NBER WORKING PAPER SERIES

HOW LEARNING ABOUT HARMS IMPACTS THE OPTIMAL RATE OF ARTIFICIAL INTELLIGENCE ADOPTION

Joshua S. Gans

Working Paper 32105
http://www.nber.org/papers/w32105

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2024

Joshua Gans has drawn on the findings of his research for both compensated speaking engagements and consulting engagements. He has written the books Prediction Machines, Power & Prediction, and Innovation + Equality on the economics of AI for which he receives royalties. He is also chief economist of the Creative Destruction Lab, a University of Toronto-based program that helps seed stage companies, from which he receives compensation. He conducts consulting on anti-trust and intellectual property matters with an association with Charles River Associates and his ownership of Core Economic Research Ltd. He also has equity and advisory relationships with a number of startup firms. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w32105

How Learning About Harms Impacts the Optimal Rate of Artificial Intelligence Adoption
Joshua S. Gans
NBER Working Paper No. 32105
February 2024
JEL No. L51,O33

ABSTRACT

This paper examines recent proposals and research suggesting that AI adoption should be delayed until its potential harms are properly understood. It is shown that conclusions regarding the social optimality of delayed AI adoption are sensitive to assumptions regarding the process by which regulators learn about the salience of particular harms. When such learning is by doing -- based on the real-world adoption of AI -- this generally favours acceleration of AI adoption to surface and react to potential harms more quickly. This case is strengthened when AI adoption is potentially reversible. The paper examines how different conclusions regarding the optimality of accelerated or delayed AI adoption influence and are influenced by other policies that may moderate AI harm.

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

Recently, advances in (generative) artificial intelligence (AI) have accelerated adoption across many sectors.[1] This rapid adoption has created calls for regulatory intervention over concerns that there are potential harms from AI adoption (Acemoglu, 2021). In one prominent case, those calls from computer scientists and AI innovators were for an immediate and quite consequential intervention.

> [W]e call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4. This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium.
>
> AI labs and independent experts should use this pause to jointly develop and implement a set of shared safety protocols for advanced AI design and development that are rigorously audited and overseen by independent outside experts. These protocols should ensure that systems adhering to them are safe beyond a reasonable doubt. This does not mean a pause on AI development in general, merely a stepping back from the dangerous race to ever-larger unpredictable black-box models with emergent capabilities.[2]

This is a call for ex ante regulation of AI as well as advocating a policy-instrument: a pause on AI adoption. In other words, with respect to AI, in contrast to the usual concerns that arise in the economics of innovation that development and adoption may be too slow relative to the social optimum (e.g., Arrow (1972) and Segal & Whinston (2007)), these calls claim that AI adoption is happening too rapidly.

However, regulation that slows down the development and adoption of technology is not new. In some cases, technologies have raised ethical dilemmas (e.g., stem-cell research,[3] human cloning and eugenics), but also concerns regarding catastrophic risk (e.g., gain-of-function research). While the concerns with regard to AI sometimes involve such high risks,[4] the recent push is with respect to potential harms from AI that are not unprecedented in recent history, including the job implications of automation, the production of misinformation and threats to cybersecurity. These harms are similar to those put forward by the

[1] OpenAI's ChatGPT was the fastest technology to reach 100 million users in history achieving that goal just two months after launch. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[2] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[3] Furman et al. (2012)

[4] https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

information technology and Internet revolutions. Importantly, it is uncertain whether those potential harms will become real or whether regulating AI alone would mitigate them.

In contrast to information regarding actual harms from AI, there is already important evidence indicating that such technologies are generating substantial benefits. All of the recent developments in AI are advances in machine learning, a subset of the broad field of computational statistics. AI today is not necessarily a new form of intelligence but what might be called *prediction machines* that allow for cheaper and superior prediction by machines than people could achieve (Agrawal et al., 2018). Numerous studies show that prediction can automate aspects of tasks and improve productivity, especially among workers with relatively less skill or experience.[5] Consequently, it is not at all clear that some of the more pessimistic forecasts of the usefulness of AI are being borne out by current AI adoption.[6]

Given the relative weights of information regarding benefits and potential harms, there is a real risk that regulation, and even the prospect of it, could delay AI adoption. As noted above, that may be the intent of these regulations. However, it should be noted that such uncertainty has had a chilling effect on innovation in other areas of innovation. Galasso & Luo (2022) found that patenting on implants decreased by 35 percent following a potential legal extension of product liability law in the class.

This paper aims to examine the optimal rate of AI adoption from a social welfare perspective, given that the potential harms from AI adoption are uncertain. This is not a new question in the literature. Jones (2023) has examined the optimal timing for the use of AI given the potential for existential risk.[7] In relation to concerns with the current generation of AI technologies, Acemoglu & Lensman (2023) examines the efficacy of delaying AI adoption in order to allow for the resolution of uncertainty regarding potential harms. They identify a precautionary motive for delaying AI adoption in order to become more confident that such adoption will not result in harm.[8] They also examine some other instruments, such as regulatory sandboxes and Pigouvian taxation, as ways of implementing the social optimum.[9]

The contribution of this paper is to provide a simple model that demonstrates the impor-

[5] See Choi & Schwarcz (2023), Noy & Zhang (2023), Dell'Acqua et al. (2023), Kanazawa et al. (2022), Brynjolfsson et al. (2023) and Agrawal et al. (2024).

[6] For discussions see, Gans & Leigh (2019) and Agrawal et al. (2023a).

[7] In the case of AI adoption, much of the concerns about catastrophic harm emerge not in relation to AI as it is currently developed as prediction machines but instead through perceived threats from artificial general intelligence (AGI). https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

[8] This is work within a standard rationality framework. Other work on precautionary motives considers bounded rationality (Grant & Quiggin (2013)) is not an approach considered here.

