\{u - \sigma^2_{s,i} - \chi\sigma^2_{\theta,e}, 0\}\overline{\mu}_2]}{\sigma^2_{s,i} + \chi\sigma^2_{\theta,e} - u}. \tag{A.16}$$

Let

$$\Lambda^{\theta} \equiv \frac{\beta}{1-\beta}\frac{\mathbb{E}^{\theta}_1[\max\{u - \mathbb{E}^s_2(\phi^2_i) - \chi\mathbb{E}^{\theta}_2(\phi^2_e), 0\}\overline{\mu}_2] - \mathbb{E}^{\theta}_1[\max\{u - \sigma^2_{s,i} - \chi\sigma^2_{\theta,e}, 0\}\overline{\mu}_2]}{\sigma^2_{s,i} + \chi\sigma^2_{\theta,e} - u}. \tag{A.17}$$

Then, the first order condition is given by

$$\pi'(\mu_1)\Lambda^{\theta} = 1 \Leftrightarrow \mu_1 = \left[\alpha\Lambda^{\theta}\frac{1}{\kappa}\right]^{\frac{1}{1-\alpha}}\kappa.$$

In this case, the developer sells to this number of people if $\left[\alpha\Lambda^{\theta}\frac{1}{\kappa}\right]^{\frac{1}{1-\alpha}}\kappa \leq \overline{\mu}_1$ and $\mu_1 = \overline{\mu}_1$ otherwise.

A.9 Proof of Proposition 6

We solve for the optimal regulation policy without commitment. At time two, the regulator chooses $\overline{\mu}_2$ to maximize

$$\begin{cases} \{u - \mathbb{E}^s_2(\phi^2_i) - (N+\chi)\mathbb{E}^s_2(\phi^2_e)\}\mathbb{E}^s_2[\mu^{\theta}_2], & \text{if } \mathcal{B} = 1, \\ \{u - \sigma^2_{s,i} - (N+\chi)\sigma^2_{s,e}\}\mathbb{E}^s_2[\mu^{\theta}_2], & \text{if } \mathcal{B} = 0, \end{cases}$$

where $\mathbb{E}_2^s[\mu_2^\theta]$ denotes the regulator's expectations over the developer's behavior given the regulator's beliefs over the developer's type at time two (which are influenced by the observed decisions at time-one).

If given the regulator's beliefs the algorithm should be released, it is optimal to set $\overline{\mu}_2 = N$, since under this cap $\mu_2^\theta = N$ for all θ:

$$\begin{cases} 0 \leq u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2) < u - \mathbb{E}_2^s(\phi_i^2) - \chi\mathbb{E}_2^s(\phi_e^2) \leq u - \mathbb{E}_2^s(\phi_i^2) - \chi\mathbb{E}_2^\theta(\phi_e^2), \quad \forall\theta \in \Theta \\ 0 \leq u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2 < u - \sigma_{s,i}^2 - \chi\sigma_{s,e}^2 \leq u - \sigma_{s,i}^2 - \chi\sigma_{\theta,e}^2, \quad \forall\theta \in \Theta. \end{cases}$$

Instead, if the regulator's beliefs are such that the algorithm should be withdrawn, then by setting $\overline{\mu}_2 = 0$ the regulator ensures that no developer type releases the algorithm. Note that welfare at time two coincides with the efficient level.

