NBER WORKING PAPER SERIES

REGULATING THE DIRECTION OF INNOVATION

Joshua S. Gans

Working Paper 32741
http://www.nber.org/papers/w32741

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2024, Revised January 2025

All correspondence to joshua.gans@utoronto.ca. The latest version of this paper is available at joshuagans.com. Thanks to Daron Acemoglu, Kevin Bryan and two anonymous referees for helpful comments. All errors remain my own. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Regulating the Direction of Innovation
Joshua S. Gans
NBER Working Paper No. 32741
July 2024, Revised January 2025
JEL No. L51, O33

ABSTRACT

This paper examines the regulation of technological innovation direction under uncertainty about potential harms. We develop a model with two competing technological paths and analyze various regulatory interventions. The optimal regulatory approach depends critically on the magnitude of potential harm relative to technological benefits. Our analysis reveals a motive to double down on harmful technologies in resource allocation across research paths, challenging common intuitions about diversification. We demonstrate that ex post regulatory instruments, particularly liability regimes, outperform ex ante restrictions in many scenarios. These insights have important implications for regulating emerging technologies like artificial intelligence, suggesting the need for informationally-responsive regulatory frameworks.

Joshua S. Gans
Rotman School of Management
University of Toronto
105 St. George Street
Toronto ON M5S 3E6
and NBER
joshua.gans@rotman.utoronto.ca

1 Introduction

While technological progress and innovation are generally presumed by economists to be about significant improvements to productivity, quality of life and general well-being, there is hardly any technology that does not involve harm (Acemoglu and Johnson, 2023). While it is often the case that those harms are outweighed by the benefits, there are prominent examples where this was not the case. The chemist Thomas J. Midgley was responsible for the invention of freon for refrigeration and anti-knock petrol. The former led to the release of chlorofluorocarbons and an expansion in the Ozone hole over the South Pole. The latter led to lead pollution with health and other consequences. In each case, governments determined that the harms outweighed the benefits and regulated a shift to alternative, but at the time, more costly technologies such as hydrofluorocarbons and unleaded petrol. For each of these alternatives, government regulation both stimulated adoption and, before that, scientific research to improve the viability of those alternatives.

In other areas, switching once the harms were understood has proved more challenging. At the beginning of the 20th century, automobiles were powered by both electric and internal combustion engines, with the latter 'winning' out in terms of adoption. However, for many decades, it has been known that petrol-powered cars resulted in more pollution than electric vehicles (McLaughlin, 1954). Yet despite some regulatory interventions, it has only recently occurred that some significant degree of switching to the alternative has occurred. Similarly, at the beginning of its deployment following World War II, nuclear power generation could be undertaken by heavy water or light water reactors. As Cowan (1990) documents, heavy water reactors initially had lower ongoing costs than light water ones and were likely to involve better safety outcomes. However, light water reactors were chosen for development in nuclear-powered submarines, and this drove private companies to invest in that path as a means of providing civilian power generation (the exception being Canada, which still has a heavy water reactor). Light-water reactors advanced technologically while other designs lagged behind. This made those reactors ubiquitous until safety concerns led to the halting of new nuclear power generation altogether following the Three Mile Island disaster in the 1970s (Bryan, 2017). The implication here is that one technological path can establish leadership, making it costly to switch to others should harm become apparent.

At present, the potential harms involved in technological innovation are being actively discussed and legislated with regard to artificial intelligence (AI). In the past decade, due to advances in computational statistics using machine learning, there have been significant advances that have allowed machines to engage more accurately and over a wider range of prediction tasks that were previously possible (Agrawal et al., 2022). These advances have the potential impact on many tasks currently performed by human workers as well as potentially to involve unintended consequences that may generate harm to security and political processes (Russell, 2019). Learning from these historical examples, some have argued that governments need to pre-emptively regulate both the adoption and direction of research associated with AI (Acemoglu, 2021). It has been noted that as AI develops along one path, it may become difficult to scale back adoption ex post (Acemoglu and Lensman, 2024). Thus, there is a call for pre-emptive regulation. While some of this regulation involves increased ex-ante assessments of the dangers associated with AI adoption, many of the proposals have involved interventions in pushing AI development toward outcomes that are 'human-centric' and more controllable (Brynjolfsson, 2022).[1]

Several theoretical papers in economics have established that market forces left alone may lead to distortions in the chosen direction of technological change away from paths that might be less harmful (Bryan and Lemus, 2017), or promote insufficient diversity in scientific effort across alternative paths (Acemoglu, 2011). Similarly, there are concerns more closely related to AI and automation that the market promotes less efficient and more harmful avenues for technological change (Acemoglu, 2023). While the mechanisms for such welfare sub-optimal technological change differ between these models, there is a common policy conclusion that the ex ante promotion of under-developed research paths would be welfare-improving. The most closely related is the work of Hopenhayn and Squintani (2021). They provide a detailed analysis of how competitive markets lead to an inefficient allocation of R&D resources, with excessive investment in high-return areas and a corresponding rent dissipation. They show how the difference between average and marginal returns leads to a market failure. The model developed here contains that particular result of Hopenhayn and Squintani (2021) and will be highlighted below. That said, the research question this paper aims to address is distinct.

This paper revisits these calls for ex ante regulation in light of the current debate regarding AI. The model presented here is inspired by that debate but is provided as an examination of a generic general-purpose technology. The modelling innovation relative to previous work explicitly takes into account that, at the outset, the potential harms regarding alternative technological paths or architectures are uncertain, and it is in the context of that uncertainty that any ex-ante regulatory intervention must be made. That said, there is the potential for learning about the harms, although it is argued that this learning occurs primarily on paths that achieve some degree of adoption so that harms, if any, can surface

[1] Jones (2024) and Trammell and Aschenbrenner (2024) examine the evaluation of broader AI risks over time and characterise the socially optimal approaches to managing such risk. They do not consider the regulatory trade-offs involved in promoting alternative technological paths as is done here.

or be dismissed. Thus, a regulator considering intervention must take into account not only uncertainty along paths ‘the market’ may be pursuing but also regarding the alternative.

Two broad findings arise from this examination. First, while it is the case that diversifying scientific resources across research paths can advance lagging technologies and reduce the costs of interventions should harms emerge on one path, that is not the only means by which scientific research can generate an option value against net harmful technologies being adopted. The other avenue is to double down and provide more resources to the leading path so that it advances to a sufficient level that its adoption involves net benefits even if harms should become apparent. This pushes regulators away from diversifying scientific resources across paths.

Second, because of this, heavy-handed ex-ante interventions, such as prohibiting adoption or scientific research along a path, can involve high costs relative to interventions that allow for ‘pricing in’ potential harm. Two such alternative interventions are considered: Pigouvian taxation and ex-post liability. It is shown that the latter, precisely because it not only internalises harm but also the prospect of harm when there is uncertainty, pushes agents in the economy to make more socially optimal adoption and research choices.

The paper proceeds as follows. The next section sets up the baseline model involving two potential technological paths that have differential appeals to different sectors in the economy and for which the potential harms are unknown for each path. Section 3 then characterises the decentralised equilibrium outcomes where both harms and their potential are not taken into account by private decision-makers. Section 4 then considers the socially optimal allocation and notes that changing scientific resource allocations across paths to take into account potential harm involves subtle optionality motives that may not be equivalent to more diversity in scientific research direction. Section 5 then examines and compares the alternative regulatory instruments, both ex ante and ex post, that can be deployed by regulators. Section 6 considers extensions, while a final section concludes.

2 Model Set-Up

The economy produces a unique final good from a continuum of sectors $i \in [0, 1]$ in each period t according to the production function:

$$Y(t) = \int_0^1 Y_i(t)di$$

A representative consumer has linear preferences over this final good and discounts the future at a rate of $\rho > 0$.

Firms in each sector choose the technology to adopt; either new technology architecture, A or B.[2] $Q_j(t) > 0$ denotes the quality of technology $j \in \{A, B\}$ at time t. For notational convenience, $x_i(t) = 1$ if sector i adopts architecture A in period t and $x_i(t) = 0$ if it, instead, adopts B.[3]

The impact of an architecture j's quality on sectoral output is captured by a parameter $\eta_{i,j} \geq 0$. For convenience, sectors are ordered so that $\eta_{i,A} = 1 - i$ and $\eta_{i,B} = i$; that is, the distribution of sector-specific productivities is uniform with half of the sectors having a higher quality-adjusted productivity in architecture A (B), $i \in [0, \frac{1}{2})$ ($i \in (\frac{1}{2}, 1]$).[4] Under these assumptions, the output of sector i can be written as:

$$Y_i(t) = x_i(t)(1-i)Q_A(t) + (1 - x_i(t))iQ_B(t)$$

Given the ordering assumptions on i, at any given time, a particular focus will be on $\hat{I}(t)$, which captures the productivity parameter for A such that all sectors, $i \leq \hat{I}(t)$ adopt A and all sectors $i > \hat{I}(t)$ adopt B, at time t.

2.1 Innovation

Improvements to the quality of any architecture j are enabled by innovation from a fixed pool of scientists, S; a continuum on the unit interval. Each scientist has one unit of effort that can be applied to one architecture or the other in each period. If scientists apply total innovative effort, $s_j(t)$, to j in period t, then with probability $h(s_j(t))$ this generates $Q_j(t+1) = Q_j(t) + \Delta$, for $\Delta > 0$, with $Q_j(t+1) = Q_j(t)$ otherwise. $h(.)$ is non-decreasing, concave, continuously differentiable and satisfies the Inada conditions, $\lim_{s_j \to 0} h'(s_j) = \infty$ and $\lim_{s_j \to 1} h'(s_j) = 0$.[5]

It is assumed that the innovation is rival but perfectly excludable for the incremental innovation; specifically, in selling an architecture with quality $Q_j(t)$, the previous quality level, $Q_j(t-1)$ is freely available to sectors. This is akin to a quality ladder model where

[2] Output for each sector in the absence of adopting the new technology is set at zero for simplicity. Acemoglu and Lensman (2024) consider a choice between an old and new technology where it is the new technology that carries potential external damage.

[3] The discussion here is of "sectoral" adoption of technology whereas, in the model below, it is individual firms within a sector choosing a technology. As each firm within a sector is identical, it will turn out, in equilibrium, that all firms in a sector make the same adoption choice in each period.

[4] This distributional assumption simplifies notation but does not play an important role in the results below. The important characteristic is that sectors can be ordered according to the magnitude of their comparative advantages of A or B adoption., holding quality constant.

[5] Note that, due to the assumptions here, $s_j(t)$, corresponds to both total effort and the share of scientific effort devoted to j.

each quality step is excludable for one period only.[6] Finally, it is assumed that, in the absence of any innovation, $Q_j = 0$.

2.2 Externalities

Note that A and B represent two distinct paths for a general-purpose technology that all sectors can use. In addition, both can potentially give rise to external effects. The size of the externality is assumed to depend on the number of sectors using a given architecture and their individual scale (using output as a proxy); that is, $E_{i,j} = -\eta_{i,j}\delta_j$ (in units of the final good) where $\delta_j \geq 0$ is common across sectors.[7] This harm is considered to be a pure externality, although, as with many technologies, it may not be a fully exogenous risk (i.e., the choices of bad actors may cause the harm to be greater). In that respect, harm here refers to harm caused by the technology's adoption including endogenous factors that may be hard to mitigate through policy interventions.

A key assumption is that the value of δ_j is ex ante uncertain. To keep the analysis simple, two assumptions are made. First, $\delta_j \in \{0, \delta\}$ for each j and $\mu \in (0, \frac{1}{2}]$ represents the (common) prior that a given architecture has $\delta_j = \delta$. The restriction that $\mu \leq \frac{1}{2}$, i.e., that harm is less likely than no harm, is made for analytical convenience, and the impact of relaxing it will be discussed below. Second, $\delta \geq \Delta$. This allows us to focus on the interesting case where if an architecture has only advanced modestly, it is optimal to abandon it.[8] Finally, to focus on a case of interest, suppose that the maximum realised externalities are such that $\delta \leq 2\Delta$.[9]

2.3 Time Structure

Timing in the model consists of two time periods, $t \in \{1, 2\}$. At the beginning of each period, scientists engage in research to determine the quality of a given architecture in that period. Following the realisation of outcomes from that research at the beginning of period $t = 1$, sectors choose whether to adopt one technology architecture or not at the end of period $t = 1$. Each sector chooses to adopt one of the architectures by paying a sector-specific price, $p_{i,j}(t)$. That adoption generates a signal of whether the technology architecture results in harm through a data-generating process described below. Then, the process repeats in period

[6] See O'Donoghue et al. (1998) for a discussion of this assumption.

[7] In a later section, the case where $E_{i,j}(t) = -\eta_{i,j}\delta_j Q_j(t)$ and damages scale with sectoral output is discussed.

[8] A case that is also the focus of Acemoglu and Lensman (2024). The impact of differing magnitudes of relative damage is discussed below.

[9] Below alternative assumptions are considered, noting that the core assumption here captures the most cases of interest.

2, except that adoption resolves any remaining uncertainty regarding potential harm. The particular focus of this paper is on the policy choices that are implemented at the beginning of period 2 following any signals generated at the end of period 1.

2.4 Learning by Using

In a decentralised context without regulation, scientists and firms do not take into account harmful externalities or their possibilities in their decisions. However, policymakers will consider this. Thus, both the social optimum and the implementation of policy options will consider how uncertainty regarding harm is resolved. Moreover, for certain policies, scientists and firms will be incentivised to take into account those signals. Therefore, it is important to specify the data-generating process for signals of harm realised before scientist allocation or adoption decisions are taken in period $t = 2$.

The core assumption made here is that actual adoption of a technology architecture is required to potentially receive a signal, $\sigma_j(I_j(1)) \in \{\varnothing, \delta\}$, i.e., no signal or a signal that the architecture involves harm of δ, where $I_j(1)$ is the scale of use of technology j in period $t = 1$.[10] That is, it is only by actually using a technology that it is possible to learn whether it is harmful. Recall that the prior probabilities that harm arises are the same, μ, across architectures. Moreover, these probabilities are independent. Given $I_A(1)$ and $I_B(1)$, if there is no signal of harm, i.e., $\sigma_j(I_j(1)) = \varnothing$, that probability is updated according to Bayes' rule as given by the following formulae:

$$\tilde{\mu}_A = \frac{\mu(1 - I_A(1))}{\mu(1 - I_A(1)) + (1 - \mu)}, \quad \tilde{\mu}_B = \frac{\mu(1 - I_B(1))}{\mu(1 - I_B(1)) + (1 - \mu)}$$

where $\tilde{\mu}_j$ is the (posterior) probability that damage could arise from the use of j in $t = 2$ given that it has not arisen before that point in $t = 1$. The learning process assumed here is that unless damage is observed during a period, the likelihood that damage will arise is updated by lowering the posterior probability that damage will arise. This is essentially an assumption that "no news is good news." More precisely, it is an assumption that a signal of harm reveals the state perfectly, while a signal there is no harm during a period involves the possibility of a false positive. Finally, observe that if there is no adoption of, say, B in period 1, then $\tilde{\mu}_B = \mu$ while $\tilde{\mu}_A = 0$; that is, if there is no indication in harm from A when it operates at "full" scale, then the probability that it is harmful in period 2 falls to 0. In contrast, if $I_A(1) = I_B(1) = \frac{1}{2}$ then $\tilde{\mu}_A = \tilde{\mu}_B = \frac{\mu}{2-\mu}$.