[9] This paper is also related to a literature on regulatory learning. The closest relevant work is Spitzer & Talley (2014), which examines how regulators assess new regulatory interventions when current interventions are already in place. While their focus is not on delay or accelerated adoption, some of the trade-offs here have similarities in their model. That said, their paper is focused on drawing out judicial versus regulatory preferences for different learning approaches.

tance of alternative assumptions regarding the learning process and how these impact on the rate of AI adoption. Section 2 develops a simple framework that starts from the assumption that regulators can learn about the potential harms of AI by learning by doing – that is, by allowing or promoting the rapid adoption of AI to provide the conditions under which potential harms may surface. If such harms do not occur, the adoption of AI can continue. However, if they do arise, then the regulator can engage in ex post policy implementations to curtail adoption and potential harm. Of course, in a case where adoption is difficult (or impossible) to reverse, the learning effect is counter-balanced by the precautionary effect identified by Acemoglu & Lensman (2023). Importantly, it is demonstrated that the notion that AI adoption may be accelerated for learning reasons arises precisely because of the assumption that learning requires real-world adoption. By contrast, Acemoglu & Lensman (2023) assume that learning is exogenous and takes time, something I term "lab learning," which is why they recommend delay as a mechanism for learning, especially when adoption may be irreversible.[10] Consequently, the framework here highlights the roles of alternative learning assumptions, the role of irreversibilities, and the differences in socially optimal adoption rates depending upon whether the planner is optimistic or pessimistic about the net social value of AI ex ante.

The paper then highlights policies that might be complementary depending on whether the social planner wants to accelerate AI adoption. These include policies that will enable ex post mitigation of potential harms that emerge from AI and those that ensure that, if delay is used, the required learning takes place. Finally, the role of having a more diverse pool of potential architectures to develop AI is explored.

2 A Simple Framework

The basic trade-offs involved regarding the rate of AI adoption can be captured in a simple model. Consider a two-period model where the social planner has a discount factor of $\delta \in (0, 1)$. Let x_t, assumed to be in $(0, 1)$, be the rate of AI adoption in $t \in \{1, 2\}$, where x_t could stand for the scale and/or scope of that adoption. The (flow) benefits from such adoption are $b > 0$ per unit of adoption. AI adoption may not only create benefits. There may be damage from AI. Damage, d, results in harm to social welfare that scales at the rate of AI adoption. Thus, total damage in each period could be as high as $x_t d$. It is assumed that $b < d$ in what follows; that is, if the damage is known to occur, it is not optimal to

[10] In a recent paper, Guerreiro et al. (2023) explore the difference between private and social returns to testing under various regulatory regimes. They do not examine the rate of adoption explicitly, although interpretations of their model may relate to the lab-learning conclusions reached below.

adopt AI.

An important element of the model is how the social planner learns about potential harms. It is assumed that the planner has a prior probability that damage occurs of μ. Harm may occur but be unobserved prior to adoption choices being made in either period – for notational convenience, if there is harm, that fact is revealed to the planner at the end of the second-period. A key assumption here is that actual adoption, x_1, is required for a signal of harm to be generated to inform adoption decisions. Thus, it is assumed, for simplicity, that if AI is actually harmful, the probability that the planner receives a signal indicating harm is x_1. Otherwise, with probability $1 - x_1$, there is no signal of harm. Thus, at the beginning of the second period, the planner's posterior probability, $\tilde{\mu}$, that AI adoption may result in damage in $t = 2$, conditional on it not having risen in $t = 1$, is given by:

$$\tilde{\mu} = \frac{\mu(1 - x_1)}{\mu(1 - x_1) + (1 - \mu)}$$

The learning process assumed here is that unless damage is observed during a period, the planner updates the likelihood that damage will arise by lowering their 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 damage reveals the state perfectly, while a signal there is no damage during a period involves the possibility of a false positive.

Importantly, the rate of learning is driven by the rate of AI adoption – a form of learning by doing. This stands in contrast to the learning process considered by Acemoglu & Lensman (2023).[11] In their formulation, the rate of learning is exogenous, and learning occurs with time, independent of the level of AI adoption. That is, damage can be signalled without adopting any AI at all. Below, the criticality of this assumption is explored as well as what happens when there are other imperfections in signals, including the possibility of false negatives.

In what follows, two broad cases of AI adoption are considered. In the first, adoption, should it occur, is reversible, while in the second, it is irreversible. It is demonstrated that this creates distinct rationales for learning and also impacts the optimal rate of AI adoption.

2.1 Reversible Adoption

Consider first the case where AI adoption is reversible; therefore, if damage is known to arise, it can be scaled back. By our earlier assumption that $b < d$, the social planner will

[11] Similarly, Guerreiro et al. (2023) examine a form of learning termed 'beta testing' an important feature of which is that it does not require actual adoption to generate a signal.

want to de-adopt AI if it is known to result in damage.[12] Thus, if the planner learns that AI adoption results in damage at the end of period 1, it is optimal to set $x_2^* = 0$ and not continue to use the AI.

As an illustration of the trade-offs involved in adopting AI, assume, for the moment, that $b > \mu d$. When this condition holds, the planner is pre-disposed to adopt AI absent any information regarding the possibility of damage. Let V_t be the present discounted social payoff in period t. Under this assumption, in period 1, the planner chooses x_1 to maximise:

$$\mathbb{E}[V_1] = x_1(b - \mu d) + (1 - \mu x_1)\delta\mathbb{E}[V_2]$$

That is, with probability, μx_1, the planner receives a signal that AI adoption leads to damage and should not be used further, while with probability $1 - \mu x_1$, they receive no signal. As $b > \mu d$, then it is optimal to set $x_1^* = 1$ in the first period and every period thereafter if damage is not observed in that period. Setting $x_1^* = 1$ improves the quality of decision-making in the second-period while increasing the first-period expected payoff.