Turning to the problem at time one, the regulator chooses $\overline{\mu}_1$ to maximize

$$(1-\beta)\{u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2\} \int_{\underline{\theta}}^{\overline{\theta}} \mu_1^\theta dF(\theta) + \beta\mathbb{E}_1^s[\mathcal{W}_2^*]. \tag{A.18}$$

First, suppose that uncertainty about the externality is small $\sigma_{s,e}^2 \leq \frac{u-\sigma_{s,i}^2}{N+\chi}$. In this case, it is efficient to release the algorithm to the entire population. So, the regulator sets $\overline{\mu}_1 = N$, which imples that $\mu_1^\theta = N$ for all θ, since

$$0 \leq u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2 < u - \sigma_{s,i}^2 - \chi\sigma_{s,e}^2 \leq u - \sigma_{s,i}^2 - \chi\sigma_{\theta,e}^2, \quad \forall\theta \in \Theta.$$

Suppose that uncertainty is large $\sigma_{s,e}^2 > \frac{u-\sigma_{s,i}^2}{N+\chi}$. Then, the regulator chooses $\overline{\mu}_1$ to maximize

$$\begin{aligned}(1-\beta)\{u - \sigma_{s,i}^2 - (N+\chi)\sigma_{s,e}^2\} \int_{\underline{\theta}}^{\overline{\theta}} \mu_1^\theta dF(\theta) + \beta \int_{\underline{\theta}}^{\overline{\theta}} \pi(\mu_1^\theta) dF(\theta) \\ \times \mathbb{E}_1^s[\max\{u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2), 0\}]N. \end{aligned} \tag{A.19}$$

Let μ_1^* denote the efficient size of the beta-test in this case. We have already established that unconstrained

$$\mu_1^\theta > \mu_1^*.$$

It follows that setting $\overline{\mu}_1 = \mu_1^*$ implies that $\mu_1^\theta = \mu_1^*$ for all θ. This policy implements the efficient outcome and is therefore optimal.

A.10 Optimal Non-Linear Taxes on Developer with Private Information

In the main text, we restrict attention to linear Pigouvian taxes. In this appendix, we generalize the analysis to considering non-linear taxes on license sales. Let the total tax payment as a function of μ_t be $t_t(\mu_t)$. The case of linear taxes obtains when $t_t(\mu_t) = \tau_t \mu_t$.

Definition 1. *A tax policy is a set of (state-contingent) tax functions $\boldsymbol{t} \equiv \{t_t(\mu_t)\}$ for $t = 1, 2$ that determine the taxes imposed on the developer as a function of their release strategy μ_t.*

We allow the tax functions to be arbitrarily non-linear, so solving the developer's general problem becomes more complex. However, it is possible to show that there is a non-linear tax policy that implements efficient allocation.

It is easy to construct tax functions that implement the efficient allocation. For example, consider the following tax policy. At time 2, set the tax function

$$t_2(\mu_2) = \begin{cases} p_2\mu_2, & \text{if } \mu_2 \neq \mu_2^*, \\ 0, & \text{if } \mu_2 = \mu_2^*, \end{cases}$$

where μ_2^* is the efficient level of release. At time 1, if the AI is sufficiently risky so that beta testing is socially optimal, the regulator announces the taxes

$$t_1(\mu_1) = \begin{cases} p_1\mu_1 + \frac{\beta}{1-\beta}\pi(\mu_1)\mathbb{E}_1^{\overline{\theta}}[\max\{u - \mathbb{E}_2^s[\phi_i^2], 0\}N], & \text{if } \mu_1 \neq \mu_1^*, \\ 0, & \text{if } \mu_1 = \mu_1^*, \end{cases}$$

where μ_1^* is the efficient level of beta testing. If the AI is not risky enough to warrant beta testing, the regulator sets $t_1(\mu_1) = 0$.

This tax policy implements the efficient allocation. Effectively, this non-linear tax policy implements the same allocation as the MBR policy.

B Limited Liability

B.1 The Impact of Limited liability on Ex-post Pigouvian Taxes

To study the consequences of limited liability for ex-post Pigouvian taxes, we consider the simple case in which the taxes paid by the developer in each period cannot exceed their revenue ($p_t \mu_t$):

$$T_t = \min\{Ne_t^2, p_t \mu_t\}.$$

This limited liability constraint is a cash-flow constraint that limits the developer's ability to pay taxes.

