Generative AI, Productivity, and Incentives for AI Twins

Catherine Wu* Arun Sundararajan†

DRAFT: March 2025

Abstract

Generative AI is an example of a technology whose automation potential depends in part on participation by humans in training and improving the technology. Greater investments by an individual could enable higher-quality output in their individual voice or style at lower individual cost. However, increasingly intelligent AI "twins" could also, over time, replace individual humans or lower their ability to command premium wages. We analyze this trade-off using a principal-agent model in which agents have the opportunity to make investments into training an AI twin that lowers their cost of effort, raises their probability of success, or both. Tasks performed vary in the extent to which AI output can be altered or improved by the human ("editability") and also vary in the extent to which a non-expert can assess the quality of human or AI output (its "verifiability.") In a single-period model we show that, paradoxically, a strategic agent will invest in training their AI twin only when this training increases the AI twin's efficacy faster than that of the human and AI working in conjunction, and further, when the agent's productivity gains from partnering with their AI twin do not increase too rapidly with training investment. The strategic agent trades off improvements in output quality and ease of effort to preserve their wage bargaining power. We explain how tasks with high verifiability and low editability are strategically most aligned with a worker's incentives to "train their twin," but that for tasks where the stakes associated with good outcomes are low, this alignment is further constrained by the agent holding back to mitigate their risk of displacement. We extend these results in a two-period model, characterizing the conditions under which non-strategic agents might overinvest into training their AI twins, and show that settings in which training investments yield the highest immediate personal productivity gains also put the agent at the greatest risk of mistakenly creating an AI twin that will displace them in the long run. We discuss extensions in progress that generalize our findings to continuous outcomes and other settings in which a human agent participates in the creation of an automation technology. Our results suggest that sustained improvements in company-specific generative AI will require nuanced design of human incentives, and that public policy which encourages balancing worker returns with generative AI improvements could yield more reliable training data within organizations and more sustained long-run productivity gains.

*NYU Stern School of Business, cw4608@stern.nyu.edu
†NYU Stern School of Business, digitalarun@nyu.edu

1 Introduction

In the early days of computer-based factory automation, a human machine operator's manual actions were recorded, often via punched cards or tapes, and then "played back" to drive machine tools automatically. These "teach-and-repeat" methods laid the foundations for the computer numerical control (CNC) systems that powered an early wave of computer-driven automation (Reintjes, 1991). Today, although generative AI is largely viewed as general-purpose source of knowledge and skills, a distinguishing characteristic of the technology is its ability to replicate the process or style of an *individual human creator*. Put differently, a human's body of work can now serve as training data to create AI that emulates their unique approaches, skills, and creative processes. From replicating an author's writing style to capturing a surgeon's precise movements, these "digital twins" (Sundararajan, 2024) generate the ability to scale human expertise in unprecedented ways, while simultaneously raising new challenges to an individual's ownership of their human capital.

The potential of a new wave of replication-driven displacement extends beyond creative domains and into highly skilled professional fields, where complex individual human capabilities may be systematically captured and partially reproduced by AI systems, reversing the trend of skill-biased technological change that has characterized digital process over the last decades (Acemoglu, 2002, 2021). This is one possible trajectory—in contrast, training an AI system to replicate individual skills may alternatively augment human potential, improving labor market outcomes and reducing inequality (Agrawal et al., 2023). As platforms like Wizly[1] that allow individuals to create "AI twins" that they control and use to scale their independent work proliferate in the coming years, custom-trained generative AI models could augment and enable humans to produce higher-quality output faster and more efficiently. At the same time, workers may view increasingly intelligent tooling as a threat that could replace them or lower their ability to command premium wages.

How will the net effect of these opposing forces affect the incentives for a knowledge worker to participate in the creation of their AI twin? This remains an open question, one that depends on factors that include the relative rate of quality improvements with or without the human in the loop, the impact of individually customized AI systems on a human's cost of effort, and the benefits that an organization enjoys from outcomes of varying levels of quality and success.