The following proposition characterises the optimal outcome for all $b < d$.

Proposition 1 *Suppose AI adoption is reversible. It is socially optimal to set $x_1^* = 1$ if $\mu \leq \frac{(1+\delta)b}{d-\delta b}$. Otherwise, it is not socially optimal to adopt AI at all.*

Proof. The case where $b \geq \mu d$ is shown in the text. Assume now that $b < \mu d$. In this case, conditional on no signal being received at $t = 1$, $\mathbb{E}[V_2] = x_2(b - \tilde{\mu} d) = x_2(b - \frac{\mu(1-x_1)}{\mu(1-x_1)+1-\mu}d)$. Note that it is only optimal to set $x_2 > 0$ if $b \geq \frac{\mu(1-x_1)}{\mu(1-x_1)+1-\mu}d)$ or $x_1 \geq \frac{b-d\mu}{(b-d)\mu}$ in which case, $x_2^* = 1$. Otherwise, $x_2^* = 0$. Thus, x_1 is chosen to maximise:

$$\mathbb{E}[V_1] = x_1(b - \mu d) + \begin{cases} \delta\left((1 - \mu x_1)b - \mu(1 - x_1)d\right) & x_1 \geq \frac{b-d\mu}{(b-d)\mu} \\ 0 & x_1 < \frac{b-d\mu}{(b-d)\mu} \end{cases}$$

This objective function is linear and decreasing in x_1 up to where $x_1 > \frac{b-d\mu}{(b-d)\mu}(> 0)$ beyond which, it is increasing in x_1. Therefore, setting $x_1 = 0$ or $x_1 = 1$ is optimal.

Setting $x_1 = 0$ maximises short-run welfare but involves no learning as to the likelihood of damage. Alternatively, by setting $x_1 = 1$ the planner receives a perfect signal of whether there will be damage from continued AI adoption in period 2. Thus, if no signal of damage is received, period 2 welfare is $V_2 = b$ (recalling that it is optimal to set $x_2 = 1$), while if

[12] If, alternatively, $b > d$, regardless of whether the planner learns if there is damage of d, it is still optimal to continue large-scale AI adoption.

there is a signal of damage, $V_2 = 0$. Therefore:

$$\mathbb{E}[V_2|x_1 = 1] = (1 - \mu)b$$

Given this, at $t = 1$:

$$\mathbb{E}[V_1] = b - \mu d + \delta(1 - \mu)b$$

This is positive if and only if $\mu < \frac{(1+\delta)b}{d-\delta b}$. Thus, for a sufficiently low prior, μ, it is optimal to accept a short-run expected loss to learn the level of damage associated with AI adoption. Note that $\frac{(1+\delta)b}{d-\delta b} > \frac{b}{d} \Leftrightarrow b + d > 0$. This completes the proof. $\blacksquare$

This proposition characterises when it is optimal to accelerate the adoption of AI. When the planner is optimistic about AI prospects (i.e., when $\mu < \frac{b}{d}$), acceleration allows for learning that can mitigate harm earlier should be observed. Importantly, acceleration can be optimal even if the planner is pessimistic (but not too pessimistic) about AI prospects (i.e., when $\frac{b}{d} < \mu < \frac{(1+\delta)b}{d-\delta b}$). AI adoption affords the opportunity to learn about potential harm and become more confident that continued adoption will not result in harm. A key assumption here, however, is that uncertainty is fully resolved by setting $x_1 = 1$. However, so long as sufficient uncertainty is resolved that would cause the planner to continue to adopt AI if there were no adverse signals, even a pessimistic planner would choose to accelerate AI adoption.

2.2 Irreversible Adoption

Consider now the case where the adoption of AI is irreversible. That is, if x_1 is the rate of adoption in period 1, then x_2 is constrained to be no less than x_1 even if that results in damage, d. This assumption is consistent with the notion that AI adoption may make switching back to an older technology difficult. As in Acemoglu & Lensman (2023), the precise reason for irreversibility is left unmodelled.[13]

Working backwards, consider the planner's problem in period 2. The planner is constrained to choose $x_2 \geq x_1$. If AI is known to result in damage, then the second-period social value is $V_2(d) = x_2(b - d)$, and it is optimal to set $x_2^* = x_1$. On the other hand, if damage has not emerged in period 1, then

$$\mathbb{E}[V_2|0] = x_2(b - \tilde{\mu}d)$$

[13] Acemoglu & Lensman (2023) also assume that irreversibility is probabilistic whereas, in this paper, the two extreme cases of reversible and irreversible adoption are considered separately.

where $\mathbb{E}[V_2|0]$ is the expected value of second-period social welfare conditional on a signal of no damage. Thus, if $b \geq \tilde{\mu}d$, $x_2^* = 1$, otherwise $x_2^* = x_1$.

Given these second-period choices, under what conditions will the planner choose to adopt AI in the first period when such adoption cannot be reversed, even if harm is known to occur? The following proposition characterises the optimal outcome.