[10] Note that this differs from Acemoglu and Lensman (2024) who assume that learning can be passive. A relaxation of this core assumption will be considered below. See also Gans (2024) for a discussion of these differing learning assumptions.

2.5 Pricing

Each architecture is assumed to be marketed by a scientist-owned monopolist in each period. Given that all sectors adopt at least one technology if it is available, we simplify notation by lettering $\hat{I}_A(t) \equiv \hat{I}(t)$ and $\hat{I}_B(t) \equiv 1 - \hat{I}(t)$. Thus, given $\hat{I}(t)$, if the number of scientists contributing to architecture A is $s_A(t)$ and those scientists share in $v_A(t)$ if an innovation is generated at the beginning of period t. Here,

$$v_A(t) = \int_0^{\hat{I}(t)} \hat{p}_{i,j}(t)\, di, \quad v_B(t) = \int_{\hat{I}(t)}^1 \hat{p}_{i,j}(t)\, di$$

Given this, scientists can charge a sector-specific price, $p_{i,j}(t)$ where:

$$\hat{p}_{i,j}(t) = \eta_{i,j} Q_j(t) - \max\{\eta_{i,j} Q_j(t-1), \eta_{i,-j} Q_{-j}(t) - \hat{p}_{i,-j}(t)\}$$

This price is set equal to the sector-specific value of technology of quality $Q_j(t)$ relative to the best alternative of firms in that sector (i.e., either adopting the previous generation of the technology freely on the sector's preferred architecture, $Q_j(t-1)$ or paying for the latest generation with the alternative architecture, $Q_{-j}(t)$ at its prevailing price $\hat{p}_{i,-j}(t)$). This price is always positive as the technology improves, and it is for the preferred architecture of firms in that sector.

All scientists contributing research effort to a given architecture share equally in these commercial returns; thus, scientists contributing to A receive $\frac{v_A(t)}{s_A(t)}$ and those contributing to B receive $\frac{v_B(t)}{s_B(t)}$. This means that scientists with an innovation on a given architecture compete with other scientists with an innovation on the alternative architecture for adoption by a sector at any given time. As the competition is in price terms, then it is clear that, in equilibrium, at least, $\hat{p}_{i,A}(t)$ *or* $\hat{p}_{i,B}(t)$ will be zero for a given sector i (with both being zero if their characteristics are identical).

3 Decentralised Equilibrium

Without regulation, there is no incentive for scientists and firms to consider the external effects of technology architectures should they exist. Let $\{q_A(t), q_B(t)\}$ be the state of each architecture where $Q_A(t) = q_A(t)\Delta$ and $Q_B(t) = q_B(t)\Delta$ and each $q_j(t) \in \{0, 1, 2\}$. Note that because there are no constraints on firms switching architectures in each period, in equilibrium, the threshold, $\hat{I}(q_A(t), q_B(t))$, defining the sectors adopting A (viz B) is defined by:

$$(1-\hat{I}(q_A(t),q_B(t))Q_A(t)=\hat{I}(q_A(t),q_B(t))Q_B(t)\implies \hat{I}(q_A(t),q_B(t))=\frac{Q_A(t)}{Q_A(t)+Q_B(t)}$$

Another threshold of interest concerns the outside options of firms in each sector when considering adopting an architecture. For instance, a firm in sector i who does not purchase an improvement to architecture A will purchase the alternative architecture only if $(1-\eta_{i,A})Q_B(t)$ is greater than $\eta_{i,A}Q_A(t-1)$. Thus, we can define $\hat{I}_A(q_A(t),q_B(t))$ as:

$$\begin{aligned}\hat{I}_A(q_A(t),q_B(t))Q_A(t-1)&=(1-\hat{I}_A(q_A(t),q_B(t)))Q_B(t)\\ \implies \hat{I}_A(q_A(t),q_B(t))&=\frac{Q_A(t-1)}{Q_A(t-1)+Q_B(t)}\end{aligned}$$

Note that $\hat{I}_A(q_A(t),q_B(t))\leq\hat{I}(q_A(t),q_B(t))$ and, thus, sectors $i\leq\hat{I}_A(q_A(t),q_B(t))$ have an outside option from not purchasing the improved quality $Q_A(t)$ as continuing to use the previous quality (which is freely available) while those $i\in[\hat{I}_A(q_A(t),q_B(t)),\hat{I}(q_A(t),q_B(t))]$ have an outside option of purchasing the current best version of architecture B.

3.1 Scientist Allocation at $t=1$

As scientist and firm choices in period $t=1$ do not constrain their choices in the next period $t=2$, they make choices to optimise their current payoffs. Consider the decision of firms to adopt architecture A. Firms in sector i, if offered a technology with quality $Q_A(1)$, will only purchase it only if $p_{i,A}\leq\eta_{i,A}Q_A(1)$. Moreover, if the current best available quality for B is $Q_B(1)$, they will purchase A only if $p_{i,A}\leq\eta_{i,A}Q_A(1)-\eta_{i,B}Q_B(1)$.

Scientists only generate a return if they produce an innovation in the current period. The magnitude of that return will depend upon whether scientists working on the alternative architecture have produced an innovation or not. Suppose that innovations arise on both paths so that $Q_A(1)=Q_B(1)=\Delta$. In this case, as an advance is required for production, the only constraint on pricing technologies based on one architecture is that based on the other. Thus, a sector, i, will adopt architecture A if $\eta_{i,A}\Delta-\hat{p}_{i,A}\geq\eta_{i,B}\Delta-\hat{p}_{i,B}$. Let $\hat{I}(1,1)=\{i|\eta_{i,A}=\eta_{i,B}\}$. Note that for all sectors, $i\leq\hat{I}(1,1)$, $\hat{p}_{i,B}=0$. This is because, in equilibrium, for those sectors, B is not their preferred architecture, and, therefore, the price of B technology falls to zero. Thus, the maximum value of $\hat{p}_{i,A}$ is $(\eta_{i,A}-\eta_{i,B})\Delta$. A similar calculation shows that for $i>\hat{I}(1,1)$, $\hat{p}_{i,B}=(\eta_{i,B}-\eta_{i,A})\Delta$. Note that because

$Q_A(1) = Q_B(1)$, $\hat{I}(1,1) = \frac{1}{2}$. Thus, the total A-scientist returns are:

$$v_A(1,1) = \int_0^{\frac{1}{2}} (\eta_{i,A} - \eta_{i,B})\Delta \, di = \int_0^{\frac{1}{2}} (1-2i)\Delta \, di = \frac{\Delta}{4}$$

where the last equality uses the fact that $\eta_{i,A} = 1 - i$ and $\eta_{i,B} = i$. $v_B(1,1)$ has the same value. Suppose that the outcomes of $t = 1$ research are $Q_A(1) = \Delta$ while B has not advanced. Using a similar calculation for the case where both paths have advanced and noting that $\hat{I}(1,0) = 1$, we can derive:

$$v_A(1,0) = \int_0^{1} (1-i)\Delta \, di = \frac{\Delta}{2}$$

Clearly, without the competitive pressure from architecture B, $v_A(1,0)$ exceeds $v_A(1,1)$. In this case, $v_B(1,0) = 0$.

To determine the equilibrium allocation of scientists to each architecture, note that the expected returns to each research path are:

$$V_A(0,0) = \frac{h(s_A(1))}{s_A(1)} \Big(h(s_B(1)) v_A(1,1) + (1 - h(s_B(1))) v_A(1,0) \Big)$$

$$V_B(0,0) = \frac{h(s_B(1))}{s_B(1)} \Big(h(s_A(1)) v_B(1,1) + (1 - h(s_A(1))) v_B(1,0) \Big)$$

Each scientist will choose a path that earns them the highest expected return. Let $s_A = s$ and $s_B = 1 - s$. Then, as $v_A(.) = v_B(.)$, the only point where $V_A(0,0) = V_B(0,0)$ is where $s = \frac{1}{2}$. Thus, scientists will allocate themselves in equal numbers to each path in equilibrium, i.e., $\hat{s}(1) = \frac{1}{2}$, as otherwise scientists would have an incentive to switch to the path offering the highest average return.

3.2 Scientist Allocation at $t = 2$

At $t = 2$, there are four possible outcomes for the state of the architecture at the beginning of the period: $\{q_A, q_B\}$ could be $\{1,1\}$, $\{0,0\}$, $\{1,0\}$ or $\{0,1\}$. If a previous innovation occurred during $t = 1$, it is now in the public domain, and any sector can utilize it at no cost.

This has an impact on the pricing that can be achieved should there be innovations in $t = 2$; in particular, sectors with a strong preference for a particular architecture may prefer to continue to use the previous generation of technology rather than the alternative architecture, even if it has advanced. To see this, suppose that $Q_A(2) = Q_B(2) = 2\Delta$ (i.e., both architectures have advanced from a starting point where $(q_A, q_B) = (1,1)$). In a price-setting game, all sectors where $i \leq \hat{I}(2,2)$ adopt $Q_A(2)$, with $\hat{p}_{i,A} = \eta_{i,A}\Delta$ for $i \leq \hat{I}_A(2,2)$ and $\hat{p}_{i,A} = (\eta_{i,A} - \eta_{i,B})\Delta$ for $i \in [\hat{I}_A(2,2), \hat{I}(2,2)]$. This outcome is depicted in Figure 1. Note that because $Q_A(2) = Q_B(2)$, $\hat{I}(2,2) = \frac{1}{2}$. Thus, total A-scientist returns are:

$$v_A(2,2) = \int_0^{\hat{I}_A(2,2)} \eta_{i,A}\Delta\, di + \int_{\hat{I}_A(2,2)}^{\frac{1}{2}} (\eta_{i,A} - \eta_{i,B})2\Delta\, di = \int_0^{\frac{1}{3}} (1-i)\Delta\, di + \int_{\frac{1}{3}}^{\frac{1}{2}} (1-2i)2\Delta\, di = \frac{\Delta}{3}$$

where the last equality uses the fact that $\eta_{i,A} = 1 - i$, $\eta_{i,B} = i$, and that $\hat{I}_A(2,2) = \{i|(1-i)\Delta = i2\Delta\} = \frac{1}{3}$. Again, considering a starting point where $\{q_A, q_B\} = \{1,1\}$, if only one architecture, say A, has an innovation, then it is easy to calculate that $\hat{I} = \frac{2}{3}$ and $\hat{I}_A(2,1) = \frac{1}{2}$. Therefore,

$$v_A(2,1) = \int_0^{\frac{1}{2}} (1-i)\Delta\, di + \int_{\frac{1}{2}}^{\frac{2}{3}} ((1-i)2\Delta - i\Delta)\, di = \frac{5}{12}\Delta$$

(Note that we state $v_A(q_A, q_B)$ and $v_A(q_A, q_B)$ both as functions of the state $\{q_A, q_B\}$.) To determine the equilibrium allocation of scientists to each architecture, note that, at the beginning of $t = 1$ when $\{q_A, q_B\} = \{1,1\}$, the expected returns to each research path are:

$$V_A(1,1) = \frac{h(s_A)}{s_A}\Big(h(s_B)v_A(2,2) + (1 - h(s_B))v_A(2,1)\Big) = \frac{h(s)}{s}\tfrac{1}{2}\Big(1 - \tfrac{1}{3}h(1-s)\Big)\Delta$$

$$V_B(1,1) = \frac{h(s_B)}{s_B}\Big(h(s_A)v_B(2,2) + (1 - h(s_A))v_B(1,2)\Big) = \frac{h(1-s)}{1-s}\tfrac{1}{2}\Big(1 - \tfrac{1}{3}h(s)\Big)\Delta$$

Given the symmetry involved, the equilibrium allocation involves $\hat{s}(2) = \frac{1}{2}$.

Note that if $\{q_A, q_B\} = \{0,0\}$, then the possible outcomes at $t = 2$ are the same as those at $t = 1$. However, if, say, $\{q_A, q_B\} = \{1,0\}$, if both paths advance at $t = 2$, then, $v_A(2,1) = \frac{5}{12}\Delta$ as derived above while

$$v_B(2,1) = \int_{\frac{2}{3}}^{1} (i\Delta - (1-i)2\Delta)\, di = \frac{\Delta}{6}$$

If only one path advances at $t = 2$, then $v_B(1,1) = \frac{\Delta}{4}$ as derived above while

$$v_A(2,0) = \int_0^1 (1-i)\Delta\, di = \frac{\Delta}{2}$$

Given this, the expected returns to each research path are:

$$V_A(1,0) = \frac{h(s_A)}{s_A}\Big(h(s_B)v_A(2,1) + (1 - h(s_B))v_A(2,0)\Big) = \frac{h(s)}{s}\tfrac{1}{2}\Big(1 - \tfrac{1}{6}h(1-s)\Big)\Delta$$

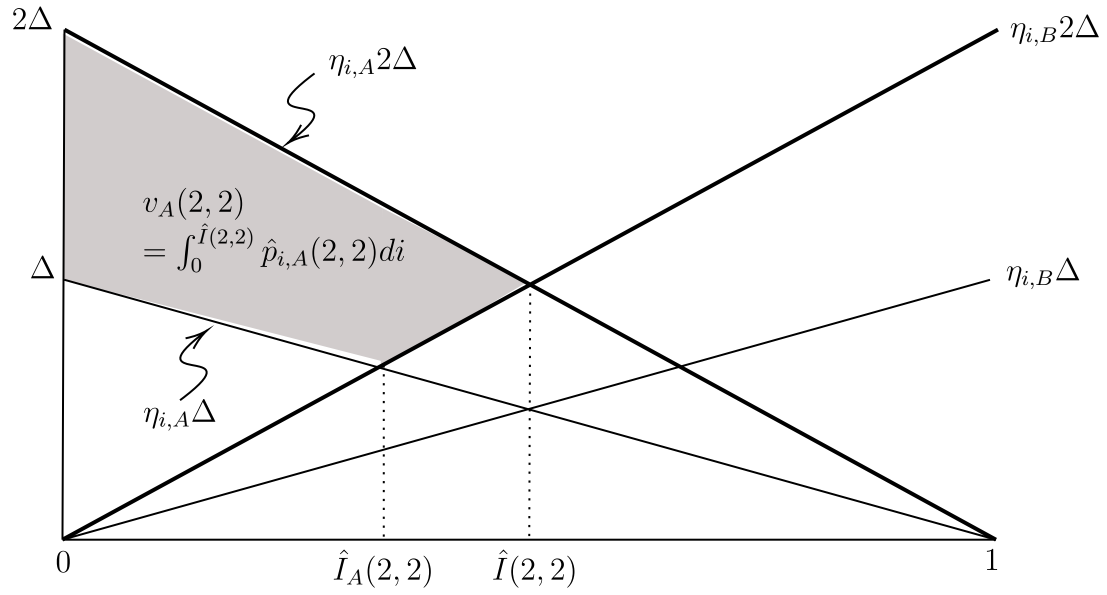

Figure 1: **Scientist Returns to Architecture** A **for** $Q_A(2) = Q_B(2) = 2\Delta$

$$V_B(1,0) = \frac{h(s_B)}{s_B}\Big(h(s_A)v_B(2,1) + (1 - h(s_A))v_B(1,1)\Big) = \frac{h(1-s)}{1-s}\tfrac{1}{4}\Big(1 - \tfrac{1}{3}h(s)\Big)\Delta$$

As $V_A(1,0) > V_B(1,0)$ for $s = \frac{1}{2}$, it is clear that the equilibrium involves $\hat{s}(2) > \frac{1}{2}$. That is, more scientists are allocated to the leading architecture than the lagging one. Intuitively, the more advanced path can potentially earn a higher return than the other path regardless of whether the other path advances or not, whereas the less advanced path can, at best, catch up commercially. Hence, more scientists will choose to research in the more advanced path. Interestingly, this outcome arises even though there are no constraints on sectors switching their adopted architecture from their previously chosen architecture nor any constraints on scientists in switching research paths.[11]

4 Socially Optimal Outcomes

Having analysed the decentralised equilibrium where scientists and firms choose their allocation and adoption decisions, respectively, we now turn to examine socially optimal outcomes. The purpose of this is to characterise the outcomes a regulator will attempt to achieve using various policy instruments that will be examined in the following section. Here, the socially

[11] In contrast to other models where innovation takes place on technologies at different steps on a quality ladder, no scientist or firm owns or is tied to a particular architecture. Moreover, past innovations are freely available. It could be imagined that there were switching costs faced by scientists. In that case, there would be a limit on adjusting scientific allocations even after uncertainty over harm is resolved, and there are asymmetries in progress along each research path. Here, it is assumed that the time frame for the two periods is sufficiently long and that scientist switching costs do not play a significant role.

optimal outcomes are characterised and the areas where these differ from the decentralised outcome are highlighted.