While there is a growing body of literature about whether artificial intelligence will substitute for or complement human labor across various sectors (Noy and Zhang, 2023; Felten et al., 2023; Brynjolfsson et al., 2023), research into the incentives a human worker might have to participate in the creation and improvement of such systems is limited. Besides, as things stand, projections about the impact of AI on worker productivity are mixed. A study conducted by The Upwork Research Institute reports that "77% [of employees] say [AI] tools have actually decreased their productivity and added to their workload" (Monahan and Burlacu, 2024). In contrast, a study by McKinsey and Company estimate that generative AI "could add the equivalent of $2.6 trillion to $4.4 trillion annually" to the economy (Chui et al., 2024). Because AI is a general purpose

[1] https://askwizly.ai/

technology, it is valuable to assess how these shifts of productivity will manifest in the long-run in market structure and governance (Tucker, 2024).

In this paper, we model a human worker who can partner with an AI system to accomplish tasks assigned by a manager. The capabilities of the AI system are inferior to those of the human worker, but the worker can choose to invest in training it to be an AI twin in order to either reduce the effort they need to exert, to improve work outcomes when partnering with the AI system, or both. We first consider a short-run problem, investigating how the worker's level of investment in the creation of their AI twin depends on intrinsic task properties—specifically, (i) how easily a human worker can "edit" AI output, which we term the *editability* of the task, and (ii) whether the manager is able to verify the quality of AI output absent the human worker, which we term the task's *verifiability*. As we discuss in Section 2, we choose these properties based on recent findings from the empirical literature connecting the use of generative AI to productivity.

We situate our problem in a two-outcome limited liability moral hazard model with hidden action. The manager (principal) chooses whether or not to offer the capability to customize an AI system to the worker (agent). If offered, the worker chooses a level of training investment in customizing the AI system. A higher level of training investment may lower the worker's cost of effort, may improve the likelihood of better work outcomes, or both. After observing the worker's training investment into customizing the AI system, the manager offers the worker a contract, the worker decides on their effort level, and outcomes and payoffs are realized.

Our results from a single period model show that a strategic worker will invest in training their AI twin only when this training increases the AI twin's efficacy growth *faster* than that of the human and AI working in conjunction, trading off improvements in output quality and ease of effort to preserve their wage bargaining power and mitigate their risk of displacement, and will underinvest in their AI twin in other settings. This growth pattern increases the informational rent a worker commands by lowering *outcome separability*—the extent to which the manager can map observed outcomes to worker effort—and we explain how it is associated with tasks that have a higher level of verifiability.

Even in the presence of this trajectory of AI capability growth, we show that a strategic worker's incentives to improve their AI twin are constrained by two additional factors. Reducing the cost of effort from using the AI twin, which is associated with a greater level of editability, has the additional effect of lowering a worker's wage bargaining power, and thus, higher levels of investment into training one's AI twin are only incentive-aligned for work with relatively low editability. Additionally, even when faced with tasks strategically best suited for AI twin improvements, a worker's training investments are sometimes constrained by the risk that excessive improvements in the AI twin's capability will lead to its performance becoming sufficiently good to cause the worker to be displaced entirely by the technology, a risk we show is greater when the task has low *quality importance*, or when there is no significant level of difference between the payoffs to the manager from good and bad outcomes.

We build on these single-period results to examine, within a two-period model, how the behavior

of non-strategic (or myopic) agents might deviate from those of a strategic agent. We characterize the conditions under which such agents might overinvest into training their AI twins and then get displaced by them, and show that these conditions are more likely to occur in settings where the training investments lead to greater immediate returns that allow the myopic worker to "shirk" in the short term. Thus, the settings in which training investments yield the highest immediate personal productivity gains also put the agent at the greatest risk of mistakenly creating an AI twin that will displace them in the long run. We further show how these risks are mitigated in part when the time-persistence of the AI is lower.

The rest of the paper is organized as follows. We use the existing literature on how AI and especially generative AI affect worker productivity as the basis to provide a framework for choosing two key work dimensions and how they are likely to be related to changes in output when human workers collaborate with AI systems. Next, we embed this framework into a simple two-outcome hidden-action principal-agent model in which, prior to contracting, the principal may choose to offer the agent the aid of an AI system, and the agent may choose to invest into customizing it to be a digital twin. We then consider a two-period extension in which we contrast the choices made by myopic and strategic agents. We conclude by discussing the implications of our results, ongoing work that generalized our model, and directions for future work.