Under limited liability, the developer's optimal algorithm release policy differs from the social optimum even if beliefs are homogeneous. This divergence arises because the developer's potential losses are capped, encouraging it to release moderately risky algorithms relying on limited liability to protect itself if significant adverse external effects occur.

In this case, we can show that

$$\mathcal{V}_t = \mathcal{W}_t - Ny_t + \mathbb{E}_t[\max\{N(\phi_e^2 + \xi_{e,t}) - p_t, 0\}]\mu_t, \tag{B.20}$$

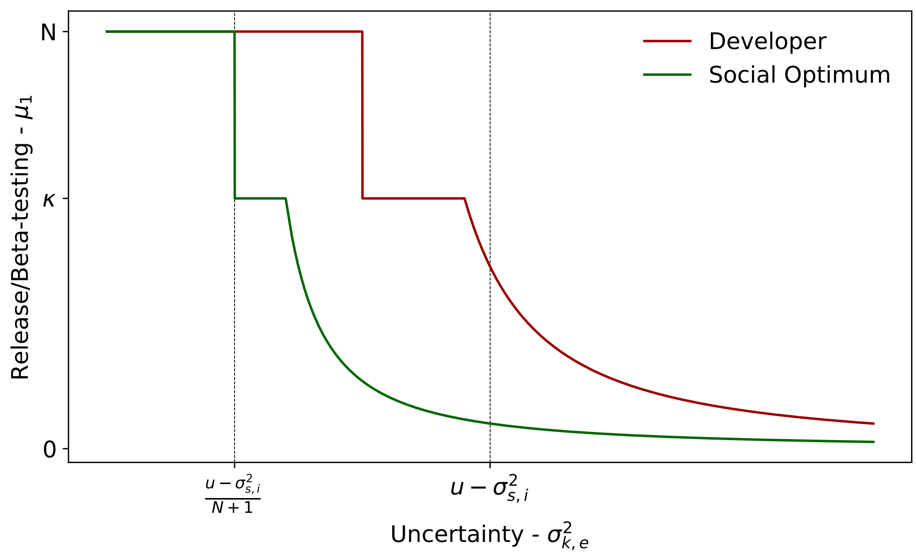
where $\xi_{e,1} = \xi_e$ and $\xi_{e,2} = 0$. To see this, note that if beliefs are homogeneous, then $\mathbb{E}_t^d(\phi_e^2) = \mathbb{E}_t^s(\phi_e^2)$. It is still optimal for the developer to set $p_t = u - \mathbb{E}_t^s(\phi_i^2)$. Replacing this price and the expected taxes into the utility of the developer we find that $\mathcal{V} = (1-\beta)\mathcal{V}_1 + \beta\mathbb{E}_1(\mathcal{V}_2)$, where

$$\begin{aligned}
\mathcal{V}_t &= \mathcal{W}_t - Ny_t - \mathbb{E}_t\left[\min\{Ne_t^2, p_t\mu_t\}\right] + N\mathbb{E}_t\left[e_t^2\right] \\
&= \mathcal{W}_t - Ny_t + \mathbb{E}_t\left[Ne_t^2 + \max\{-Ne_t^2, -p_t\mu_t\}\right] \\
&= \mathcal{W}_t - Ny_t + \mathbb{E}_t\left[\max\{0, Ne_t^2 - p_t\mu_t\}\right] \\
&= \mathcal{W}_t - Ny_t + \mathbb{E}_t\left[\max\{0, N(\phi_e^2 + \xi_{e,t}) - p_t\}\mu_t\right].
\end{aligned}$$