The social planner chooses $\{s_A(t), s_B(t), \{x_i(t)\}_{i\in[0,1]}\}_{t=1,2}$ to maximise:

$$V_S = \sum_{t=1}^{2} \rho^{t-1}\mathbb{E}[Y(t) - E(t)]$$

where $\rho \in (0, 1]$ is the planner's discount factor and $E(t) = \int_0^1 \eta_{i,j}\delta_j di$. Note, however, that since the level of the externality is unknown at $t = 1$, allocations at that time will be based on its expected value, while those at $t = 2$ may take into account the externality's realised value.

To analyse the socially optimal outcomes, we work backwards from $t = 2$. To do this, note our earlier assumptions that $2\Delta > \delta$ and $\mu \leq \frac{1}{2}$ imply that $\Delta > \mu\delta$; that is, if there is uncertainty regarding harm, it is socially optimal to adopt an architecture. This means that there is no cost to adopting technologies that have advanced in period $t = 1$ in the face of no additional information regarding potential harms. Below, the implications of this assumption will be discussed, but for the moment, this implies that the planner will engage in some research and adoption of technologies that successfully advance in period $t = 1$. Relaxing this assumption adds complexity to the analysis but does not alter the qualitative choices of the planner so long as μ is not too high.

4.1 Adoption at $t = 2$

Consider the social planner's choice of which sectors should adopt which architecture following a realisation of the magnitude of any externalities. Suppose that the expected values of the relevant parameters are $\{\mathbb{E}[\delta_A], \mathbb{E}[\delta_B]\}$; recalling that these are common across sectors and conditional on adoption at $t = 1$. Let I denote the threshold whereby sectors, $i \leq I$ adopt A and the remainder adopt B. In this case, the social planner chooses I to maximise:

$$\int_0^I \eta_{i,A}(Q_A(2) - \mathbb{E}[\delta_A])di + \int_I^1 \eta_{i,B}(Q_B(2) - \mathbb{E}[\delta_B])di$$

The optimum has a simple structure. If, for any architecture, j, $Q_j(2) < \delta_j$, then it is not optimal to adopt that architecture. Otherwise, the adoption of that architecture by some sectors is optimal.

If, for one architecture, say A, $Q_A(2) \geq \mathbb{E}[\delta_A]$ while for the other $Q_B(2) < \mathbb{E}[\delta_B]$, then all sectors adopt architecture A. This results in social welfare at $t = 2$ of $\frac{1}{2}(Q_A(2) - \mathbb{E}[\delta_A]))$.

By contrast, if $Q_j(2) > \delta_j$ for both architectures, j, then the optimum satisfies:

$$I^* = \frac{Q_A(2) - \mathbb{E}[\delta_A]}{Q_A(2) - \mathbb{E}[\delta_A] + Q_B(2) - \mathbb{E}[\delta_B]}$$

This results in social welfare realised at $t = 2$ of:

$$v_S(q_A, q_B, \mathbb{E}[\delta_A], \mathbb{E}[\delta_B]) = \frac{1}{2}\left(\frac{(Q_A(2)-\mathbb{E}[\delta_A])^2+(Q_B(2)-\mathbb{E}[\delta_B])^2+(Q_A(2)-\mathbb{E}[\delta_A])(Q_B(2)-\mathbb{E}[\delta_B])}{Q_A(2)-\mathbb{E}[\delta_A]+Q_B(2)-\mathbb{E}[\delta_B]}\right)$$

As can readily be seen, the planner takes into account their knowledge of potential harms in choosing the allocation at $t = 2$, whereas, in a decentralised equilibrium, these do not factor into the resulting outcomes. Thus, there is *distorted adoption* relative to the social optimum.

4.2 Scientist Allocation at $t = 2$

At the beginning of period $t = 2$, the planner will know the state of each architecture $(q_A, q_B) \in \{(0,0), (1,0), (0,1), (1,1)\}$ and the expected potential harms, $\mathbb{E}[\delta_A|(q_A, q_B), \sigma_A(I_A(1))]$ and $\mathbb{E}[\delta_B|(q_A, q_B), \sigma_B(I_B(1))]$ that are conditional as the technology state and its adoption impact on the signal of harm to the planner.

There are three potential cases of interest: two symmetric and one asymmetric. First, if $(q_A, q_B) = (0,0)$, then neither technology has advanced or was adopted in period $t = 1$. Therefore, $\mathbb{E}[\delta_A|(0,0), \varnothing] = \mathbb{E}[\delta_B|(0,0), \varnothing] = \mu\delta$ and the planner will act as they did in period $t = 1$, setting $s^*(2) = \frac{1}{2}$ and adopting any technology that advances. Note that this involves the same scientist allocation and $t = 2$ adoption outcomes as the decentralised equilibrium.

Second, if $(q_A, q_B) = (1,1)$, then both technologies advanced and were adopted in period $t = 1$. For j, either $\mathbb{E}[\delta_j|(1,1), \varnothing] = \tilde{\mu}\delta$ or $\mathbb{E}[\delta_j|(1,1), \delta] = \delta$. It is only where expected harm between the two architectures differ that the scientist allocations and adoption outcomes differ from the decentralised equilibrium. To see this, suppose that $\mathbb{E}[\delta_B] = \delta$ while $\mathbb{E}[\delta_A] = \tilde{\mu}\delta$. It is easy to see that if, in $t = 2$, if B does not advance, it will not be adopted and $I^*(2) = 1$. If only B advances, $I^*(2) > \frac{1}{2}$ only if $\Delta - \tilde{\mu}\delta > 2\Delta - \delta$ or $\mu < \frac{2(\delta - \Delta)}{2\delta - \Delta}$. This can hold given our assumption that $\mu < \frac{1}{2}$ so long as $\frac{3}{2}\Delta \geq \delta$. Regardless, an advance in B generates social value even if B is known to be harmful and, therefore, $s^*(2) < 1$.

The following proposition confirms that $s^*(2) > \frac{1}{2}$ as it is the case that allocating scientists to the advancement of A (a technology with unknown harm) would involve a higher social return than allocating scientists to B.

Proposition 1 *Suppose that $(q_A, q_B) = (1,1)$, $\mathbb{E}[\delta_B] = \delta$ and $\mathbb{E}[\delta_A] = \tilde{\mu}\delta$, then the planner sets $s^*(2) \geq \frac{1}{2}$ for all $\mu \in [0, \frac{1}{2}]$.*

The proof is in the appendix. It shows that the only asymmetric effect of moving scientists from researching on B to A is that the latter reduces the probability of generating $\mathbb{E}[v_S(1,2),\delta_A,\delta)]$ while the former increases the probability of generating $\mathbb{E}[v_S(2,1),\delta_A,\delta)]$ where expectations are taken with respect to δ_A based on a posterior probability of $\tilde{\mu}$. Note, however, that if only A advances, B will not be adopted and, in fact, $\mathbb{E}[v_S(2,1),\delta_A,\delta)] = \Delta - \frac{1}{2}\tilde{\mu}\delta)$ (with $I^*(2) = 1$). By contrast, as was already shown, if only B advances, both technologies are used. The proof of the proposition shows, importantly, that $\mathbb{E}[v_S(1,2,\delta_A,\delta)] < \mathbb{E}[v_S(2,1),\delta_A,\delta)]$; that is, the social planner, if choosing between advancing each technology prefers to advance the technology with unknown harm.

The result in Proposition 1, however, embeds a non-intuitive effect: the planner has an incentive to devote some scientific resources to advancing the harmful technology. This is because, as $2\Delta > \delta$, if B advances, all sectors, including those that prefer B the most, adopt a technology aligned with their preferences, whereas, if only A advances, there is misalignment. This *doubling-down effect* arises precisely because the sectors that most value B above A would receive very little value from an advanced A and a large value from an advanced B given that it is only where it has advanced that it is adopted at all given its known harm. By contrast, those sectors who prefer A are guaranteed some value regardless.

This effect that research can generate option value in mitigating a known harmful architecture arises in a more pronounced fashion in the third case where only one architecture, say A advances in period $t = 1$; i.e., $(q_A, q_B) = (1, 0)$. In that case, only A would have been adopted in period 1. Thus, $\mathbb{E}[\delta_A|(1,0),\delta] = \delta$ or $\mathbb{E}[\delta_A|(1,0),\varnothing] = 0$ while $\mathbb{E}[\delta_B|(1,0),\sigma_B(0)] = \mu\delta$.

Proposition 2 *Suppose that $(q_A, q_B) = (1, 0)$, then the planner sets $s^*(2) > \frac{1}{2}$ if A is known not to be harmful or if μ is sufficiently high (i.e., $\mu > \frac{\delta-\Delta}{\delta}$). If A is known to be harmful and μ is sufficiently low (i.e., $\mu < \frac{\delta-\Delta}{\delta}$), then $s^*(2) < \frac{1}{2}$.*

The proof demonstrates that the allocation of scientists is driven by the expected social value if only A advances versus if only B advances. When A is known to be safe, then it will be adopted in either scenario, whereas B is only adopted if it advances and, even in this case, more sectors adopt A. Thus, the returns to researching to advance A are correspondingly higher, although the planner diversifies the research effort despite B's harm being unknown. Interestingly, when A is known to be harmful, the planner might still double down on research towards A if the likelihood, μ, of B's potential harm is high. This is because the planner is trading off scenarios where only A is used (at high value with harm) or only B is used because A is too harmful to use if it does not advance. It is only when B is forecast to likely be safe that the planner favours B in research, given A's known harm.

Propositions 1 and 2 demonstrate that the planner will engage in research to advance an architecture known or more likely to be harmful at the expense of research on the alternative. This *doubling-down effect* arises because, in the absence of an advance in that technology, it will not be adopted, and those sectors that value it the most will generate far less value. This can most clearly be seen in Proposition 2, where only one technology has advanced, and the planner has a clear signal of its harmfulness. However, in this case, the planner, in order to boost the productivity of the sectors underserved by the leading architecture, allocates scientific resources to the riskier technology to *diversify* research outcomes.

4.3 Scientist Allocation and Technology Adoption at $t = 1$

It has already been noted that because μ is assumed to be less than $\frac{1}{2}$, that $\Delta \geq \mu\delta$. This means that a planner who did not have any information regarding the likely harm from an architecture will adopt the technology in period $t = 1$ as there is no cost to so doing. This implies that the planner will also allocate scientists to advance both architectures, setting $s^*(1) = \frac{1}{2}$. In effect, learning is not costly from a social perspective, even though technology adoption can possibly result in harm in period $t = 1$ even if this only becomes apparent at the end of period $t = 2$.

It is instructive to discuss what happens if μ can exceed $\frac{1}{2}$ and, hence, $\Delta < \mu\delta$, making the expected social return from the adoption of a technology that advances during $t = 1$ negative. Even in this case, it may be worthwhile both researching towards and adopting technology at $t = 1$ as this creates an option value to advance the technology further. Here, we discuss this possibility in more detail.

To begin, suppose that after research is conducted at $t = 1$, only A advances and is available for adoption in that period. The expected social value in $t = 1$ is $\frac{1}{2}(\Delta - \mu\delta)$. If A is adopted, then the expected social value in $t = 2$ is:

$$(1 - \mu)\rho\mathbb{E}[V_S(1, 0, 0, \delta_B)] + \mu\rho\mathbb{E}[V_S(1, 0, \delta, \delta_B)]$$

Note that if $\Delta < \mu\delta$, as B does not advance in period $t = 1$, it is not worthwhile to set $s < 1$ at $t = 2$ (as B would never be adopted). Therefore, $\mathbb{E}[V_S(1, 0, 0, \delta_B)] = h(1)\Delta + (1 - h(1))\frac{1}{2}\Delta$ and $\mathbb{E}[V^*(1, 0, \delta, \delta_B)] = h(1)\frac{1}{2}(2\Delta - \delta)$ (as A is not adopted at $t = 2$ if it does not advance in that period). Therefore, if $\Delta < \mu\delta$, it will only be worthwhile adopting A in period $t = 1$ if:

$$\mu \leq \frac{\delta + 2\rho\mathbb{E}[V_S(1, 0, 0, \delta_B)]}{\Delta + 2\rho(\mathbb{E}[V^*(1, 0, 0, \delta_B)] - \mathbb{E}[V_S(1, 0, \delta, \delta_B)])} = \frac{(1 + \rho(1 + h(1))\Delta}{(1 - h(1))\rho\Delta + (1 + h(1)\rho)\delta}$$

Therefore, even where $\Delta < \mu\delta$, for μ not too high, there is option value in adopting A at

$t = 1$ in order to learn about the potential harms from adoption at $t = 2$. Note that, even if such harms exist, continuing research into A is valuable because if A advances, it is still worthwhile adopting it at $t = 2$ per the double-down effect explored above.

What if both A and B advance during $t = 1$? The above analysis demonstrates that it is potentially worthwhile to adopt at least one architecture to learn about its harms. Is it potentially worthwhile to adopt both? The $t = 1$ cost of doing this is $\frac{3}{4}(\mu\delta - \Delta)$, which is higher than the cost of adopting both architectures. The value of adopting both is that learning about the potential harms of each is possible, which has value as either one or both might advance at $t = 2$. When $\Delta > \tilde{\mu}\delta$, learning allows the technology to be adopted in period $t = 2$ and is, thus, valuable. If this condition does not hold, learning does not have value as it can only lead to the certainty that the technology is harmful to avoid adoption, which is the default in this case, regardless. Thus, so long as μ is not too high, it is worthwhile to adopt both technologies during $t = 1$ to learn about their potential harms and also provide value to sectors with extreme preferences. Further analysis of this case is involved, and so is not provided here.