Proposition 2 *Suppose AI adoption is irreversible. The socially optimal level of x_1 is:*

$$x_1^* = \begin{cases} 1 & \mu \leq \frac{b}{(1+\delta)d-\delta b} \\ \frac{b-\mu d+b(1-\delta\mu)}{2\delta\mu(d-b)} & \mu \in [\frac{b}{(1+\delta)d-\delta b}, \frac{(1+2\delta)b}{(1+\delta)d+\delta b}] \\ 0 & \mu \geq \frac{(1+2\delta)b}{(1+\delta)d+\delta b} \end{cases}$$

Proof. There are several cases to consider. First, suppose that $b > \mu d$. Note that this implies that $b \geq \tilde{\mu}d$, so it is optimal to set $x_2 = 1$ so long as no adverse signal of harm is received. Otherwise, the optimal $x_2 = x_1$. Taking this into account, in period 1, the planner chooses x_1 to maximise:

$$\begin{aligned} \mathbb{E}[V_1] &= x_1(b-\mu d) + (1-\mu x_1)\delta(b-\tilde{\mu}d) + \mu\delta x_1^2(b-d) \\ &= (x_1+\delta)(b-\mu d) - \mu\delta x_1(1-x_1)(b-d) \end{aligned}$$

Maximising this with respect to x_1 gives:

$$x_1^* = \max\{\min\{\tfrac{b-\mu d+b(1-\delta\mu)}{2\delta\mu(d-b)}, 1\}, 0\}$$

Observe that $x_1^* \in (0,1)$ if $\mu \in (\frac{b}{(1+\delta)d-\delta b}, \frac{b}{(1-\delta)d+\delta b})$. It is easy to check that this interval is non-empty. Note that when $x_1^* = \frac{b-\mu d+b(1-\delta\mu)}{2\delta\mu(d-b)}$, $b \geq \tilde{\mu}d$ requires that:

$$\mu \leq \frac{(1+2\delta)b}{(1+\delta)d+\delta b}$$

The RHS is strictly less than $\frac{b}{(1-\delta)d+\delta b}$ and strictly greater than $\frac{b}{(1+\delta)d-\delta b}$.

Second, suppose that $b < \mu d$. If $b > \tilde{\mu}d$, the analysis for $b \geq \mu d$ applies and x_1^* can be positive. However, unlike where $b \geq \mu d$, when $b < \mu d$, it is possible that $\mu > \bar{\mu}$ so that $b < \tilde{\mu}d$ for all x_1. If this is the case, then $x_2^* = x_1^*$. Thus, the planner chooses x_1 to maximise:

$$\mathbb{E}[V_1] = x_1(b-\mu d) + (1-\mu x_1)\delta x_1(b-\tilde{\mu}d) + \mu\delta x_1^2(b-d)$$

The derivative of this objective with respect to x_1 is $(1+\delta)(b-\mu d)$, which is always negative.

Thus, $x_1^* = 0$. ■

In contrast to the reversible adoption case, while setting $x_1 = 1$ resolves the uncertainty by period 2, it involves a cost in terms of having to continue to incur some social loss if AI turns out to be harmful. Thus, for a planner with optimistic beliefs (i.e., $b \geq \mu d$), this creates a precautionary motive potentially slowing the adoption of AI in the expectation that sufficient "good news" may be delivered to be confident of full adoption in period 2. When the planner is pessimistic (i.e., $b < \mu d$), absent any learning, it is not optimal to adopt AI. However, Proposition 2 shows that, even with irreversibilities, it can be optimal to set $x_1^* > 0$, therefore accelerating AI adoption relative to the case without any learning.

Interestingly, in a model that extended beyond period 2, the set-up here has an interesting implication about the broader dynamics of adoption and learning. It suggests that it may be optimal when $b < \mu d$ to choose a positive level of x_1 initially, with the expectation that a sufficient number of periods (or rounds) of observation may arise to eventually given the planner confidence to increase adoption. Thus, an initial level of adoption may be regarded as the 'cost' of engaging in learning similar to the lab-learning model to be discussed next if such learning involved per period costs.[14]

2.3 Lab Learning

Thusfar, the model has considered learning by doing whereby uncertainty regarding damage is only resolved if AI is adopted and that a greater scale of adoption is more likely to surface the possibility of damage if such damage is possible. By contrast, Acemoglu & Lensman (2023) examines a model whereby time alone – independent of adoption – resolves uncertainty regarding any harms of AI. In particular, translating their assumption into the structure of the present model, they assume that:

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

where $\lambda < 1$ is a constant. In other words, the safety of AI can be evaluated in a lab instead of requiring any AI adoption in the market. Hence, this is termed "lab learning."[15] The process for such lab learning is that if AI involves harm, then there is a probability, λ, that a lab will discover evidence of that harm. A higher λ implies that if there is no evidence of

[14] Guerreiro et al. (2023) do consider such costs but also allow them to be minimised by setting what would be in the model here, x_1, as close to zero as possible.

[15] Spitzer & Talley (2014) in examining financial regulation examine a similar dichotomy. Their "analytical learning" is similar to "lab learning" while their "field experimentation" is similar to "learning by doing."

harm found during a period in the lab, then a more precise inference can be drawn that AI is, in fact, safe. Of course, in contrast to learning by doing, the precision of the signal from lab learning is exogenous.

Under this assumption, there is no separate reason to engage in AI adoption for the purposes of learning. If the signal from lab learning is precise enough (that is, if $b \geq \tilde{\mu}d \Leftrightarrow \lambda > \frac{b-\mu d}{\mu(b-d)}$ so that it is optimal to adopt AI in period 2 if there is no signal of harm), then the expected social value under lab learning when AI adoption is reversible is:

$$\mathbb{E}[V_1] = x_1(b - \mu d) + \delta(1 - \mu\lambda)(b - \tilde{\mu}d)$$

When adoption is reversible, the first-period adoption decision does not impact second-period outcomes, so AI is adopted then if $b > \mu d$ and delayed one period if $b < \mu d$. When adoption is irreversible, the expected social value under lab learning is:

$$\mathbb{E}[V_1] = \begin{cases} x_1(b - \mu d) + \delta(1 - \mu\lambda)(b - \tilde{\mu}d) + \delta\mu\lambda x_1(b - d) & \mu \leq \frac{b}{d+\lambda(b-d)} \\ x_1(b - \mu d) + \delta(1 - \mu\lambda)x_1(b - \tilde{\mu}d) + \delta\mu\lambda x_1^2(b - d) & \mu > \frac{b}{d+\lambda(b-d)} \end{cases}$$

In this case, the precautionary motive (as discussed above) limits AI adoption in period 1, even if $b > \mu d$.