2 Related literature and framework

We briefly summarize some of the recent empirical literature that informs our work. Although our model is also connected to an extensive body of work on contract theory, placing our results in this literature remains work-in-progress.

A rapidly growing body of research has examined generative AI's potential to enhance human productivity. Recent studies have found that generative AI tools increased throughput by an average of 13.8% for customer support workers (Brynjolfsson et al., 2023), 59% for professionals writing business documents (Noy and Zhang, 2023), and 126% for programmers on personal projects (Peng et al., 2023). Dell'Acqua et al. (2023) report that the use of AI improved the output quality of management consultants by nearly 40% and allowed teams to perform tasks 25.1% faster. However, the results are not all positive—Yeverechyahu et al. (2024) report that the release of Github Copilot, a code automation tool, did not significantly increase the number of "originative tasks" completed in open-source software projects.

The empirical evidence also remains mixed regarding AI's differential impact across skill levels. Some studies suggest that AI tools provide greater benefits to highly skilled workers, who can better direct and contextualize AI outputs for entrepreneurial tasks (Otis et al., 2024). Others find that AI may have an equalizing effect by providing greater relative assistance to less experienced workers in the context of customer service (Brynjolfsson et al., 2023). These apparently contradictory findings highlight the need for a more nuanced theoretical framework.

Work that analyzes the risks of worker displacement and job exposure to AI typically involves

some form of task categorization. For example, Eloundou et al. (2023) and Felten et al. (2023) assess tasks based on their similarity to LLM training data or generative abilities. However, it is likely that as the capabilities of AI technology evolve, the attributes of work that are predictive of automation will change. Chain-of-thought reasoning and other recent developments (Wei et al., 2022; Yao et al., 2023; Bubeck et al., 2023) have already expanded LLM capabilities into domains that require more advanced reasoning and creativity than previously considered possible. Furthermore, assessments that focus solely on raw AI output miss the important role of human-AI collaboration. Distinguishing standalone AI capabilities from performance under human-AI partnerships seems especially important in light of the new capabilities of generative AI.

To frame our analytical model, we sought a classification for tasks that was agnostic of the AI technological fit, while also making provisions to incorporate the skill of the human user. As a simple starting point, we have categorized the tasks in work on two dimensions. First, a task is *verifiable* if the correctness of the output can be confirmed by an unskilled operator. For example, it is easy to verify if AI-generated art is realistic or not. By contrast, it is more difficult for an average human to evaluate the correctness of a mathematical proof. Second, a task is *editable* if the output can be edited by an unskilled operator. For example, an AI-drafted letter that gets the addressee's name wrong can easily be edited. In contrast, editing AI-generated art may require custom tools and be time-consuming for even skilled artists.

When we apply this framework to the empirical literature we have surveyed, it appears to suggest that tasks that are more verifiable and editable are more exposed to AI, implying that AI tools built for such tasks will increase productivity or quality. Furthermore, for tasks that are more verifiable, AI will improve the output quality of low-skilled workers, and for tasks that are more editable, AI will improve the productivity of high-skilled workers. Table 1 lists examples of tasks in each category and implications for each task.

Prior work on skill-biased technical change by Autor et al. (2003), has classified tasks as either "routine" or "nonroutine." Much like computers in the past were predicted to replace "routine" tasks, we suggest that it is easier for AI to substitute for workers in "verifiable" tasks. On the other hand, AI will complement workers in more complex, "nonverifiable" tasks. The editability of the task will mediate the trade-off in switching between AI and human workers between high quality and low cost. Independent of the economic modeling that follows, we continue to expand and refine our categorization, including drawing skill-task-ability databases like those used in Felten et al. (2023), but for now, we frame the model that follows through this simple two-dimensional framework.