In summary, for the base case where $\mu < \frac{1}{2}$, the allocation of scientists to research paths involves $s^*(1) = \frac{1}{2}$. Even where adoption is costly in the face of uncertainty, the planner will choose to diversify research and adoption in order to learn about harms and make adjustments in $t = 2$ so long as μ is not too high.[12]

4.4 Comparison with the Decentralised Equilibrium

There are three dimensions upon which the decentralised equilibrium generates different outcomes than those that would be chosen by a social planner. Two of these are driven by the harm or potential harm that are externalities from the perspective of scientists and firms and not taken into account by them. One of these involves distortions to the adoption decisions of firms in period $t = 2$ while the remaining dimensions generate distinct distortions to the allocation of scientists to research paths in period $t = 2$. As was already noted, as $\Delta > \mu\delta$, the decentralised choices in $t = 1$ are the same as the planner's choices.

First, as already noted, firms do not take into account harm or potential harm when making decisions to adopt either technology. If one architecture involves a higher known or likely harm than another, that architecture will be adopted by too many sectors relative to the social optimum. Even if architectures involve the same harm or expected harm, There will be over-adoption of the architecture that is less advanced. These cases include the fact

[12] This mirrors conclusions found by Gans (2024) regarding the value of learning by using to determine the extent of harm from adopting a technology.

that, in a decentralised equilibrium, lower quality but harmful architectures are adopted when they should not be.

Second, these considerations impact the allocation of scientists to research paths in period $t = 2$. Proposition 1 considers a situation where both architectures have advanced and will be utilised in period $t = 2$ in a decentralised equilibrium but where asymmetric harms cause the planner to aim to reduce the adoption of the known harmful technology. In this case, the decentralised equilibrium would involve $\hat{s}(2) = \frac{1}{2}$, but as already shown, the planner wants to favour the technology with uncertain harm and allocate more resources to its advancement. Thus, the decentralised equilibrium involves too few scientific resources devoted to the (likely) less harmful technology.

When only one architecture has advanced in period $t = 1$, more resources will be allocated to its advancement in the decentralised equilibrium. Proposition 2 demonstrates that this same direction will be favoured by the social planner under certain circumstances. Nonetheless, the double-down and diversity effects remain something that is absent from scientists' decisions.

Finally, the model contains a distortion to the allocation of scientists identified by Hopenhayn and Squintani (2021) that arises even if there are no potential or actual harms. In a decentralised allocation, scientists choose their research paths based on relative average expected returns, whereas the socially optimal allocation depends on relative marginal expected returns.

To see this, suppose that $\mu = 0$. This distortion only arises if only one architecture has advanced in period $t = 1$; i.e., when $q_A = q_B$, $s^*(1) = \hat{s}(1) = \frac{1}{2}$. If, say, $q_A = 1$, while q_B remains at 0, then, the allocation of scientists in the decentralised equilibrium is determined by:

$$\frac{V_A(1,0)}{V_B(1,0)} = 1 \implies \frac{h(s)/s}{h(1-s)/(1-s)} = \frac{h(s)v_B(2,0) + (1-h(s))v_B(1,0)}{h(1-s)v_A(1,1) + (1-h(1-s))v_A(1,0)}$$

By contrast, the socially optimal allocation is determined by:

$$\frac{h'(s)}{h'(1-s)} = \frac{h(s)(v_S(2,0) - v_S(2,1) + (1-h(s))(v_S(1,0) - v_S(1,1))}{h(1-s)(v_S(1,1) - v_S(2,1)) + (1-h(1-s))(v_S(1,0) - v_S(2,0))}$$

It can be demonstrated that, in this case, the RHS of each expression is equal to:

$$\frac{1 - \frac{1}{3}h(s)}{2 - \frac{1}{3}h(1-s)}$$

Thus, the differences in the scientist allocation will be driven by the LHS of each expression.

What this implies is that, compared to the decentralised equilibrium, asymmetries lead to an amplification in the concentration of scientists to a leading sector. In the context of the present model, it can be demonstrated that:

Lemma 1 *If $\frac{h'(s)}{h(s)/s}$ is decreasing in s, then (i) if $s^* = \frac{1}{2}$, $\hat{s} = s^*$; and (ii) if $s^* > (<)\frac{1}{2}$, $\hat{s} > (<)s^*$.*

As Hopenhayn and Squintani (2021) show, there is too much entry as scientists do not take into account what their allocation of research effort has on other scientists. When scientists choose between alternative research paths, they consider the relative *average* probabilities of success on that path and so neglect the impact that their own choice has on the relative *marginal* probabilities of success. In particular, the probability of success for each path depends on each scientist's choice, but when a scientist switches from the lagging to the leading path, the marginal negative impact on the less advanced path is higher than the marginal positive impact on the probability that the more advanced path succeeds in advancing further. The social planner takes these impacts into account, while scientists only consider the latter positive impact on the path they pursue.[13] Note, however, that with a functional form assumption for $h(.)$, the distortion caused by scientists choosing their research path based on the average probability of success as opposed to the marginal probability that governs the socially optimal allocation can be removed; that is, $\frac{h'(s)}{h(s)/s} = a$. What this shows is for a case where the distortion from the decentralised allocation of scientists is not present, the impact of direct externalities is clearer.

This section demonstrates that, from the perspective of allocating scientific resources, there are no clear courses of action when uncertainty is resolved and even more subtle interactions when uncertainty remains. This means that there are challenges and trade-offs involved between alternative intervention instruments, which will be addressed next.

[13] This result that there is socially too high a degree of concentration on the leading research path than the lagging in a decentralised equilibrium arises from a distinct rationale than similar results in the literature. For instance, Acemoglu (2011) provides a model where a leading path has an advantage in that products along this path can be commercialised immediately while those on a lagging path may have to wait until some event, such as changing tastes, makes them commercially viable. The market then underprovides diversity due to the asymmetric nature of private appropriation along the two competing paths. Bryan and Lemus (2017) show that racing distortions, the value of being first to advance a technological path, confers a negative externality on research incentives on the other path that is not taken into account by scientists, causing them to allocate too many resources to "quick wins." Similarly, scientists do not necessarily place sufficient value on the depth of research paths – in terms of how many future innovations they might yield – and so may inefficiently devote too many scientific resources to paths where innovation opportunities are front-loaded in time. While Bryan and Lemus (2017)'s results do not specifically address issues of the optimal level of diversity in research, they do identify some key distortions that arise. It is easy to imagine that a different model specification could emphasise a different externality in scientist allocation without any change to the broad conclusions reached below.

5 Regulatory Interventions

We now turn to consider how various regulatory instruments can be used to generate more socially efficient outcomes, given the inefficiencies examined above. First, it is demonstrated that perfectly applied sector-specific taxes targeting only technology providers can achieve the socially optimal outcome under certain assumptions. Second, various instruments are evaluated under restrictions that can impact the practical regulatory environment. These restrictions are limited liability and informational requirements that trigger intervention. Four policy options are available: (i) a ban on the adoption of an architecture, (ii) a ban on research advancing an architecture, (iii) a tax on the adoption of an architecture of $E_{i,j}$ and (iv) subjecting adopters to ex post liability of $E_{i,j}$ if harm occurs. The first three of these are forms of ex ante regulation as intervention is prior to the realisation of research outcomes in period 2. The final outcome is an ex post regulation because regulation only proceeds after all activity has occurred. This section will evaluate and compare each of these options, which are contingent on the receipt of various policy signals outlined next. As is standard in the regulation literature, it will be assumed that the regulator's goal is to achieve improvements in social welfare outcomes.

5.1 Optimal Intervention

The analysis of socially optimal outcomes demonstrated that both the adoption of technologies by sectors and the allocation of scientists to research paths were distorted by the harm and potential harm associated with each architecture. While some of the distortions to scientist allocation are generated by the Hopenhayn and Squintani (2021) effects that arise even in the absence of harmful externalities, the following proposition demonstrates that when those effects are not present, the only decisions a regulator needs to target are adoption decisions.

Proposition 3 *Suppose that $h(s_j) = s_J^a$ for $a \in (0, 1)$. If the provider of an architecture, j, to a sector, i, pays a tax, $\eta_{i,j}\tau_j = \eta_{i,j}\mathbb{E}[\delta_j|\sigma_j(I_j(1)]$ (expected harm at the beginning of $t = 2$, then expected social welfare at the beginning of $t = 2$, contingent on signals received, is maximised.*

This proposition examines the imposition of a tax equal to expected harm from adoption of an architecture. Note that the tax is committed to by the regulator at the beginning of period $t = 2$.[14] If harm is known to arise from the adoption of an architecture, each sector

[14] Recall that there are no inefficiencies in scientist allocation and adoption in period $t = 1$.

pays a sector-specific tax of $\eta_{i,j}\delta$. However, even if harm is expected, rather than known with certainty, the sector-specific tax is $\eta_{i,j}\tilde{\mu}\delta$. Importantly, this expected tax is not adjusted ex post once actual harm is revealed at the end of $t = 2$.[15]

The proof of this proposition (in the appendix) highlights some potentially non-trivial aspects of this result. Importantly, intervention is only required via a tax on technology providers. That not only means the adoption decisions are socially optimal but that the scientist allocation to research in the face of those prospective taxes is optimal as well. The latter arises because the functional form assumption on $h(.)$ suppresses the Hopenhayn and Squintani (2021) effect but also that the relative expected payoffs to scientists are such that their choice of which research path to take is optimal. This latter result is not trivial since the profits from technology adoption are mitigated by both inter-architecture competition but in-architecture competition. The latter represents a potential distortion in the adoption of a higher quality technology when a lower quality technology, facing the same tax, is present. For an adopting firm in a sector more favourable disposed to a particular architecture, the price paid by those firms is $p_{i,j} = 2\Delta - (\Delta - \tau_{i,j})$ where the second term is the firm's surplus from adopting the lower quality technology (that is freely available but whose providers are still subject to the tax of $\tau_{i,j}$. Thus, the high quality technology provider's profits from that sector are $p_{i,j} - \tau_{i,j} = \Delta$ (independent of the tax). The proof demonstrates that, despite this, the taxes induce a socially optimal allocation of scientific resources.

5.2 Practical Constraints

Two practical constraints—informational limitations and limited liability—can impact the implementation of various policy instruments. We discuss each in turn.

Information limitations arise because, at the end of $t = 1$, it may not be known if a technology is harmful or not. The optimal approach is to base a tax on providers based on expected harm. However, this requires that a tax by imposed even on a provider whose technology has not been previously adopted (i.e., where $q_j = 0$). Moreover, it is possible that intervention may be triggered by relatively "good news" if one architecture only advances but receives no adverse signal. In this case, the posterior probability of harm is lower for the architecture that has advanced, raising the question of whether intervention would be welfare-improving directed at the architecture that has not advanced.It is arguable that this might be practically or politically difficult to do. Thus, we evaluate policy options by assuming that taxes or liability can only be triggered if there is evidence of known harm, i.e., if $\tilde{\mu} = 1$, in which case $\tau_j = \delta$. When there is no adverse signal, either because neither

[15] It could be adjusted but it would have to be adjusted downwards (i.e., through a rebate) should no harm be proven to arise.

architecture advances at $t = 1$ (and is therefore not adopted) or when both advance without an adverse signal despite adoption (leading to $\tilde{\mu}_j < \mu$ for both A and B), it is assumed that there is no basis for intervention.[16]

Given this, suppose, without loss in generality, that A has advanced in period 1 (i.e., $q_A = 1$). This implies that there are three possible scenarios at the end of period 1 that would trigger potential interventions aimed at architecture A at the beginning of period 2:

1. (B harmful) $q_B = 1$ and $\delta_A = \delta_B = \delta$: B has advanced and both are known to be harmful;

2. (B harm uncertain) $q_B = 1$ and $\Pr[\delta_B = \delta] = \tilde{\mu}_B < 1$ and $\delta_A = \delta$: B has advanced but is not known to be harmful while A is known to be harmful;

3. (B has not advanced) $q_B = 0$ and $\delta_A = \delta$: B has not advanced and A is known to be harmful.

Note that it is also possible that $q_B = 1$ and $\delta_B = \delta$ and $\Pr[\delta_A = \delta] = \tilde{\mu}_A < 1$ where B has advanced and is known to be harmful while A's harm is uncertain. However, this is technically equivalent to Scenario 2 above (B harm uncertain), and so is omitted from consideration. In what follows, each policy option is considered for each of these four scenarios.[17]

One way of dealing with informational limitations is only to enact policy interventions ex post; for instance, through a liability regime. This leads to a second practical constraint of *limited liability.* If an architecture is low quality and found to be harmful, then the optimal $\tau_j = \delta$. However, as $\Delta < \delta$, a technology provider would not pay that tax if it was imposed ex post. It is assumed, therefore, in this case, that limited liability binds and the sanction to the firm is limited to δ.

In what follows, four policy instruments are evaluating assuming both of these constraints hold. Table 1 summarises the outcomes in terms of the allocation of scientists and the adoption of A and B architectures in period 2 under various scenarios. Note that endogenous

[16] For a recent investigation into how learning impacts on regulatory choices see Koh and Sanguanmoo (2024)

[17] Ot is worthwhile noting another type of intervention for the scenario where A only advances and does not receive a signal of harm. Given that this scenario involves $I_1 = 1$, it is known with certainty that A is 'safe' (i.e., $\tilde{\mu}_A = 0$) whereas the probability that B is harmful remains at μ. In this situation, policymakers could preemptively enact regulations targeting B. For instance, regulators could prevent any potential harm from B by either banning its adoption, banning further research on B or taxing B as if the harm would occur. In each case, $I_2 = 1$ (with all sectors adopting A) and $s_A = 1$ as a result. Expected social welfare would be $h(1) \int_0^1 2\Delta \, di + (1 - h(1)) \int_0^1 \Delta \, di = (1 + h(1))\frac{1}{2}\Delta$ in each case. By contrast, if there was no intervention, if B advances, this leads to some value from its use but also a probability, μ, that there is harm resulting from that use. If μ and/or δ is relatively high, it may be worthwhile to intervene to 'not take a risk' on B. In what follows, this type of intervention is set aside, although it does highlight the continuing theme here that the trade-offs regarding intervention can be subtle.