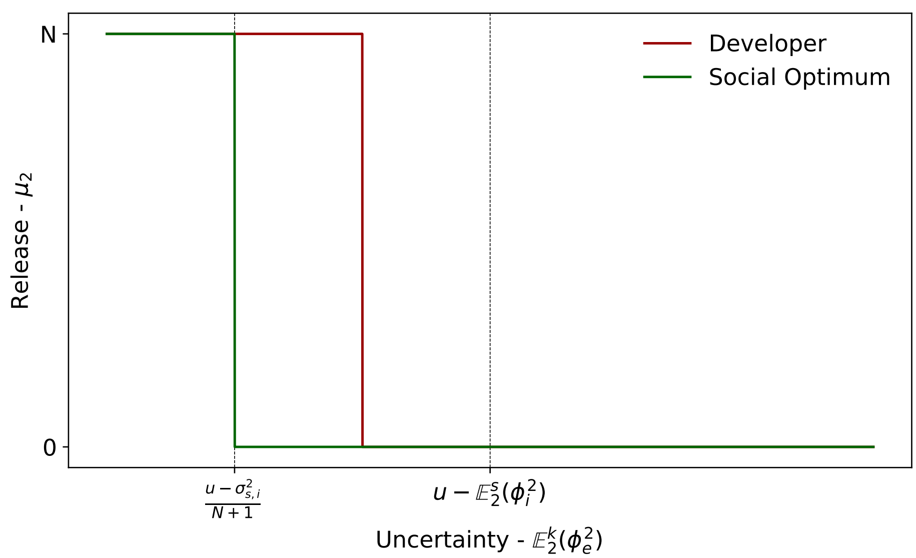
The expected value of the taxes on the developer is lower than the expected social welfare cost of the externality. It follows that the developer is more likely to release the algorithm in period two than the planner, being protected by limited liability should very negative external effects materialize. The same logic implies that the developer may forgo beta testing in period one and release the algorithm immediately, knowing it is protected by limited liability if dire external effects materialize. As a result, the developer may act with less caution than would be socially optimal. Figure 3 illustrates the release decisions under ex-post Pigouvian taxes with limited liability.

Figure 3: Release decisions in the first and second periods under Ex-post Pigouvian Taxes with Limited Liability

Panel A: Time 1



Panel B: Time 2



B.2 The Impact of Limited liability on Ex-ante Pigouvian Taxes

We now show that ex-ante Pigouvian taxes fail to implement the social optimum when limited liability is present, even under homogeneous beliefs. While they ensure that the release policy in period 2 remains optimal, they lead to beta test sample sizes in period one that exceed the optimal level.

Imposing limited liability means that

$$T_t^{ex-ante} = \min\left\{N\mathbb{E}_t(\phi_e^2)\mu_t, p_t\mu_t\right\}. \tag{B.21}$$

Consider the problem at time two, after a successful beta test ($\mathcal{B} = 1$). If limited liability is not binding, then private and social incentives coincide. What happens when limited liability does bind? Given that $p_2 = u - \mathbb{E}_2^s(\phi_i^2)$, limited liability binds whenever $u - \mathbb{E}_2^s(\phi_i^2) < N\mathbb{E}_2^s\left(\phi_e^2\right)$. In this scenario, the developer makes no profit from selling the algorithm and still experiences the consequences of the negative externality when $\chi > 0$; when $\chi = 0$, the developer is indifferent and our tie-breaking selects withdrawal. Limited liability binds only in cases where the regulator would also strictly prefer not to release the algorithm, since $u - \mathbb{E}_2^s(\phi_i^2) < N\mathbb{E}_2^s(\phi_e^2) \leq (N+\chi)\mathbb{E}_2^s(\phi_e^2)$. In other words, the developer and the regulator agree not to release the algorithm whenever limited liability binds.

If the beta test is unsuccessful ($\mathcal{B} = 0$), limited liability binds in period two whenever $u - \sigma_{s,i}^2 < N\sigma_{s,e}^2$. In this case, the developer strictly prefers not to release the algorithm when $\chi > 0$ and is indifferent when $\chi = 0$. The regulator strictly prefers not to release the algorithm because $u - \sigma_{s,i}^2 < N\sigma_{s,e}^2 \leq (N+\chi)\sigma_{s,e}^2$. Therefore, whenever limited liability is binding, both the developer and the regulator agree that the algorithm should not be released.