3 A single period model

A manager (principal) wishes to hire a worker (agent) to perform a task. Both parties are risk neutral. The agent chooses between two effort levels e, high ($e = 1$) and low ($e = 0$). Exerting high effort involves the agent incurring a cost $c > 0$. (We discuss the relationship between cost of effort

Table 1: Task verifiability and editability: a summary of some prior empirical findings

	Easy to verify	Hard to verify
Easy to edit	*Example*: Customer service, writing *Outcome*: Lower-skilled workers saw a greater quality boost compared to high-skilled workers in customer service.	*Example*: Entrepreneurship advice, financial modeling *Outcome*: High-skilled workers saw quality improvement when provided AI advice for entrepreneurship, whereas low-skilled workers saw worse performance.
Hard to edit	*Example*: Software maintenance, art *Outcome*: Fast adoption among unskilled workers (i.e. non-artists) at the risk of professional artists.	*Example*: Software feature development, mathematical proofs *Outcome*: No increased improvement observed for unskilled open-source contributors.

Source: Yeverechyahu et al. (2024); Brynjolfsson et al. (2023); Otis et al. (2024)

and the use of AI a little later in this section.) The principal's payoff is shaped by outcomes that are realized from the agent's exertion of effort.

We follow the standard approach to modeling this setting of moral hazard. If the principal could observe the agent's effort, either directly or by inferring it perfectly from observable output, they would pay the agent conditional on the effort observed (the "first-best" setting). When the principal cannot observe the agent's effort, the observed work outcome depends probabilistically on the agent's effort and the principal can only pay the agent based on outcome rather than on effort (the "second-best" setting).

We assume two possible outcomes that, respectively, provide benefit $\bar{S}$ and $\underline{S}$ to the principal, where $\bar{S} > \underline{S}$. We designate π_1 as the probability that high effort produces the better outcome, and π_0 as the probability that low effort produces the better outcome. We reinterpret these probabilities and discuss how they are altered by the use of AI a little later in this section. We also discuss an ongoing generalization to multiple or continuous outcomes in our concluding section. The agent's outside option is set to 0.[2]

Since the principal can observe the agent's outcome, but not their effort, the contract offered by the principal will depend on the outcome observed. Denote by $\bar{t}$, $\underline{t}$ the payments for the better ($\bar{S}$) and worse ($\underline{S}$) outcomes respectively. We analyze contracts where $\bar{t} > \underline{t}$, since if $\bar{t} \leq \underline{t}$, the agent will choose effort $e = 0$.

Additionally, the principal can choose to offer the agent a customizable generative AI system that we refer to in what follows as a *trainable AI twin*. The baseline level of capability of the

[2]In the language of contract theory, the agent has "limited liability," meaning they cannot pay more than L to the principal. To simplify the analysis, we assume that $L = 0$.

trainable AI twin can be improved by the agent. These improvements involve an investment $v \geq 0$ on the part of the agent, where $v = 0$ reflects the status quo, or the absence of an AI twin.

For simplicity and to focus our analysis on the strategic aspects of the agent's choices, we assume that this "training investment" v involves no direct cost to the agent or the principal. Examples of this kind of investment could involve the agent providing examples of their writing style to fine-tune an LLM, the agent annotating data about their decisions and communications to improve an AI system's understanding of them, or the agent participating in rating the quality of a system's output in response to their prompts for use in reinforcement learning with human feedback (RLHF).

The choice of v may alter the agent's cost of effort, now denoted $c(v) > 0$, as well as the relative likelihoods of better and worse outcomes, now denoted $\pi_0(v)$ and $\pi_1(v)$ respectively. A larger training investment v weakly increases the likelihood of better outcomes, and weakly lowers the agent's cost of exerting effort once the AI twin is trained, and thus, $\pi_1'(v) \geq 0$, $\pi_0'(v) \geq 0$, $c'(v) \leq 0$. We return to interpretations of these rates of change, whose relative values are central to our results, later in this paper.

One can now also think of $\pi_0(v)$ as characterizing the "AI only" setting, or the outcome distribution induced by AI without additional human effort, and $\pi_1(v)$ as characterizing the "human+AI" setting, or the outcome distribution induced when human effort complements the AI twin. Importantly, we are considering a setting in which a higher level of training investment in the customizable AI twin can also improve *the standalone performance* of the AI twin.