2.4 Comparing Learning Modes

It is instructive to compare what different learning modes – learning by doing versus lab learning – imply for whether AI adoption should be accelerated or delayed relative to a counterfactual where there is no learning mode available. When there is no learning mode then $\mathbb{E}[V_1] = x_1(b - \mu d) + \max_{x_2} \delta x_2(b - \mu d)$. Adoption in the first period is driven solely by the impact of adoption on $x_1(b - \mu d)$. If $b \leq \mu d$, there is no adoption in either period and $\mathbb{E}[V_1] = 0$, otherwise if $b > \mu d$, then adoption in each period is 1 and $\mathbb{E}[V_1] = (1 + \delta)(b - \mu d)$.

Learning is said to "accelerate" ("delay") AI adoption if the marginal return to x_1 is higher (lower) when learning is available compared to when it is not. Thus, with learning by doing, in the reversible case, the marginal return to x_1 is $b - \mu d - \delta\mu(b - d)$ compared with $b - \mu d$ when there is no learning. The difference is $-\delta\mu(b - d) > 0$, so optimal AI adoption is accelerated by learning. By contrast, with lab learning, in the reversible case, with or without learning, the marginal return to x_1 is $b - \mu d$ and so there is no impact from learning on the rate of AI adoption in the first period.

When AI adoption is irreversible, beliefs, in terms of the level of μ, factor into the calculation. Note that, because of irreversibility, without learning, the marginal return to

	Reversible	Irreversible
Learning by Doing	Accelerate	Delay if optimistic Accelerate if pessimistic
Lab Learning	Same	Delay if optimistic Same if pessimistic

Table 1: Impact of Learning on AI Adoption in the First Period

x_1 is $(1+\delta)(b-\mu d)$. Under learning by doing with irreversible AI adoption, the marginal return to x_1 is: $b-\mu d-\delta\mu(1-2x_1)(b-d)$ so long as $\mu \leq \frac{(1+2\delta)b}{(1+\delta)d+\delta b}$ (that is, $b \geq \tilde{\mu}d$). Here, the difference in marginal returns is:

$$-\delta(b-\mu d+\mu(1-2x_1)(b-d))$$

The sign of this depends on the value of x_1 chosen under learning. From Proposition 2, it is easy to see this is negative if $x_1^*=1$ (as this also requires $b>\mu d$). While, if $x_1^*=\frac{b-\mu d+b(1-\delta\mu)}{2\delta\mu(d-b)}$ it is negative only if $b>\mu d$. Otherwise, for $b<\mu d$, if $\mu<\frac{(1+2\delta)b}{(1+\delta)d+\delta b}$, the difference in marginal returns is positive. However, $\mu>\frac{(1+2\delta)b}{(1+\delta)d+\delta b}$, the difference in marginal returns is unchanged. Thus, if $b>\mu d$, adoption is delayed under learning compared with no learning, while if $b<\mu d$, adoption may be accelerated or the same depending on μ.

For lab learning, when AI adoption is irreversible, the marginal return to x_1 is:

$$\begin{cases} b-\mu d+\delta\mu\lambda(b-d) & \mu \leq \frac{d}{d+\lambda(b-d)} \\ (1+\delta)(b-\mu d) & \mu > \frac{d}{d+\lambda(b-d)} \end{cases}$$

Subtracting $(1+\delta)(b-\mu d)$ gives, for μ low, $\delta((1-\lambda)d-(1-\mu\lambda)b)$ which is always negative and 0 for μ high. Thus, learning (weakly) causes delay.

Given this, the impact of learning on the rate of AI adoption is summarised in Table 1. When adoption is reversible, learning by doing accelerates the rate of adoption in the first period, regardless of whether the planner is optimistic or pessimistic. By contrast, lab learning does not alter behaviour at all and adoption in the first period will be determined by other factors. When AI adoption is irreversible, the Acemoglu & Lensman (2023) result that adoption should be (weakly) delayed emerges – again, regardless of beliefs. However, under learning by doing, the adoption decisions of optimistic versus pessimistic planners are tempered. An optimistic planner becomes more cautious, while a pessimistic planner has an incentive to accelerate adoption in order to learn if such learning could lead to further adoption in the future.

3 Interaction with other Policies

Given this simple framework, it is useful to consider how this impacts other policy interventions beyond simply accelerating or delaying AI adoption. This section explores those options that potentially complement learning and the consequent acceleration or delay, respectively.

3.1 Improving Learning

While this paper has distinguished between learning by doing and lab learning in terms of the implication for the optimal rate of AI adoption, it is instructive to consider what factors may improve learning in terms of lowering learning costs or raising the quality of learning.

For learning by doing, the precision of the signal comes from the rate of AI adoption in the first period, x_1. Choosing a higher x_1, only involves costs if the planner is pessimistic (creating a marginal cost per unit of adoption of $-(b-\mu d)$ and/or AI adoption is irreversible (leading to a marginal cost per unit of adoption of $-\delta\mu(b-d)$). By comparison, for lab learning, the precision of the signal is exogenous (i.e., λ) but otherwise only has a cost in terms of time. While, in the model, the potential precision of the signal from learning by doing is higher than lab learning, a careful ranking of the two on this dimension is likely to be subtle and difficult to consider within the theoretical scope of this paper.