Scenario	B harmful	B harm uncertain	B not advanced
$(q_A, q_B, \delta_A, \delta_B)$	$(1, 1, \delta, \delta)$	$(1, 1, \delta, \delta_B)$	$(1, 0, \delta, \delta_B)$
Adoption Ban	No Adoption of A or B	$I = \hat{s} = 0 \implies$ wp $h(1)$: B advances	$I = \hat{s} = 0 \implies$ wp $h(1)$: B advances
Research Prohibition	$s_A = s_B = 0$ $\hat{I} = \frac{1}{2}$	$s_A = 0$, $\hat{s}_B = 1 \implies$ wp $1 - h(1)$: $\hat{I} = \frac{1}{2}$ $h(1)$: $\hat{I} = \frac{1}{3}$	$s_A = 0$, $\hat{s}_B = 1 \implies$ wp $1 - h(1)$: $\hat{I} = 1$ $h(1)$: $\hat{I} = \frac{1}{2}$
Pigovian Tax	$\hat{s} = \frac{1}{2} \implies$ wp $h(\frac{1}{2})^2$: $\hat{I} = \frac{1}{2}$ $h(\frac{1}{2})(1 - h(\frac{1}{2}))$: $\hat{I} = 1$ $(1 - h(\frac{1}{2}))h(\frac{1}{2})$: $\hat{I} = 0$	$\hat{s} \geq 0 \implies$ wp $h(\hat{s})h(1 - \hat{s})$: $\hat{I} = \frac{2\Delta - \delta}{4\Delta - \delta}$ $h(\hat{s})(1 - h(1 - \hat{s}))$: $\hat{I} = \frac{2\Delta - \delta}{3\Delta - \delta}$ $(1 - h(\hat{s}))h(1 - \hat{s})$: $\hat{I} = 0$	$\hat{s} \geq 0 \implies$ wp $h(\hat{s})h(1 - \hat{s})$: $\hat{I} = \frac{2\Delta - \delta}{3\Delta - \delta}$ $h(\hat{s})(1 - h(1 - \hat{s}))$: $\hat{I} = 1$ $(1 - h(\hat{s}))h(1 - \hat{s})$: $\hat{I} = 0$
Ex Post Liability	$\hat{s} = \frac{1}{2} \implies$ wp $h(\frac{1}{2})^2$: $\hat{I} = \frac{1}{2}$ $h(\frac{1}{2})(1 - h(\frac{1}{2}))$: $\hat{I} = 1$ $(1 - h(\frac{1}{2}))h(\frac{1}{2})$: $\hat{I} = 0$	$\hat{s} \geq 0 \implies$ wp $h(\hat{s})h(1 - \hat{s})$: $\hat{I} = \frac{2\Delta - \delta}{4\Delta - (1 + \tilde{\mu})\delta}$ $h(\hat{s})(1 - h(1 - \hat{s}))$: $\hat{I} = \frac{2\Delta - \delta}{3\Delta - \delta + \tilde{\mu}\Delta}$ $(1 - h(\hat{s}))h(1 - \hat{s})$: $\hat{I} = 0$	$\hat{s} \geq 0 \implies$ wp $h(\hat{s})h(1 - \hat{s})$: $\hat{I} = \frac{2\Delta - \delta}{3\Delta - \delta + \mu\Delta}$ $h(\hat{s})(1 - h(1 - \hat{s}))$: $\hat{I} = 1$ $(1 - h(\hat{s}))h(1 - \hat{s})$: $\hat{I} = 0$

Table 1: **Research and Adoption Outcomes**

outcomes are depicted with a 'hat' while policy requirements are hat-free. The endogenous items in blue are those that are socially optimal. In what follows, each outcome is described before turning to consider their rankings in terms of social welfare realised. In the appendix, expected social welfare for each option is derived and is the basis for the results to follow.

5.3 Bans on Adoption and Research

An *adoption ban* involves banning the adoption of any technology architecture for which it is known that $\delta_j = \delta$. If A is known to involve harm, its adoption is prohibited while, for B, prohibition only arises if it has advanced and adopted in period 1, as this is the only case where a signal of the harmfulness of adoption is generated and that signal realisation is that it is harmful.

Recalling that there can be no perfect signal that B is safe given that $1 - I_1 < 1$, B's harm remains uncertain in two cases and all scientific resources are allocated to B because there is no return from researching on A. Thus, with probability $h(1)$, B advances and, in so doing, competes with A at quality $Q_A(2) = \Delta$. Note, however, that there remains a risk that B is harmful, but this is not taken into account in the adoption decision. In the one case where B is known to be harmful, the adoption of both A and B are barred, so there is no research, and total welfare is 0. The social welfare calculations are derived in the appendix.[18]

A *prohibition on research* involves preventing research from advancing any technology architecture for which it is known that $\delta_j = \delta$. If A is known to involve harm, research on A is prohibited while, for B, prohibition could only arise if it has advanced and adopted in period 1, where a prohibition is put in place if the resulting signal is that it is harmful.

In this case, where B is still not known to be harmful, all scientific resources are allocated to B. If B research is successful, then this creates competition for A and adoption moves towards a greater number of sectors using B. That competition is beneficial as any sector adopting A is welfare-reducing as $\Delta < \delta$. When B is known to be harmful, there is no further research. In this case, adopting either technology is welfare reducing but because there is no ban on adoption that occurs in all sectors.

We are now in a position to compare the policies of a ban on adoption and a prohibition on research. We can establish the following.

Proposition 4 *A prohibition of research always results in lower expected social welfare than a ban on adoption.*

The intuition is straightforward and can be seen in Table 1. An adoption ban leads to the same research outcome as a research prohibition but ensures that there is no adoption of A as $Q_A(2) = \Delta$, which would be strictly welfare-reducing. Such adoption can still occur when there is only a research prohibition.

5.4 Pigouvian Taxation

Pigouvian taxation involves levying a sector-specific charge, $\tau_{i,j} = -E_{i,j} = \eta_{i,j}\delta$, on each sector adopting a technology j that is *known* to be harmful. As it is assumed that A is known to be harmful, then a full internalisation of the externality would imply a tax of $\tau_{i,A} = -E_{i,A} = \eta_{i,A}\delta$ for all $i < \hat{I}$. Note that while this tax would eliminate the adoption of A if $Q_A(2) = \Delta$, A will still be adopted if $Q_A(2) = 2\Delta$. This plays an important role in driving the social and private incentives to research in period 2 along the A path.

[18] Note that there is a potential time inconsistency issue. Because research on A is not prohibited, should it occur and should A advanced to $Q_A(2) = 2\Delta$, then it is possible that a policy-maker who has not committed to the ban may have an incentive to reverse the ban on A as $2\Delta > \delta$. If this reversal were anticipated, this would justify the research being conducted on the A architecture. This possibility is not evaluated here as it is assumed that the planner's policy implementations are time-consistent.

Once again, the full social welfare outcomes are stated in the appendix. It is instructive to compare the outcomes under Pigouvian taxation with those from a ban on adoption. It is often the case that Pigouvian taxation that fully internalises the external harm for a decision-maker, here a sector adopting a harmful technology, leads to higher social welfare outcomes than a ban on the decision that may lead to harm. This certainly is the case when B is also known to harmful. In that situation, the Pigouvian tax leaves open the possibility of adoption should there be further advances in one or both architectures and so expected social welfare is higher than zero; the level that results from a complete ban on the adoption of both architectures.

This simple intuition breaks down, however, when there continues to be uncertainty regarding whether the B technology is harmful or not. In this case, while the A technology is only adopted if that technology advances in period 2 (as $\Delta < \delta < 2\Delta$), the B technology might be adopted even if $\Delta < \tilde{\mu}\delta$. In this case, it is possible that in certain states expected social welfare may be negative even under a Pigouvian tax. Nonetheless, it remains the case that expected social welfare at the beginning of period 2 is higher under a Pigouvian tax than a ban on adoption.

Proposition 5 *A Pigouvian tax always results in a (weakly) higher expected social welfare than a ban on adoption.*

The reason that a Pigouvian tax socially dominates a ban on technology adoption even when B harm is uncertain is because the B technology might be adopted even when A has been banned. When $\hat{s} \to 0$ under a Pigouvian tax, expected social welfare is the same as under a ban on adoption. However, it is possible that in some circumstances, $\hat{s} > 0$ under a Pigouvian tax, leading to some A adoption that is superior to B adoption for some sectors if $2\Delta - \delta > \Delta - \tilde{\mu}\delta$.

The potential social inefficiency from a Pigouvian tax arises because B adopters do not internalise the *expected* harm from their adoption. This insight leads to the following result that demonstrates that a Pigouvian tax that fully internalises the harm from A adopters leads to too little A adoption and too few scientists allocated to advancing A further.

Proposition 6 *Under a Pigouvian tax of $\tau_{i,A} = \eta_{i,A}\delta$ for all i adopting A, when the harmfulness of B remains unknown, increasing the allocation of scientists to the A architecture would raise expected social welfare.*

The proof is a straightforward examination of the expected social welfare calculations in the Appendix. Intuitively, when the harm to B remains unknown, its expected harm of $\tilde{\mu}\delta$ or $\mu\delta$ as the case may be, is not taken into account in either the adoption of B by sectors or the returns to B research. If it were taken into account, both would be reduced. Thus, from a social perspective, a Pigouvian tax whereby A adopters internalised fully the known harm would result in too little research being directed towards improving A. One mechanism that could achieve this would be to lower $\tau_{i,A}$.

A natural question following this result is what the optimal Pigouvian tax would be given that it can only be imposed on the adopters of technology known to be harmful. A precise characterisation of this would be complex and really only establish that a second-best outcome could be achieved with a tax that does not fully internalise A's external harm; that is, is lower than $\eta_{i,A}\delta$. The real issue is that the optimal tax needs to be contingent upon the realised level of both architectures, which means it must be determined ex-post after the realisation of research outcomes during period 2. The following result characterises the optimal (ex post) tax on A:

Proposition 7 *Suppose that B is not known to be harmful at the beginning of period 2 and that $h(s_j) = s_j^a$ ($a \in (0,1)$). The optimal Pigouvian taxes following the realisation of period 2 research outcomes are as follows:*

1. *if $Q_A(2) = Q_B(2) = 2\Delta$, $\tau^*_{i,A} = \eta_{i,j} \max\{\frac{\Delta(1-3\tilde{\mu})\delta}{\Delta-\tilde{\mu}}\delta, 0\}$;*
2. *if $Q_A(2) = 2\Delta$ and $Q_B(2) = \Delta$, $\tau^*_{i,A} = \eta_{i,j} \max\{\frac{\Delta(1-2\tilde{\mu})\delta}{\Delta-\tilde{\mu}}\delta, 0\}$; and*
3. *if $Q_A(2) = \Delta$ and/or $Q_B(2) = 0$, $\tau^*_{i,A} = \eta_{i,A}\delta$.*

The proof follows directly from maximising expected social welfare as calculated in the appendix and noting that the assumption on $h(.)$ guarantees that the scientist allocation is optimal as per Lemma 1.

These taxes are optimal because they are such that the adoption of A is optimal ex post (that is, so that $\hat{I} = \frac{Q_A(2)-\delta}{Q_A(2)-\delta+Q_B(2)-\tilde{\mu}\delta}$). Note that they may be contingent on the posterior probability, $\tilde{\mu}$, which itself depends on whether any B adoption occurred in period 1. The higher is $Q_B(2)$, the lower is $\tau^*_{i,A}$. This is because it is when B is more competitive that, at the margin, A's competitiveness needs to be strengthened the most.

Interestingly, the proposition demonstrates that there are situations when it may be optimal not to have a tax at all. That is, if either $2\Delta - \delta \geq \Delta - \tilde{\mu}\delta \Leftrightarrow \tilde{\mu} \geq \frac{\delta-\Delta}{\delta}$ or $\tilde{\mu} \geq \frac{1}{3}$ depending on the case. In these situations, the optimal policy would be a subsidy to A rather than a tax. This reflects the observation made earlier that if μ is sufficiently high, then the socially optimal optionality against harm from a technology is to develop a technology sufficiently advanced to 'pay' for that harm rather than prevent the adoption of harmful technology per se.

5.5 Ex post liability

When the Pigouvian tax can be applied ex post and tailored to the realisation of research outcomes, the socially optimal outcome can be produced (at least when $h(s_j) = s_j^a$). There may be practical difficulties in doing this if the precise research outcomes cannot be observed easily by the social planner. However, the final policy option of ex post liability is designed to allow for some degree of tailoring to realised outcomes and, therefore, may result in higher expected social welfare than Pigouvian taxation.

Ex post liability involves imposing a penalty on technology adopters if there is adoption that results in realised harm. While in the model presented here, this possibility does not change research and adoption decisions in period 1 (that is, a penalty may be imposed, but its expectation does not alter the relative allocation of scientific resources nor the adoption decisions), it does impact on period 2 decisions both in terms of realised outcomes prior to that point (akin to the impact of a Pigouvian tax) but also in expectation of potential harm (unlike a Pigouvian tax). Therefore, as is summarised in Table 1, in numerous scenarios, the expectation of a penalty ex post generates a socially optimal adoption decision.

What restricts a socially optimal adoption decision is that it is assumed here that adopters have limited liability in that their realised penalty cannot exceed the realised surplus when a technology is adopted. Note that surplus is the liability metric because profits accrue to both adopters of a technology and also providers of the technology (i.e., scientists). It is assumed that both are liable for any realised harm.[19] For instance, if technology adoption results in surplus of $\eta_{i,j}2\Delta$, this can always fund a penalty of $\eta_{i,j}\delta$. Hence, when both technologies generate this surplus, the limited liability constraint is not binding, and adoption is socially optimal.

However, if the adoption of a technology results in a surplus of $\eta_{i,j}\Delta$, this also defines the maximum penalty, which is less than $\eta_{i,j}\delta$. The limited liability constraint binds, and thus, there may be too much adoption prior to the resolution of uncertainty regarding harm. The only time, however, when this leads to sub-optimal adoption is when adopting technology A generates a surplus of $\eta_{i,A}2\Delta$ while B generates a surplus of $\eta_{i,B}\Delta$. In this case, the limited liability constraint applies for B but not for A, resulting in too little adoption of A feeding into too few scientific resources devoted to advancing A. If liability were unlimited, this distortion would not arise, and adoption would be socially optimal.

Nonetheless, even limited liability pushes adoption closer to a socially optimal level than does Pigouvian taxation. For that reason, the following result can be demonstrated:

Proposition 8 *An ex post liability regime always results in (weakly) higher expected social*

[19] This does not necessarily reflect how tort law might be applied, which may only find one of these agent types liable, in which case the liability constraint will bind more strongly.

welfare than a Pigouvian tax.

Thus, this completes the comparison of each of the four policy options. Put simply, the more a policy instrument can adjust in its application to the realisation of uncertainty along both the harm and research outcome dimensions, the closer the outcome will be to what would be socially optimal. This favours policy options that are ex-post in nature and only determined in their application following the realisation of uncertainty but where, in anticipation of that adjustment, it impacts the expectations of the relevant decision-makers – scientists and technology adopters.[20]

6 Extensions

Various assumptions were made in deriving the above results. In this section, these assumptions are re-examined, and the implications of generalising them are considered.

6.1 Learning from Research

The model thus far assumes that signals regarding harm arise from AI adoption. That is, harm levels are surfaced through direct marketplace testing or 'learning by doing.' Harms can also be signalled through research or what Gans (2024) calls 'lab learning.' This is the type of learning considered by Acemoglu and Lensman (2024). The analogue to the earlier updating formula would be that if there is no signal of harm following a research period, the posterior probability of harm becomes:

$$\tilde{\mu}_A = \frac{\mu(1-s)}{\mu(1-s)+(1-\mu)}$$

$$\tilde{\mu}_B = \frac{\mu s}{\mu s+(1-\mu)}$$

Thus, if all research is devoted to one path, the signal of harm will be perfect; otherwise, there is some learning, but uncertainty remains.