With ex-ante Pigouvian taxes $\mathbb{E}_1^s(\mathcal{V}_2^*) = \mathbb{E}_1^s(\mathcal{W}_2^*) - Ny_2$, even in the presence of limited liability, so private and social incentives are aligned.

Turning to the problem at time one, private and social incentives coincide if limited liability does not bind. When limited liability binds, both the developer and the planner choose strictly positive values for μ_1, but they select different values. For $\chi > 0$, the developer's information benefit-cost ratio is

$$\Lambda^d = \Lambda^s \times \frac{\sigma_{s,i}^2 + (N+\chi)\sigma_{s,e}^2 - u}{\frac{T_1^{ex-ante}}{\mu_1} + \sigma_{s,i}^2 + \chi\sigma_{s,e}^2 - u} = \Lambda^s \times \frac{\sigma_{s,i}^2 + (N+\chi)\sigma_{s,e}^2 - u}{\chi\sigma_{s,e}^2}, \tag{B.22}$$

where Λ^s is given by (7). Since limited liability binds, we have $\sigma_{s,i}^2 + N\sigma_{s,e}^2 - u > 0$, which implies

$$\frac{\sigma_{s,i}^2 + (N+\chi)\sigma_{s,e}^2 - u}{\chi\sigma_{s,e}^2} > \frac{\chi\sigma_{s,e}^2}{\chi\sigma_{s,e}^2} = 1.$$

When $\chi = 0$, the current private cost of beta testing under a binding limited-liability constraint is zero, so the same force is even stronger. Thus, under limited liability, the developer's information benefit-cost ratio exceeds that of the regulator. Consequently, the developer adopts a more aggressive approach, choosing to beta test the algorithm on a larger sample than the regulator would.

When limited liability binds, the developer's after-tax sales revenue is zero. However, since the tax does not fully internalize the externality imposed on households, the developer still has an incentive to conduct beta testing on a larger sample to increase the likelihood of obtaining valuable information for period two (see Figure 4).

We summarize these results in the following proposition.

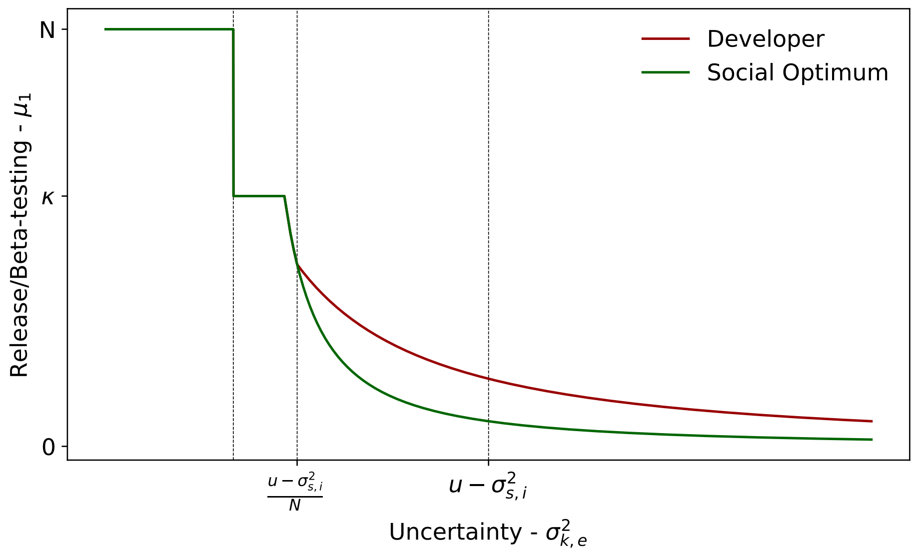
Proposition 7 (Ex-Ante Pigouvian Taxes with Homogeneous Beliefs and Limited Liability)**.** *Suppose developers and society share the same beliefs. If the regulator implements ex-ante Pigouvian taxes subject to the limited liability constraints in equation* (B.21)*, then:*