Some factors that determine the efficacy of each learning mode can be identified. Learning by doing involves real-world learning that, while potentially imperfect, has the advantage that the harms identified are clearly real and applicable, motivating a strong ex post regulatory response. Lab learning translates when there is model stability between the underlying processes being studied and real-world adoption. Such pre-screening of products is common in many domains, including pharmaceutical drugs, medical treatments, building codes and broad areas of product safety, especially regarding physical products. Such lab learning could apply where AI is used in those existing domains, and these would be covered by existing regulations.[16]

Learning issues are more challenging when AI is interacting with human behaviour. Model stability may be an issue because of heterogeneous behavioural responses in this case. However, lab learning can potentially only identify direct, and not equilibrium, effects of AI adoption. In an educational setting, generative AI may make assessment more difficult and, hence, could demotivate students. Still, over the longer term, such AI allows for new and personalised sources of learning. This may lead to delays in the deployment of non-harmful AI. Similarly, it may be that a new social media algorithm's short-term effects are innocuous but have broader effects on behaviour when deployed. In these cases, learning

[16] See, for instance, the series of papers in Agrawal et al. (2023b).

itself can be accelerated by accelerating AI adoption, and ex post interventions can be put into play quicker than if such AI had been pre-cleared.[17] Nonetheless, because it has the potential to be a least-cost method of learning, finding ways to improve the precision of lab learning (λ) would be a potentially fruitful policy direction.[18]

3.2 Investment in Harm Mitigation

In the analysis thusfar, if AI turns out to be harmful, if possible, the policy response is to reduce the adoption of AI. Learning provides a means of enabling that policy response to occur and, in this respect, reduces the expected harm from AI adoption. However, there are also policies that can pre-emptively reduce the magnitude of AI harm should it be identified. For instance, investments could be made in the capacity to mitigate harms that arise. This is a form of insurance.[19]

There are numerous examples of policies that could mitigate harm; in the model, reducing d. The concerns with respect to AI involve the automation and replacement of jobs, the use of AI to spread misinformation and the potential issues that might arise if bad actors utilise AI. In each of these cases, there are ex post policies that can be implemented to mitigate the adverse consequences of AI should these harms arise. This would involve, inter alia, investments in social security and education, tools to counter misinformation and develop trusted information sources, and general investments in cyber-security.[20]

Other policies are potentially broader in their potential scope. For instance, there may be institutional changes that increase the speed by which policymakers can react to mitigate harm. Hadfield & Clark (2023) propose the establishment of an independent sector of licensed private regulators to manage potential harms from AI, serving as an investment in mitigating unforeseen risks. In their conception, governments define the required outcomes for AI regulation, which could be either metrics-based, like the frequency of fraudulent transactions, or principles-based, such as maintaining a low incidence of accidents in autonomous vehicles. Private regulators, licensed by governments, are evaluated based on their ability to achieve these government-defined objectives, like non-discrimination in AI hiring practices. Entities subject to regulation, such as social media platforms and banks, are mandated to procure

[17] Agrawal et al. (2022) argue that there may be both private and public motivations for an FDA-like pre-clearance assessment for AI in terms of its safety. The idea here is to establish the capacity to give AI adopters assurances of safety before deployment.

[18] A recent paper by Guerreiro et al. (2023) explores some of the potential policy implications that arise when the cost of lab learning is reduced.

[19] Ehrlich & Becker (1972) define insurance as investments that can be made to reduce the losses from adverse outcomes.

[20] Research in other domains has shown that when concerns regarding harms surface, there can be a substantial industry response in developing risk-mitigating technologies; see Galasso & Luo (2021).

the services of these licensed private regulators. The government also sets rules for private regulators to foster competition and maintain the integrity of regulation.

Hadfield & Clark (2023) provide several examples that relate to industries where the adoption of AI may be ex post irreversible. For instance,

> A private regulator in the banking industry might require a bank using ML to analyze customer data and develop new products to implement differential privacy techniques (Dwork & Roth 2014) to minimize the likelihood that a customer is harmed by the use of their data. The regulator could prescribe the specific techniques/algorithms to use; or it could establish a procedure for the banks that it regulates to propose techniques that survive tests conducted by the regulator. (Hadfield & Clark (2023), p.16)

Note that the goal here is not to supplant what a regulator might want but to provide incentives to build up the capacity to implement regulations and do so in a timely manner. While there can be issues associated with regulatory capture that emerge with such a model, if effective, its existence can play an insurance role that can be triggered ex post after the extent of potential harm is revealed.

While the above example considers regulatory capabilities that can be built up to mitigate realised AI harms, it is also possible to consider infrastructure investments that can be provided as ways of actually protecting against hypothesised avenues for harm. As an example, consider one of the implications of generative AI being able to produce images, documents or audio/video recordings that are distortions of historical events.[21] The concern is that sometime in the future, a bad actor may distort the historical record for their own purposes. It may not be able to distinguish whether the AI-doctored images were accurate or not. Potentially, this could lead to a wholesale re-writing of the written record.

One possible way to counter this is to establish a record of the truth pre-emptively. One option is to create a digital library that stores these records that would be the trusted authority on these matters. However, the problem is that such trust would require that a bad actor could not alter the library records or otherwise censor them. To provide some infrastructure that is immune from such challenges will likely require a more active response today. One possibility would be to register the library records with a timestamp that could not be adjusted. A technique for achieving this was developed by Haber & Stornetta (1991) and was the foundation for the modern blockchain. The idea is that today, a library could be created with digital assets time-stamped on a decentralised blockchain. It would not be

[21] This is not a new phenomenon and was extensively used in the former Soviet Union; https://www.history.com/news/josef-stalin-great-purge-photo-retouching

possible to alter those time stamps; hence, if someone claimed a record was, at the very least, from pre-2024, without that time stamp, that claim would not be trusted. By contrast, the library records could easily pass that verification check.[22]

In summary, these investments that either allow policy-makers to more effectively mitigate AI harms ex-post or to reduce the probability of those harms arising in the first place, all raise the returns for AI adoption, in general, and in the case where learning by doing is being utilised by regulators, lowering the cost of associated with such learning. However, there is an interesting question of how policies for harm mitigation, as described here, interact with the use of learning to drive AI adoption levels.