From a policy perspective, the post-research probabilities of harm can be taken into account when deciding whether to adopt technologies should they have been developed. As harm only arises from adoption, this generates an incentive to conduct research in order

[20] Guerreiro et al. (2023) examines regulatory options with respect to AI and finds that a liability regime is socially optimal in their context if liability is unlimited. They examine different margins of AI adoption than the focus on the direction of technological change here. It should also be noted that even imposing ex post liability is a relatively optimistic practical assumption. This is one reason why bans are easier to impose. The modelling of such constraints is left for future researchers.

to evaluate harm and potentially avoid incurring any harm. This learning and the option it affords are valuable for regulatory interventions that are contingent upon those signals, such as banning adoption (or further research) and a Pigouvian tax. For ex post liability, however, the new information can be used in decentralised scientist and adoption decisions and will reduce errors.

Apart from these details, however, having lab learning rather than learning by doing changes the overall regulatory picture along the lines outlined by Gans (2024) in that it potentially introduces a precautionary motive to any AI adoption and would prioritise research to surface harms should that be possible.

6.2 Higher and Lower Damage

If harm occurs, it has been assumed that $\delta \in [\Delta, 2\Delta)$. Within this range, it is socially optimal to adopt technology along a path if it is sufficiently advanced (i.e., $q_j = 2$) and not otherwise. It is this assumption that generated a potential incentive to continue research along a path even if the risk of harm had increased over time.

There are two changes that may arise if this assumption is relaxed. If $\delta < \Delta$, then it is never optimal not to adopt a technology, even if it is known to be harmful. Thus, there is no case for intervention to prohibit adoption, although there remains an incentive to push adoption towards a less harmful path should it exist. Otherwise, the incentives to choose regulatory instruments that internalise externalities and risk remain.

If $\delta > 2\Delta$, then it is never optimal to adopt a technology known to be harmful. This strengthens the case for banning adoption and/or research along a harmful path, as it is no longer the case that advancing the technology sufficiently can outweigh the costs it imposes. Thus, this paper could be interpreted as favouring policies that allow the pricing-in of externalities into private decisions so long as the harm evaluated is not too high. If that harm is known to be substantial, this bolsters the case for prohibitions as regulatory instruments.

6.3 Damage that Scales

The final assumption worth examining is that the extent of damage is independent of the quality of the technology adoption. By contrast, suppose that $E_{i,j}(t) = \eta_{i,j}\delta_j Q_j(t)$; that is, damage scales with the quality of the technology. This is the main case considered by

Acemoglu and Lensman (2024).[21] This implies that total social welfare at a given time is:

$$\int_0^I \eta_{i,A}(1-\delta_A)Q_A(2)di + \int_I^1 \eta_{i,B}(1-\delta_B)Q_B(2)di$$

The optimum sectoral adoption threshold satisfies:

$$I^*(t) = \frac{Q_A(t)(1-\delta_A)}{Q_A(t)(1-\delta_A) + Q_B(t)(1-\delta_B)}$$

Importantly, this implies that it is optimal to adopt a technology with $Q_j(t) > 0$ regardless of its quality if and only if $\delta_A < 1$.

It can readily be seen that this specification simplifies the analysis of the model akin to the cases of higher ($\delta_j > 1$) and lower ($\delta_j < 1$) damage considered in the previous subsection. However, it does not allow the more complex trade-offs that arise in the intermediate case, which is the focus of this paper.

However, this specification could open up various policy commitment issues that may be pursued in future work. For instance, for δ always less than 1, a policy-maker may want to commit to not adopting a technology with known harm so as to encourage scientific research on the other technology path. However, when policymakers learn about the level of harm, they may be unable to commit to de-adopting the technology. Thus, there may be a time inconsistency issue that, in turn, makes certain decisions irreversible. As Gans (2024) argues, irreversibility can change the value of learning about harm. Examining this would require a more detailed model of the regulator than is provided here, and so it is left for future work.

7 Conclusion

This paper has examined the complex trade-offs involved in regulating the direction of technological innovation, with a particular focus on situations where potential harms from new technologies are uncertain. The analysis reveals several key insights. First, the socially optimal allocation of scientific resources between competing technological paths is not always straightforward. While diversification can provide option value in the face of potential harm, there are also cases where concentrating resources on advancing a leading technology may be preferable, even if that technology carries some risk of harm. Second, market forces alone tend to result in an inefficiently high concentration of research effort on leading technological paths. This stems from scientists failing to account for the negative externality their choice imposes on the probability of success for the lagging path. Third, when regulating in the

[21] The case without such scaling is considered in the main model here is in an appendix in their work.

face of uncertainty about potential harms, ex post policy instruments that can adjust to realized outcomes tend to outperform ex ante prohibitions or restrictions. Specifically, the analysis suggests that ex post liability regimes are likely to produce better outcomes than Pigouvian taxes, which in turn outperform bans on adoption or research. Fourth, the optimal regulatory approach depends critically on the magnitude of potential harm relative to the benefits of technological progress. For very high levels of harm, prohibitions may become optimal, while for lower levels, instruments that allow for pricing-in of risk are preferable. Finally, there is an important distinction between learning about harms through research versus through adoption. The possibility of"lab learning" introduces additional complexity to the optimal regulatory strategy.

These findings have important implications for current debates surrounding the regulation of emerging technologies like AI. They suggest that policymakers should be cautious about implementing heavy-handed ex ante restrictions on research or adoption paths. Instead, the focus should be on developing robust mechanisms for ongoing assessment of potential harms and flexible policy instruments that can adjust as uncertainty is resolved.

However, several important questions remain for future research. These include exploring how different liability regimes might be structured to optimize incentives, examining the implications of strategic behaviour by firms or researchers in anticipation of future regulation, and investigating how international coordination (or lack thereof) impacts the efficacy of different regulatory approaches.

8 Appendix

8.1 Proof of Proposition 1

The planner chooses s to maximise:

$$\begin{aligned}\mathbb{E}[V^*(1,1,\delta_A,\delta)] =& h(s)h(1-s)\mathbb{E}[v_S(2,2,\delta_A,\delta)] \\ &+ (1-h(s))h(1-s)\mathbb{E}[v_S(1,2,\delta_A,\delta)] \\ &+ h(s)(1-h(1-s))\mathbb{E}[v_S(2,0,\delta_A)] \\ &+ (1-h(s))(1-h(1-s))\mathbb{E}[v_S(1,0,\delta_A,0)]\end{aligned}$$

where the expectations on the RHS are with respect to δ_A based on $\tilde{\mu}$ and $\mathbb{E}[v_S(2,1,\delta_A,\delta)] = \mathbb{E}[v_S(2,0,\delta_A,0)]$ and $\mathbb{E}[v_S(1,1,\delta_A,\delta)] = \mathbb{E}[v_S(1,0,\delta_A,0)]$ as $I^*(2) = 1$ in those cases. Let $\Omega(.)$ is the degree of substitutability between A and B:

$$\Omega(1,1) \equiv \mathbb{E}[v_S(1,2,\delta_A,\delta)] + \mathbb{E}[v_S(2,0,\delta_A,0)] - \mathbb{E}[v_S(2,2,\delta_A,\delta)] - \mathbb{E}[v_S(1,0,\delta_A,0)]$$

The derivative of the planner's objective with respect to s is:

$$h'(1-s)\Big(\Omega(1,1)h(s) + \mathbb{E}[v_S(1,0,\delta_A,0)] - \mathbb{E}[v_S(1,2,\delta_A,\delta)]\Big)$$

$$-h'(s)\Big(\Omega(1,1)h(1-s) + \mathbb{E}[v_S(1,0,\delta_A,0)] - \mathbb{E}[v_S(2,0),\delta_A,0)]\Big)$$

At $s = \frac{1}{2}$, the first term exceeds the second if:

$$\mathbb{E}[v_S(2,0,\delta_A,0)] \geq \mathbb{E}[v_S(1,2,\delta_A,\delta)]$$

$$\implies \tfrac{1}{2}(2\Delta - \tfrac{\mu}{2-\mu}\delta) \geq \tfrac{\delta^2((\mu-2)\mu+4)-\delta\Delta(\mu-10)(\mu-2)+7\Delta^2(\mu-2)^2}{2(\mu-2)(2\delta+3\Delta(\mu-2))} \implies \tfrac{(\delta\Delta(6-4\mu)-\Delta^2-\delta^2)(2-\mu)}{6\Delta(2-\mu)-4\delta} \geq 0$$

which always holds strictly as $\mu < \frac{1}{2}$. Hence, given the concavity of $h(.)$, $s^*(2) > \frac{1}{2}$.

8.2 Proof of Proposition 2

First, assume that $\mathbb{E}[\delta_A|(1,0),\varnothing]=0$. The planner chooses s to maximise:

$$\begin{aligned}\mathbb{E}[V^*(1,0,0,\delta_B)] =& h(s)h(1-s)\mathbb{E}[v_S(2,1,0,\delta_B)]\\ &+(1-h(s))h(1-s)\mathbb{E}[v_S(1,1,0,\delta_B)]\\ &+h(s)(1-h(1-s))\mathbb{E}[v_S(2,0,0,\delta_B)]\\ &+(1-h(1-s))(1-h(s))\mathbb{E}[v_S(1,0,0,\delta_B)]\end{aligned}$$

where the expectations on the RHS are with respect to δ_B based on μ and:

$$\Omega(1,0)\equiv\mathbb{E}[v_S(1,1,0,\delta_B)]+\mathbb{E}[v_S(2,0,0,\delta_B)]-\mathbb{E}[v_S(2,1,0,\delta_B)]-\mathbb{E}[v_S(1,0,0,\delta_B)]$$

The derivative of the planner's objective with respect to s is:

$$h'(1-s)\Big(\Omega(1,0)h(s)+\mathbb{E}[v_S(1,0,0,\delta_B)]-\mathbb{E}[v_S(1,1,0,\delta_B)]\Big)$$

$$-h'(s)\Big(\Omega(1,0)h(1-s)+\mathbb{E}[v_S(1,0,0,\delta_B)]-\mathbb{E}[v_S(2,0),0,\delta_B)]\Big)$$

Following the same procedure as the proof of Proposition 1, $s^*(2)>\frac{1}{2}$ if and only if:

$$\mathbb{E}[v_S(2,0),0,\delta_B)]>\mathbb{E}[v_S(1,1,0,\delta_B)]\implies\Delta>\tfrac{-\delta^2\mu^2+\delta\Delta\mu+\Delta^2}{2(2\Delta-\delta\mu)}$$

which always holds.

Second, assume that $\mathbb{E}[\delta_A|(1,0),\delta]=\delta$. The planner chooses s to maximise:

$$\begin{aligned}\mathbb{E}[V^*(1,0,\delta,\delta_B)] =& h(s)h(1-s)\mathbb{E}[v_S(2,1,\delta,\delta_B)]\\ &+(1-h(s))h(1-s)\mathbb{E}[v_S(1,1,\delta,\delta_B)]\\ &+h(s)(1-h(1-s))\mathbb{E}[v_S(2,0,\delta,\delta_B)]\\ &+(1-h(1-s))(1-h(s))\mathbb{E}[v_S(1,0,\delta,\delta_B)]\end{aligned}$$

where the expectations on the RHS are with respect to δ_B based on μ and:

$$\Omega(1,0)\equiv\mathbb{E}[v_S(1,1,\delta,\delta_B)]+\mathbb{E}[v_S(2,0,\delta,\delta_B)]-\mathbb{E}[v_S(2,1,\delta,\delta_B)]-\mathbb{E}[v_S(1,0,\delta,\delta_B)]$$

The derivative of the planner's objective with respect to s is:

$$h'(1-s)\Big(\Omega(1,0)h(s)+\mathbb{E}[v_S(1,0,\delta,\delta_B)]-\mathbb{E}[v_S(1,1,\delta,\delta_B)]\Big)$$

$$-h'(s)\Big(\Omega(1,0)h(1-s) + \mathbb{E}[v_S(1,0,\delta,\delta_B)] - \mathbb{E}[v_S(2,0),\delta,\delta_B)]\Big)$$

Following the same procedure as the proof of Proposition 1, $s^*(2) > \frac{1}{2}$ if and only if:

$$\mathbb{E}[v_S(2,0),\delta,\delta_B)] > \mathbb{E}[v_S(1,1,\delta,\delta_B)] \implies \Delta - \tfrac{1}{2}\delta > \tfrac{1}{2}(\Delta - \mu\delta) \implies \tfrac{1}{2}(\Delta - (1-\mu)\delta) > 0$$

which holds if $\mu > \frac{\delta - \Delta}{\delta}$. Note that the RHS is less than $\frac{1}{2}$ so it could hold when if $\mu < \frac{1}{2}$.

8.3 Proof of Lemma 1

Suppose $s^* = \frac{1}{2}$, then $h'(s^*) = h'(1-s^*)$ which implies that $\frac{h(\hat{s})/\hat{s}}{h(1-\hat{s})/(1-\hat{s})}$ or $\hat{s} = \frac{1}{2}$. If $q_A = q_B = q$, then $\frac{h'(s^*(q,q))}{h'(1-s^*(q,q))} = \frac{h(\hat{s}(q,q))/\hat{s}(q,q}{h(1-\hat{s}(q,q))/(1-\hat{s}(q,q))} = 1$. In this case, $s^* = \frac{1}{2}$.

Next, note that $\lim_{s\to 0}\left(\frac{h'(s)}{h'(1-s)} - \frac{h(s)/s}{h(1-s)/(1-s)}\right) > 0$ and $\lim_{s\to 1}\left(\frac{h'(s)}{h'(1-s)} - \frac{h(s)/s}{h(1-s)/(1-s)}\right) < 0$ by the assumed Inada conditions on $h(.)$. If $\frac{h'(s)}{h(s)/s}$ is decreasing in s, $\frac{h'(s)}{h'(1-s)}$ and $\frac{h(s)/s}{h(1-s)/(1-s)}$ cross at exactly one point, which, as already demonstrated, is where $s = \frac{1}{2}$.

Note that $q_A \neq q_B$ if $\{q_A, q_B\} = \{1,0\}$ or $\{0,1\}$. Consider the case where $\{q_A, q_B\} = \{1,0\}$. Note that $\frac{1-\frac{1}{3}h(s)}{2-\frac{1}{3}h(1-s)}$ has the following properties: (i) it is decreasing in s; (ii) as $s \to 0$, this becomes $\frac{1}{2-\frac{1}{3}h(1)} < 1$; (iii) as $s \to 1$, this becomes $\frac{1-\frac{1}{3}h(1)}{2} < \frac{1}{2-\frac{1}{3}h(1)}$; and (iv) that $\frac{\partial \frac{h'(s)}{h'(1-s)}}{\partial s} < \frac{\partial \frac{1-\frac{1}{3}h(s)}{2-\frac{1}{3}h(1-s)}}{\partial s}$, as $\frac{h'(1)}{h'(0)} = 0 < \frac{1-\frac{1}{3}h(1)}{2}$. This implies that $\frac{1-\frac{1}{3}h(s^*(1,0))}{2-\frac{1}{3}h(1-s^*(1,0))} < 1$ and so $s^* > \frac{1}{2}$ and $\frac{h(s^*)/s^*}{h(1-s^*)/(1-s^*)} > \frac{1-\frac{1}{3}h(s^*(1,0))}{2-\frac{1}{3}h(1-s^*(1,0))}$. Condition (iv) above then implies that $\hat{s}(1,0) > s^*(1,0)$. An analogous argument holds where $s^* < \frac{1}{2}$.