- *In the first period, the developer's choices align with the socially optimal outcomes only if limited liability does not bind. When it binds, the developer has an incentive to beta test the algorithm on a sample that is larger than socially optimal.*

- *In the second period, the developer's decisions regarding release and withdrawal are socially optimal.*

Figure 4 illustrates the release decisions under ex-ante Pigouvian taxes with limited liability for the first period. We ommit the case of the second period, since they coincide.

Figure 4: Time 1 testing decisions with Ex-Ante Taxes and Limited Liability



B.2.1 Proof of Proposition 7

Consider first the problem in the second period. If limited liability does not bind, then release and withdrawal incentives are aligned. Instead, suppose that limited liability binds. Then, the developer makes zero profits from releasing the algorithm, but still suffers from the externality when $\chi > 0$; when $\chi = 0$, the developer is indifferent and our tie-breaking selects withdrawal. Limited liability binds when

$$\begin{cases} u - \mathbb{E}^s_2(\phi^2_i) < N\mathbb{E}^s_2(\phi^2_e) \leq (N+\chi)\mathbb{E}^s_2(\phi^2_e), & \text{if } \mathcal{B} = 1 \\ u - \sigma^2_{s,i} < N\sigma^2_{s,e} \leq (N+\chi)\sigma^2_{s,e}, & \text{if } \mathcal{B} = 0. \end{cases}$$

So, when limited liability binds, the regulator also wants the algorithm to be withdrawn.

Turning to the problem of the first period, if limited liability does not bind, then incentives are aligned as before. If limited liability binds, then the developer's problem is

$$\max_{\mu_1} \left\{ (1-\beta) \left\{ u - \sigma_{s,i}^2 - \chi\sigma_{s,e}^2 - \frac{T_1^{ex-ante}}{\mu_1} \right\} \mu_1 + \beta\pi(\mu_1)\mathbb{E}_1^s[\max\{u - \mathbb{E}_2^s(\phi_i^2) - (N+\chi)\mathbb{E}_2^s(\phi_e^2), 0\}]N \tag{B.23}$$

For $\chi > 0$, the first order condition is given by

$$\begin{aligned}
&(1-\beta)\left\{u - \sigma_{s,i}^2 - \chi\sigma_{s,e}^2 - \frac{T_1^{ex-ante}}{\mu_1}\right\} + (1-\beta)\pi'(\mu_1)\Lambda^s N\left\{\sigma_{s,i}^2 + (N+\chi)\sigma_{s,e}^2 - u\right\} = 0 \\
\Leftrightarrow &\pi'(\mu_1)N\Lambda^s \frac{\sigma_{s,i}^2 + (N+\chi)\sigma_{s,e}^2 - u}{\frac{T_1^{ex-ante}}{\mu_1} + \sigma_{s,i}^2 + \chi\sigma_{s,e}^2 - u} = 1 \Leftrightarrow \pi'(\mu_1)N\Lambda^s \frac{\sigma_{s,i}^2 + (N+\chi)\sigma_{s,e}^2 - u}{\chi\sigma_{s,e}^2} = 1 \\
\Leftrightarrow &\mu_1 = \left[\alpha\Lambda^s \frac{\sigma_{s,i}^2 + (N+\chi)\sigma_{s,e}^2 - u}{\chi\sigma_{s,e}^2}\frac{N}{\kappa}\right]^{\frac{1}{1-\alpha}}\kappa.
\end{aligned}$$

When $\chi = 0$, the developer's current private cost of beta testing is zero under a binding limited-liability constraint, so the same conclusion holds. Thus, when limited liability binds, the developer generally beta tests the algorithm in a larger pool than the social optimal size.