To examine this, consider the complementarity between harm mitigation and the use of learning to adjust AI adoption rates. Intuitively, reducing d reduces the direct cost of learning by doing, but it also reduces the benefit from additional learning as the value of such learning, the ability to avoid incurring d, is lowered. Thus, the nature of the relationship between these policies is not necessarily straightforward. However, for lab learning, learning costs are not a factor. In that case, when adoption is reversible, the rate of AI adoption is not driven by learning and so there is no policy interaction. However, if adoption is irreversible, as noted earlier, the marginal return to first-period adoption as a result of learning is decreasing in d. Therefore, policies that reduce d will increase the amount of learning-driven adoption.

For learning by doing, when AI adoption is reversible, the marginal return to first-period adoption as a result of learning is $-\delta\mu(b-d)$; increasing in d. In this case, reducing d will reduce the return from such learning and hence, lower the rate of AI adoption. When AI adoption is irreversible, that marginal return is $-\delta(b-\mu d+\mu(1-2x_1)(b-d))$. It is decreasing (increasing) in d if x_1 is high (low). Recall that high x_1 is associated with delayed adoption as a result of learning while low x_1 is associated with accelerated adoption as a result of learning. What this demonstrates is that when the policy approach to learning-based AI adoption favours acceleration (delay), this decreases (increases) the return to harm mitigation investments.

This analysis suggests that, in the simple model, harm mitigation and learning-based acceleration of AI adoption are substitute policies for learning by doing and complements for lab learning. More broadly, this analysis indicates that the more a planner intends to invest in mitigating harm, the less it needs to learn to avoid such harm by either adopting AI at an accelerated rate or by waiting for some other resolution of uncertainty.

[22] Catalini & Gans (2020) provide other examples of how blockchain technologies can be used to provide verifiable information integrity.

3.3 Diversity in Research Trajectories

While the theme in this paper has been that learning regarding the potential harms of AI is complicated and that blunt regulatory instruments are unlikely to have a large effect on learning, it has also been noted that the ability to make use of learning involves there being more ways of implementing AI in a safe manner. In other words, regulatory learning is complemented by a greater variety of regulatory options.

If AI harm surfaces, one of the issues that can make it difficult to mitigate that harm is that there are no viable, productive options that a regulator can incentivise ex post. For instance, even within machine learning, AI is being developed on the basis of distinct architectures based on predictive versus generative capabilities, interpretability, bias, automation and the extent of data required. These alternative architectures represented, at times, potentially viable trajectories of AI development, but pressures from both the market and scientist incentives to exceed established benchmarks can cause development efforts to be focussed on relatively few architectures.

This has happened in the economics of technology previously. For instance, Bryan (2017) documents the path dependence that arose from nuclear power choices driven by military criteria rather than those that might have promoted power safety and efficiency. Several models demonstrate that market forces create incentives for too few research architectures to be explored and developed. This includes Acemoglu (2011) and Bryan & Lemus (2017). The former paper shows how the incentives of scientists can be distorted by the salience of rewards towards more recent advances along a technological trajectory.

The implication here is that there may be other architectures associated with AI that may turn out to have lower harmful effects than the ones being pursued prior to those harms being surfaced. The regulatory challenge is that by the time those harms have emerged, the current architecture has progressed so far that it is difficult or impossible to switch to alternative options with lower damage prospects.

To counter this, there is value in policy interventions that are directed at promoting a diverse set of architectures being developed, even though one may be dominant in adoption. The tools to promote this are familiar to the economics of basic science but, broadly speaking, range from funding for basic science to an allocation of those funds to support architectures that are not being developed at present.

Some of these interventions may also be directed at some of the potential equilibrium harms from the adoption of AI. For instance, large language models are trained on the writing of people while, at the same time, themselves reduce incentives for people to provide that writing.[23] To counter this, there may be interventions to generate the data that would allow future LLMs to be trained.

4 Conclusion

This paper has demonstrated that whether AI adoption should be delayed or accelerated from a social perspective is highly sensitive to the nature of assumptions regarding how regulators learn about the adverse consequences of AI adoption. When learning about potential harms from AI requires real-world AI deployment, learning generally favours accelerated adoption, while if learning can be achieved away from real-world deployment, i.e., in the lab, this favours delaying or slowing down AI adoption. Of course, these conclusions are tempered by what policymakers would do in the absence of being able to learn about harm. For instance, an optimistic policymaker facing an irreversible adoption decision may adopt AI with some precaution, while a pessimistic one may be pushed to accelerate adoption in order to facilitate learning. This suggests that policies that may improve learning outcomes or allow more options for policy-makers to respond to what they learn will be important. Interestingly, insurance-like policies such as harm mitigation may reduce rather than increase the returns to learning by doing.

The paper here is a normative one. It discusses what a planner might do in the face of uncertainty regarding AI harm and the ability to learn about that harm through the actual adoption of AI. The motivation was calls for a pause in AI adoption as a precautionary manner. The analysis here implies that this conclusion is sensitive to the way in which policy-makers and others might learn about the harms of AI. When such learning requires doing, this implies that a pause or delay is not necessarily a cautious approach to learning about AI harms and that being more proactive can surface harms more readily. Policymakers need to consider the types of learning outcomes they are trying to encourage and whether real-world adoption may be the most efficient way in which learning, even about harms, can be achieved.