8.4 Proof of Proposition 3

Let τ_A and τ_B be the expected tax of each path. Consider the following cases and derivation of relevant provider profits based on the realisation of research outcomes in $t = 2$:

1. $Q_A(2) = Q_B(2) = 2\Delta$. Then $\hat{I}(2,2) = \frac{2\Delta - \tau_A}{4\Delta - \tau_A - \tau_B}$.

    - If $\Delta < \tau_A$, then no in-architecture competition, so A profits are $\int_0^{\hat{I}(2,2)}((1-2i)2\Delta + i\tau_B - (1-i)\tau_A)di = \frac{(2\Delta - \tau_A)^2}{8\Delta - 2\tau_A - 2\tau_B}$.
    - If $\Delta > \tau_A$, there is in-architecture competition with $\hat{I}_A(2,2) = \{i|(1-i)(\Delta - \tau_A) = i(2\Delta - \tau_B)\} = \frac{\Delta - \tau_A}{3\Delta - \tau_A - \tau_B}$, so A profits are $\int_0^{\hat{I}_A(2,2)}(1-i)(2\Delta - \tau_A - (\Delta - \tau_A))di + \int_{\hat{I}_A(2,2)}^{\hat{I}(2,2)}((1-2i)2\Delta + i\tau_B - (1-i)\tau_A)di = \frac{\Delta\big(\tau_A(2\tau_B - 7\Delta) + \tau_A^2 + \Delta(8\Delta - 3\tau_B)\big)}{2(4\Delta - \tau_A - \tau_B -)(3\Delta - \tau_A - \tau_B -)}$.

2. $Q_A(2) = 2\Delta$ and $Q_B(2) = \Delta$.

- If $\Delta < \tau_B$, $\hat{I}(2,1) = 1$ and there is no inter-architecture competition:
  - if $\Delta > \tau_A$, profits are $\int_0^1 (1-i)\Delta di = \frac{1}{2}\Delta$.
  - if $\Delta < \tau_A$, $\int_0^1 (1-i)(2\Delta - \tau_A)di = \frac{1}{2}(2\Delta - \tau_A)$.
- If $\Delta \geq \tau_B$, $\hat{I}(2,1) = \frac{2\Delta - \tau_A}{3\Delta - \tau_A - \tau_B}$, and there is inter-architecture competition:
  - if $\Delta > \tau_A$, $\hat{I}_A(2,1) = \frac{\Delta - \tau_A}{2\Delta - \tau_A - \tau_B}$ and A profits are $\int_0^{\hat{I}_A(2,1)} (1-i)\Delta di + \int_{\hat{I}_A(2,1)}^{\hat{I}(2,1)} ((2-3i)\Delta + i\tau_B - (1-i)\tau_A)di = \frac{\Delta\left(\tau_A(2\tau_B - 5\Delta) + \tau_A^2 + \Delta(5\Delta - 3\tau_B)\right)}{2(3\Delta - \tau_A - \tau_B)(2\Delta - \tau_A - \tau_B)}$.
  - If $\Delta < \tau_A$, A profits are $\int_0^{\hat{I}(2,1)} ((2-3i)\Delta + i\tau_B - (1-i)\tau_A)di = \frac{(\tau_A - 2\Delta)^2(-\tau_A - 2\tau_B + 3\Delta)}{2(3\Delta - \tau_A - \tau_B)^2}$.

3. $Q_A(2) = 2\Delta$ and $Q_B(2) = 0$.
   - if $\Delta > \tau_A$, A profits are $\int_0^1 (1-i)\Delta di = \frac{1}{2}\Delta$.
   - if $\Delta < \tau_A$, A profits are $\int_0^1 (1-i)(2\Delta - \tau_A)di = \frac{1}{2}(2\Delta - \tau_A)$.
4. $Q_A(2) = Q_B(2) = \Delta$.
   - If $\Delta > \tau_B$, and $\Delta > \tau_A$, $\hat{I}(1,1) = \frac{\Delta - \tau_A}{2\Delta - \tau_A - \tau_B}$, B profits are $\int_{\hat{I}(1,1)}^1 (i(\Delta - \tau_B) - (1-i)(\Delta - \tau_A))di = \frac{(\Delta - \tau_B)^2}{4\Delta - 2\tau_A - 2\tau_B}$.
   - If $\Delta < \tau_A$, B profits are $\int_0^1 i(\Delta - \tau_B)di = \frac{1}{2}(\Delta - \tau_B)$.
5. $Q_A(2) = \Delta$ and $Q_B(2) = 0$. B's profits are 0.

Now, we derivate the resulting social welfare outcomes for each of these cases:

1. $Q_A(2) = Q_B(2) = 2\Delta$ then $\hat{I}(2,2) = \frac{2\Delta - \tau_A}{4\Delta - \tau_A - \tau_B}$. $v_S(2,2) = \int_0^{\hat{I}(2,2)} (1-i)(2\Delta - \tau_A)di + \int_{\hat{I}(2,2)}^1 i(2\Delta - \tau_B)di = \frac{(\tau_A - 2\Delta)^2}{8\Delta - 2\tau_A - 2\tau_B} - \frac{\tau_B}{2} + \Delta$.
2. $Q_A(2) = 2\Delta$ and $Q_B(2) = \Delta$.
   - If $\Delta < \tau_B$, $v_S(2,1) = v_S(2,0) = \int_0^1 (1-i)(2\Delta - \tau_A)di = \frac{1}{2}(2\Delta - \tau_A)$.
   - If $\Delta > \tau_B$, $v_S(2,1) = \int_0^{\hat{I}(2,1)} (1-i)(2\Delta - \tau_A)di + \int_{\hat{I}(2,1)}^1 i(\Delta - \tau_B)di = \frac{1}{2}\left(\frac{(2\Delta - \tau_A)^2}{3\Delta - \tau_A - \tau_B} - \tau_B + \Delta\right)$.
3. $Q_A(2) = 2\Delta$ and $Q_B(2) = 0$. $v_S(2,0) = \int_0^1 (1-i)(2\Delta - \tau_A)di = \frac{1}{2}(2\Delta - \tau_A)$.
4. $Q_A(2) = Q_B(2) = \Delta$.
   - $\Delta > \tau_A$ and $\Delta > \tau_B$, $\hat{I}(1,1) = \frac{\Delta - \tau_A}{2\Delta - \tau_A - \tau_B}$ and $v_S(1,1) = \int_0^{\hat{I}(1,1)} (1-i)(\Delta - \tau_A)di + \int_{\hat{I}(1,1)}^1 i(\Delta - \tau_B)di = \frac{1}{2}\left(\frac{(\Delta - \tau_A)^2}{2\Delta - \tau_A - \tau_B} - \tau_B + \Delta\right)$
   - $\Delta > \tau_A$ and $\Delta < \tau_B$, $v_S(1,0) = \frac{1}{2}(\Delta - \tau_A)$

- $\Delta < \tau_A$ and $\Delta > \tau_B$, $v_S(0,1) = \frac{1}{2}(\Delta - \tau_B)$
- $\Delta < \tau_A$ and $\Delta < \tau_B$, $v_S(1,1) = 0$

First, suppose that $(q_A, q_B) = (1,1)$. Note that the first order condition for the socially optimal scientist allocation is:

$$\frac{h'(s)}{h'(1-s)} = \frac{(v_S(1,2) + v_S(2,1) - v_S(1,1) - v_S(2,2))h(s) + v_S(1,1) - v_S(1,2)}{(v_S(1,2) + v_S(2,1) - v_S(1,1) - v_S(2,2))h(1-s) + v_S(1,1) - v_S(2,1)}$$

And the condition defining the scientist allocation in the decentralised equilibrium (contingent on τ_A and τ_B) is:

$$\frac{h(s)/s}{h(1-s)/(1-s)} = \frac{h(s)v_B(2,2) + (1-h(s))v_B(1,2)}{h(1-s)v_A(2,2) + (1-h(1-s))v_A(2,1)}$$

For each case, we can compare the two conditions and observe that the RHS of each is identical for each case:

1. $\Delta > \tau_A$ and $\Delta > \tau_B$:

$$\frac{h(s)/s}{h(1-s)/(1-s)} = \frac{\Delta h(s)(\tau_A(2\tau_B-3\Delta)+\Delta(4\Delta-3\tau_B))-(-\tau_A-\tau_B+4\Delta)\left(\tau_A(2\tau_B-3\Delta)-5\Delta\tau_B+\tau_B^2+5\Delta^2\right)}{\Delta h(1-s)(\tau_A(2\tau_B-3\Delta)+\Delta(4\Delta-3\tau_B))-(-\tau_A-\tau_B+4\Delta)\left(\tau_A(2\tau_B-5\Delta)+\tau_A^2+\Delta(5\Delta-3\tau_B)\right)} = \frac{h'(s)}{h'(1-s)}$$

2. $\Delta > \tau_A$ and $\Delta < \tau_B$ (and $\Delta < \tau_A$ and $\Delta > \tau_B$):

$$\frac{h(s)/s}{h(1-s)/(1-s)} = \frac{\Delta\left((\tau_A+\tau_B)(-\tau_A-\tau_B+7\Delta)+h(s)(\tau_A-2\Delta)^2-12\Delta^2\right)}{(\tau_A-2\Delta)^2(\tau_A+\tau_B-4\Delta+\Delta h(1-s))} = \frac{h'(s)}{h'(1-s)}$$

3. $\Delta < \tau_A$ and $\Delta < \tau_B$:

$$\frac{h(s)/s}{h(1-s)/(1-s)} = \frac{(2\Delta-\tau_B)(h(s)(2\Delta-\tau_A)+\tau_A+\tau_B-4\Delta)}{(2\Delta-\tau_A)(\tau_A+h(1-s)(2\Delta-\tau_B)+\tau_B-4\Delta)} = \frac{h'(s)}{h'(1-s)}$$

Second, suppose that $(q_A, q_B) = (1,0)$. The first order condition for the socially optimal scientist allocation is:

$$\frac{h'(s)}{h'(1-s)} = \frac{(v_S(1,1) + v_S(2,0) - v_S(1,0) - v_S(2,1))h(s) + v_S(1,0) - v_S(1,1)}{(v_S(1,1) + v_S(2,0) - v_S(1,0) - v_S(2,1))h(1-s) + v_S(1,0) - v_S(2,0)}$$

And the condition defining the scientist allocation in the decentralised equilibrium (contingent on τ_A and τ_B) is:

$$\frac{h(s)/s}{h(1-s)/(1-s)} = \frac{h(s)v_B(2,1) + (1-h(s))v_B(1,1)}{h(1-s)v_A(2,1) + (1-h(1-s))v_A(2,0)}$$

For each case, we can compare the two conditions and observe that the RHS of each is identical for each case:

1. $\Delta > \tau_A$ and $\Delta > \tau_B$:

$$\frac{h(s)/s}{h(1-s)/(1-s)} = \frac{(\Delta-\tau_B)^2(\tau_A+\tau_B-3\Delta+\Delta h(s))}{\Delta((\tau_A+\tau_B)(5\Delta-\tau_A-\tau_B)+h(1-s)(\Delta-\tau_B)^2-6\Delta^2)} = \frac{h'(s)}{h'(1-s)}$$

2. $\Delta < \tau_A$ and $\Delta > \tau_B$:

$$\frac{h(s)/s}{h(1-s)/(1-s)} = \frac{(\Delta-\tau_B)(h(s)(2\Delta-\tau_A)+\tau_A+\tau_B-3\Delta)}{(2\Delta-\tau_A)(\tau_A+h(1-s)(\Delta-\tau_B)+\tau_B-3\Delta)} = \frac{h'(s)}{h'(1-s)}$$

Noting that, under our functional form assumptions:

$$\frac{h(s)/s}{h(1-s)/(1-s)} = \frac{h'(s)}{h'(1-s)} = \left(\frac{s}{1-s}\right)^{a-1}$$

completes the proof.

8.5 Social Welfare under Each Policy Option

Here, for completeness, social welfare under each policy option is calculated reflecting the main comparator outcomes in Table 1. The results from Propositions 4 - 7 follow from these calculations as outlined in the text.

8.5.1 Ban on Adoption

1. (B harmful) $q_B = 1$ and $\delta_A = \delta_B = \delta$: No research would be undertaken in period 2 (or more specifically, any research would be inconsequential), and social welfare would be $V^{NoAd}(1, 1, \delta, \delta) = 0$.

2. (B harm uncertain) $q_B = 1$ and $\delta_A = \delta$ and δ_B remains uncertain with posterior probability of $\tilde{\mu}_B = \frac{\mu}{2-\mu}$ as $\hat{I}_1 = \frac{1}{2}$. As the adoption of A is prohibited, then there will be no research devoted to extending it at $t = 2$ as no scientist can earn a return on any advance. Therefore, s_B will be as high as possible, resulting in $\hat{s} = 0$. Expected social welfare will, therefore, be as $\hat{I}_2 = 0$:

$$\begin{aligned}\mathbb{E}[V^{NoAd}(1, 1, \delta, \delta_B)] &= h(1)\int_0^1 i(2\Delta - \tilde{\mu}_B\delta)\,di + (1 - h(1))\int_0^1 i(\Delta - \tilde{\mu}_B\delta)\,di \\ &= \tfrac{1}{2}\left((1 + h(1))\Delta - \delta\tilde{\mu}_B\right)\end{aligned}$$

3. (B has not advanced) $q_B = 0$ and $\delta_A = \delta$: Again this implies that $s_A = 0$ and so $s_B = 1$. Thus, $\mathbb{E}[V^{NoAd}(1, 0, \delta, \delta_B] = h(1)\left(\mu\frac{1}{2}(\Delta - \delta) + (1-\mu)\frac{1}{2}\Delta\right) = h(1)\frac{1}{2}(\Delta - \mu\delta)$.

8.5.2 Prohibition on Research

1. (B harmful) $q_B = 1$ and $\delta_A = \delta_B = \delta$: No research would be undertaken in period 2; however, because both architectures have advanced, $\hat{I}_2 = \frac{1}{2}$ and so social welfare will be $V^{NoRes}(1, 1, \delta, \delta) = \frac{1}{2}(\Delta - \delta) < 0$.