That said, this paper does not consider the operation of more decentralised actors, such as firms and other agents, in this type of environment. There is, indeed, a very rich literature on how social learning impacts the diffusion and adoption of new technologies (Foster & Rosenzweig, 2010). In that regard, the analysis is aimed at addressing what the goal of AI adoption policy should be rather than whether there is a case for intervention per se and so sidesteps the research questions of that literature.[24] Moreover, it considers a single

[23] https://medium.com/@duanevalz/the-paradoxes-of-generative-ai-alignment-48de8135714b

[24] Guerreiro et al. (2023) do, however, examine the difference between private and social incentives to

planner or policy-maker. At best, in today's economy, policy-makers exist at a national level, and, therefore, some of the learning that could arise could come from AI adoption policies pursued elsewhere. This creates a positive spillover in learning that, in most contexts, tends to generate an undersupply of such learning (e.g., King (1995)).

Finally, while the paper here encompasses many potential harms of AI, it has not examined concerns that AI adoption will lead to an increased concentration of power – both market and politically. Acemoglu & Johnson (2023) have put forward the thesis that many technologies have proven to increase power concentration and, in their view, AI will follow their thesis's lines. This is a set of issues where it is not at all clear whether there are learning mechanisms that can play a role in guiding policy towards the impact of technologies on power concentration. Instead, this is likely to require a broader institutional intervention in society to address power inequities rather than something that can be evaluated by a deeper exploration of a technological class – whether in lab or market settings. Hence, this particular potential consequence of AI is outside the scope of the present paper, although addressing these institutional issues may turn out to impact the efficacy of learning about potential AI harm in other domains.

learn about AI harms finding that private incentives are generally lower than social incentives.

References

Acemoglu, D. (2011). Diversity and technological progress. In *The Rate and Direction of Inventive Activity Revisited* (pp. 319–356). University of Chicago Press.

Acemoglu, D. (2021). Harms of ai. In *The Oxford Handbook of AI Governance*: Oxford University Press.

Acemoglu, D. & Johnson, S. (2023). *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. Hachette UK.

Acemoglu, D. & Lensman, T. (2023). *Regulating Transformative Technologies*. Technical report, National Bureau of Economic Research.

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The simple Economics of Artificial Intelligence*. Harvard Business Press.

Agrawal, A., Gans, J., & Goldfarb, A. (2022). *Power and Prediction: The Disruptive Economics of Artificial Intelligence*. Harvard Business Press.

Agrawal, A., Gans, J., & Goldfarb, A. (2023a). Do we want less automation? *Science*, 381(6654), 155–158.

Agrawal, A., Gans, J., & Goldfarb, A. (2024). The turing transformation: artificial intelligence, intelligence augmentation, and skill premiums. *Harvard Data Science Review*.

Agrawal, A., Gans, J., Goldfarb, A., & Tucker, C. (2023b). *The Economics of Artificial Intelligence: Health Care Challenges*. University of Chicago Press.

Arrow, K. J. (1972). *Economic welfare and the allocation of resources for invention*. Springer.

Bryan, K. A. (2017). The perils of path dependence. *Survive and Thrive: Winning Against Strategic Threats to Your Business*, 111.

Bryan, K. A. & Lemus, J. (2017). The direction of innovation. *Journal of Economic Theory*, 172, 247–272.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at work*. Technical report, National Bureau of Economic Research.

Catalini, C. & Gans, J. S. (2020). Some simple economics of the blockchain. *Communications of the ACM*, 63(7), 80–90.

Choi, J. H. & Schwarcz, D. (2023). Ai assistance in legal analysis: An empirical study. *Available at SSRN 4539836*.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Ehrlich, I. & Becker, G. S. (1972). Market insurance, self-insurance, and self-protection. *Journal of Political Economy*, 80(4), 623–648.

Foster, A. D. & Rosenzweig, M. R. (2010). Microeconomics of technology adoption. *Annu. Rev. Econ.*, 2(1), 395–424.

Furman, J. L., Murray, F., & Stern, S. (2012). Growing stem cells: The impact of federal funding policy on the us scientific frontier. *Journal of Policy Analysis and Management*, 31(3), 661–705.

Galasso, A. & Luo, H. (2021). Risk-mitigating technologies: The case of radiation diagnostic devices. *Management Science*, 67(5), 3022–3040.

Galasso, A. & Luo, H. (2022). When does product liability risk chill innovation? evidence from medical implants. *American Economic Journal: Economic Policy*, 14(2), 366–401.

Gans, J. & Leigh, A. (2019). *Innovation+ Equality: How to Create a Future that is More Star Trek Than Terminator*. Mit Press.

Grant, S. & Quiggin, J. (2013). Bounded awareness, heuristics and the precautionary principle. *Journal of Economic Behavior & Organization*, 93, 17–31.

Guerreiro, J., Rebelo, S., & Teles, P. (2023). *Regulating Artificial Intelligence*. Technical report, National Bureau of Economic Research.

Haber, S. & Stornetta, W. S. (1991). *How to time-stamp a digital document*. Springer.

Hadfield, G. K. & Clark, J. (2023). Regulatory markets: The future of ai governance.

Jones, C. (2023). The ai dilemma: Growth versus existential risk.

Kanazawa, K., Kawaguchi, D., Shigeoka, H., & Watanabe, Y. (2022). *AI, Skill, and Productivity: The Case of Taxi Drivers*. Technical report, National Bureau of Economic Research.

King, S. P. (1995). Search with free-riders. *Journal of Economic Behavior & Organization*, 26(2), 253–271.

Noy, S. & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Available at SSRN 4375283*.

Segal, I. & Whinston, M. D. (2007). Antitrust in innovative industries. *American Economic Review*, 97(5), 1703–1730.

Spitzer, M. & Talley, E. (2014). On experimentation and real options in financial regulation. *The Journal of Legal Studies*, 43(S2), S121–S149.