2. (B harm uncertain) $q_B = 1$ and $\delta_A = \delta$ and δ_B remains uncertain with a posterior probability of $\tilde{\mu}_B$: The prohibition on A research implies that all scientists will be allocated to research on B. Thus, $s_B = 1$ (or equivalently), $\hat{s} = 0$. Expected social welfare will, therefore, be:

$$\begin{aligned}\mathbb{E}[V^{NoRes}(1, 1, \delta, \delta_B)] &= (1 - h(1))\left(\int_0^{\frac{1}{2}}(1-i)(\Delta - \delta)\,di + \int_{\frac{1}{2}}^1 i(\Delta - \tilde{\mu}_B\delta)\,di\right)\\ &+ h(1)\left(\int_0^{\frac{1}{3}}(1-i)(\Delta - \delta)\,di + \int_{\frac{1}{3}}^1 i(2\Delta\tilde{\mu}_B - \delta)\,di\right)\\ &= \frac{1}{72}(54\Delta - 27(1 + \tilde{\mu}_B)\delta + h(1)(\delta(7 - 5\tilde{\mu}_B) + 30\Delta))\end{aligned}$$

3. (B has not advanced) $q_B = 0$ and $\delta_A = \delta$: If research on A is prohibited, then all scientists will research on B. Note, however, as the adoption of A is not prohibited, and externalities are not internalised, then $\hat{I}(2) = \frac{1}{2}$ if there is an advance in B; otherwise, all sectors continue to adopt A. It is clear that this involves lower social welfare than prohibiting the adoption of A as any use of A lowers social welfare given that $\Delta < \delta$. Social welfare is:

$$\begin{aligned}\mathbb{E}[V^{NoRes}(1, 0, \delta, \delta_B)] &= (1 - h(1))\int_0^1 (1-i)(\Delta - \delta)\,di\\ &+ h(1)\left(\int_0^{\frac{1}{2}}(1-i)(\Delta - \delta)\,di + \int_{\frac{1}{2}}^1 i(\Delta - \mu\delta)\,di\right)\\ &= \tfrac{1}{2}(\Delta - \delta) + h(1)\frac{2\Delta + (1 - 3\mu)\delta}{8}\end{aligned}$$

8.5.3 Pigouvian Tax

1. (B harmful) $q_B = 1$ and $\delta_B = \delta$: In this case, a tax of $\tau_B = E_{i,B}(2) = -\eta_{i,B}\delta$ is imposed on the adoption of B leading to an ex post optimal adoption of technologies

with $\hat{I}(2)=\frac{1}{2}$ if both A and B advance to $Q_j(2)=2\Delta$, $\hat{I}(2)=1$ if only A advances, $\hat{I}(2)=0$ if only B advances and no adoption if neither advance. Given this, the decentralised allocation of scientists will be $\hat{s}=\frac{1}{2}$, which is also the socially optimal allocation. In this case, social welfare is:

$$\begin{aligned}\mathbb{E}[V^{Tax}(1,1,\delta,\delta)] =&h(\tfrac{1}{2})^2\left(\int_0^{\frac{1}{2}}(1-i)(2\Delta-\delta)\,di+\int_{\frac{1}{2}}^1 i(2\Delta-\delta)\,di\right)\\ &+h(\tfrac{1}{2})(1-h(\tfrac{1}{2}))\left(\int_0^1(1-i)(2\Delta-\delta)di+\int_0^1 i(2\Delta-\delta)di\right)\\ =&h(\tfrac{1}{2})(4-h(\tfrac{1}{2}))\tfrac{1}{4}(2\Delta-\delta)\end{aligned}$$

2. (B harm uncertain) $q_B=1$ and δ_B remains uncertain with posterior probability of $\tilde{\mu}$: Social welfare is:

$$\begin{aligned}\mathbb{E}[V^{Tax}(1,1,\delta,\delta)] =&h(1-s)h(s)\left(\int_{\frac{2\Delta-\delta}{4\Delta-\delta}}^1 i(2\Delta-\delta\tilde{\mu}_B)\,di+\int_0^{\frac{2\Delta-\delta}{4\Delta-\delta}}(1-i)(2\Delta-\delta)\,di\right)\\ &+h(1-s)(1-h(s))\int_0^1 i(2\Delta-\tilde{\mu}\delta)\,di\\ &+(1-h(1-s))h(s)\left(\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}}(1-i)(2\Delta-\delta)\,di+\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1 i(\Delta-\tilde{\mu}\delta)\,di\right)\\ &+(1-h(s))(1-h(1-s))\int_0^1 i(\Delta-\tilde{\mu}\delta)\,di\end{aligned}$$

3. (B has not advanced) $q_B=0$ while $\tilde{\mu}=\mu$: Research potentially occurs on both paths in period 2. If neither advances, the social welfare (and private return) will be 0 as A will not be adopted under Pigouvian taxation. If A advances while B does not, then $\hat{I}(2)=1$ and social welfare (and A return) is $\frac{1}{2}(2\Delta-\delta)$. If B advances while A does not, then $\hat{I}(2)=0$ and the B return is $\frac{1}{2}\Delta$. Finally, if both advance, then $\hat{I}(2)=\frac{2\Delta-\delta}{3\Delta-\delta}$ with $v_A(2,1)=\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}}\left((1-i)(2\Delta-\delta)-i\Delta\right)\,di=\frac{(2\Delta-\delta)^2}{2(3\Delta-\delta)}$ and $v_B(2,1)=\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^1\left(i\Delta-(1-i)(2\Delta-\delta)\right)\,di=\frac{\Delta^2}{2(3\Delta-\delta)}$. At the beginning of period 2, $\hat{s}(2)$ will equate the average returns to scientists researching on each path; that is:

$$\frac{h(\hat{s}(2))}{\hat{s}(2))}\left(h(1-\hat{s}(2))\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}}\left((1-i)(2\Delta-\delta)-i\Delta\right)\,di+(1-h(1-\hat{s}(2)))\int_0^1(1-i)(2\Delta-\delta)di\right)$$

$$= \frac{h(1-\hat{s}(2))}{1-\hat{s}(2))} \left(h(\hat{s}(2)) \int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^{1} (i\Delta - (1-i)(2\Delta-\delta)) \; di + (1-h(\hat{s}(2))) \int_{0}^{1} i\Delta di \right)$$

or

$$\frac{h(\hat{s}(2))/\hat{s}(2)}{h(1-\hat{s}(2))/(1-\hat{s}(2))} = \frac{\int_0^1 i\Delta \, di - h(\hat{s}(2)) \left(\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^{1} (1-i)(2\Delta-\delta) \, di + \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} i\Delta \, di \right)}{\int_0^1 (1-i)(2\Delta-\delta) di - h(1-\hat{s}(2)) \left(\int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^{1} (1-i)(2\Delta-\delta) \, di + \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} i\Delta \, di \right)}$$

Thus, at the beginning of period 2, expected social welfare is:

$$\begin{aligned}
\mathbb{E}[V^{Tax}(1,0,\delta,\delta_B)] =& h(\hat{s}(2))h(1-\hat{s}(2)) \left(\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} (1-i)(2\Delta-\delta) \, di + \int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^{1} i(\Delta-\mu\delta) \, di \right) \\
&+ h(\hat{s}(2))(1-h(1-\hat{s}(2))) \int_0^1 (1-i)(2\Delta-\delta) di \\
&+ (1-h(\hat{s}(2)))h(1-\hat{s}(2)) \int_0^1 i(\Delta-\mu\delta) di + (1-h(1-s))(1-h(s))0 \\
=& h(\hat{s}(2)) \left(\int_0^1 (1-i)(2\Delta-\delta) di - h(1-\hat{s}(2)) \int_{\frac{2\Delta-\delta}{3\Delta-\delta}}^{1} (1-i)(2\Delta-\delta) \, di \right) \\
&+ h(1-\hat{s}(2)) \left(\int_0^1 i(\Delta-\mu\delta) di - h(\hat{s}(2)) \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta}} i(\Delta-\mu\delta) \, di \right) \\
=& \tfrac{1}{2} \left(h(1-\hat{s}(2)) \left(\Delta - \mu\delta + h(\hat{s}(2)) \tfrac{(2\Delta-\delta)((2\Delta-\delta)\mu\delta - \Delta(3\Delta-\delta))}{(3\Delta-\delta)^2} \right) + h(\hat{s}(2))(2\Delta-\delta) \right)
\end{aligned}$$

8.5.4 Ex Post Liability

1. (B harmful) $q_B = 1$ and $\delta_B = \delta$. From Table 1, it can be seen that this generates the same outcome as under a Pigouvian tax.

2. (B harm uncertain) $q_B = 1$ and δ_B remains uncertain with posterior probability of $\tilde{\mu}$:

Social welfare is:

$$
\begin{aligned}
\mathbb{E}[V^{Liab}(1,1,\delta,\delta)] =& h(1-s)h(s)\left(\int_{\frac{2\Delta-\delta}{4\Delta-(1+\tilde{\mu})\delta}}^{1} i(2\Delta-\delta\tilde{\mu}_B)\,di + \int_0^{\frac{2\Delta-\delta}{4\Delta-(1+\tilde{\mu})\delta}} (1-i)(2\Delta-\delta)\,di\right) \\
&+ h(1-s)(1-h(s))\int_0^1 i(2\Delta-\tilde{\mu}\delta)\,di \\
&+ (1-h(1-s))h(s)\left(\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\tilde{\mu}\Delta}} (1-i)(2\Delta-\delta)\,di + \int_{\frac{2\Delta-\delta}{3\Delta-\delta+\tilde{\mu}\Delta}}^{1} i(\Delta-\tilde{\mu}\delta)\,di\right) \\
&+ (1-h(s))(1-h(1-s))\int_0^1 i(\Delta-\tilde{\mu}\delta)\,di
\end{aligned}
$$

3. (B has not advanced) $q_B = 0$ while $\tilde{\mu} = \mu$: Research potentially occurs on both paths in period 2. If neither advances, the social welfare (and private return) will be 0 as A will not be adopted under ex post liability. If A advances while B does not, then $\hat{I}(2) = 1$ and social welfare (and A return) is $\frac{1}{2}(2\Delta-\delta)$. If B advances while A does not, then $\hat{I}(2) = 0$ and the B return is $\frac{1}{2}\Delta$. Finally, if both advance, then $\hat{I}(2) = \frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}$ with $v_A(2,1) = \int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}} ((1-i)(2\Delta-\delta) - i\Delta)\ di = \frac{(2\Delta-\delta)^2}{2(3\Delta-\delta+\mu\Delta)}$ and $v_B(2,1) = \int_{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}}^{1} (i\Delta - (1-i)(2\Delta-\delta))\ di = \frac{\Delta^2}{2(3\Delta-\delta+\mu\Delta)}$. At the beginning of period 2, $\hat{s}(2)$ will equate the average returns to scientists researching on each path; that is:

$$
\frac{h(\hat{s}(2))}{\hat{s}(2))}\left(h(1-\hat{s}(2))\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}} ((1-i)(2\Delta-\delta)-i\Delta)\ di + (1-h(1-\hat{s}(2)))\int_0^1 (1-i)(2\Delta-\delta)di\right)
$$

$$
= \frac{h(1-\hat{s}(2))}{1-\hat{s}(2))}\left(h(\hat{s}(2))\int_{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}}^{1} (i\Delta-(1-i)(2\Delta-\delta))\ di + (1-h(\hat{s}(2)))\int_0^1 i\Delta di\right)
$$

Thus, at the beginning of period 2, expected social welfare is:

$$\begin{aligned}\mathbb{E}[V^{Liab}(1,0,\delta,\delta_B)] =& h(\hat{s}(2))h(1-\hat{s}(2))\left(\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}}(1-i)(2\Delta-\delta)\,di+\int_{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}}^1 i(\Delta-\mu\delta)\,di\right)\\ &+h(\hat{s}(2))(1-h(1-\hat{s}(2)))\int_0^1(1-i)(2\Delta-\delta)di\\ &+(1-h(\hat{s}(2)))h(1-\hat{s}(2))\int_0^1 i(\Delta-\mu\delta)di+(1-h(1-s))(1-h(s))0\\ =& h(\hat{s}(2))\left(\int_0^1(1-i)(2\Delta-\delta)di-h(1-\hat{s}(2))\int_{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}}^1(1-i)(2\Delta-\delta)\,di\right)\\ &+h(1-\hat{s}(2))\left(\int_0^1 i(\Delta-\mu\delta)di-h(\hat{s}(2))\int_0^{\frac{2\Delta-\delta}{3\Delta-\delta+\mu\Delta}} i(\Delta-\mu\delta)\,di\right)\\ =& \tfrac{1}{2}h(1-\hat{s}(2))\left(\Delta-\mu\delta+h(\hat{s}(2))\tfrac{(2\Delta-\delta)((2\Delta-\delta)\mu\delta-\Delta(3\Delta-\delta+\mu\Delta))}{(3\Delta-\delta+\mu\Delta)^2}\right)\\ &+\tfrac{1}{2}h(\hat{s}(2))(2\Delta-\delta)\end{aligned}$$

References

Acemoglu, Daron, "Diversity and technological progress," in "The Rate and Direction of Inventive Activity Revisited," University of Chicago Press, 2011, pp. 319–356.

_ , "Harms of AI," Technical Report, National Bureau of Economic Research 2021.

_ , "Distorted Innovation: Does the Market Get the Direction of Technology Right?," *AEA Papers and Proceedings*, 2023, *113*, 1–28.

_ **and Simon Johnson**, *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity*, Hachette UK, 2023.

_ **and Todd Lensman**, "Regulating Transformative Technologies," *American Economic Review: Insights*, 2024.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Press, 2022.

Bryan, Kevin A, "The Perils of Path Dependence," in Joshua S. Gans and Sarah Kaplan, eds., *Survive and Thrive: Winning Against Strategic Threats to Your Business*, Dog Ear Publishing, 2017.

_ **and Jorge Lemus**, "The direction of innovation," *Journal of Economic Theory*, 2017, *172*, 247–272.

Brynjolfsson, Erik, "The Turing Trap," *Daedalus*, 2022, *151* (2), 272–287.

Cowan, Robin, "Nuclear power reactors: a study in technological lock-in," *Journal of Economic History*, 1990, *50* (3), 541–567.

Gans, Joshua S., "How Learning About Harms Impacts the Optimal Rate of Artificial Intelligence Adoption," *Economic Policy*, 2024.

Guerreiro, Joao, Sergio Rebelo, and Pedro Teles, "Regulating Artificial Intelligence," Technical Report, National Bureau of Economic Research 2023.

Hopenhayn, Hugo and Francesco Squintani, "On the Direction of Innovation," *Journal of Political Economy*, 2021, *129* (7), 1991–2022.

Jones, Charles I., "The AI Dilemma: Growth versus Existential Risk," *American Economic Review: Insights*, 2024, *6* (4), 575–590.

Koh, Andrew and Sivakorn Sanguanmoo, "Robust Technology Regulation," *arXiv preprint arXiv:2408.17398*, 2024. `https://arxiv.org/abs/2408.17398`.

McLaughlin, C.C., "The Stanley Steamer: A Study in Unsuccessful Invention," *Explorations in Entrepreneurial History*, 1954, *7.*

O'Donoghue, Ted, Suzanne Scotchmer, and Jacques-Francois Thisse, "Patent Breadth, Patent Life, and the Pace of Technological Progress," *Journal of Economics & Management Strategy*, 1998, *7* (1), 1–32.

Russell, Stuart, *Human Compatible: AI and the Problem of Control*, Penguin Uk, 2019.

Trammell, Philip and Leopold Aschenbrenner, "Existential Risk and Growth," December 2024. Unpublished manuscript, accessed December 7, 2024.