It seems natural to assume that the standalone AI performance is always weakly inferior to the setting where the human exerts high effort and collaborates with the AI, or that $\pi_1(v) > \pi_0(v)$ for all v. Let there be some $v_{\max} > 0$ that bounds how much an agent can invest in their AI twin. This limit could be the result of a predetermined model size: the performance and sample efficiency of a language model scales with the size of the model (Kaplan et al., 2020), and the size of the training data corpus is limited by the agent's productive capacity.

The principal provides the AI twin if, taking into account the anticipated training investment v by the agent and the ensuing choices in effort by the agent, the presence of the AI twin increases the expected surplus of the principal.

Before proceeding further, we note that since $\pi_1'(v) > 0$, $\pi_0'(v) > 0$ and $c'(v) < 0$, the socially optimal level of training investment is $v = v_{\max}$.

Lemma 1 *The socially optimal level of training investment is* $v = v_{\max}$.

3.1 Optimal contract

For a fixed level of training investment v, the principal will choose a menu of payments $\{\bar{t}(v), \underline{t}(v)\}$ to maximize their surplus, subject to the usual constraints we describe in what follows:

$$\max_{\{\bar{t}(v),\underline{t}(v)\}} \pi_1(\bar{S} - \bar{t}(v)) + (1 - \pi_1(v))(\underline{S} - \underline{t}(v)). \tag{1}$$

In order to induce both participation and high effort from the agent, the principal's payments must satisfy the agent's participation and incentive compatibility constraints, which are, respectively, equations (2) and (3)

$$\pi_1(v)\bar{t}(v) + (1-\pi_1(v))\underline{t}(v) - c(v) \geq 0, \tag{2}$$

$$\pi_1\bar{t}(v) + (1-\pi_1(v))\underline{t}(v) - c \geq \pi_0(v)\bar{t}(v) + (1-\pi_1(v))\underline{t}(v). \tag{3}$$

The optimal contract follows from standard textbook contract theory dealing with hidden action settings that have two effort levels and two outcomes:

Lemma 2 *Given a fixed level of investment v by the agent, the optimal contract offered by the principal is $\bar{t}(v) = \frac{c(v)}{\pi_1(v)-\pi_0(v)}, \underline{t}(v) = 0$.*

(A proof of this lemma is available on request.)

Under this contract, the principal has an expected surplus of

$$\pi_1(\bar{S} - \frac{c(v)}{\pi_1(v)-\pi_0(v)}) + (1-\pi_1(v))\underline{S}, \tag{4}$$

or of

$$\pi_1(v)\bar{S} + (1-\pi_1(v))\underline{S} - \frac{\pi_1(v)}{\pi_1(v)-\pi_0(v)}c(v). \tag{5}$$

The principal thus offers a contract to induce a high level of effort if the incremental expected payoff from a higher effort level versus a lower effort level is at least as high as the expected payment to the agent, or if

$$(\pi_1(v)-\pi_0(v))(\bar{S}-\underline{S}) \geq \pi_1(v)\frac{c(v)}{\pi_1(v)-\pi_0(v)}. \tag{6}$$

The agent's expected surplus is

$$U(v) = \frac{\pi_1(v)}{\pi_1(v)-\pi_0(v)}c(v) - c(v) = \frac{\pi_0(v)}{\pi_1(v)-\pi_0(v)}c(v). \tag{7}$$

We focus on settings for which in the absence of an AI twin, the principal wishes to offer a contract to induce high agent effort, or that

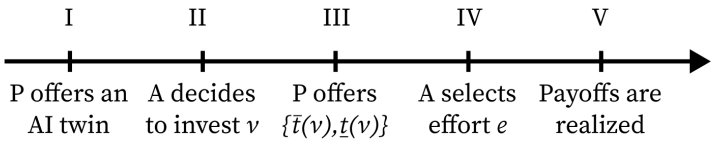
$$(\pi_1(0)-\pi_0(0))(\bar{S}-\underline{S}) \geq \pi_1(0)\frac{c(0)}{\pi_1(0)-\pi_0(0)}. \tag{8}$$

Put differently, we imagine this baseline as the pre-AI status quo, and assume that in this status quo, there are gains from contracting between the principal and the agent that can be implemented.

3.2 AI twin: Incentives for both parties

We now consider the agent's incentives for making a training investment into their AI twin and the principal's incentives for offering this AI twin. The sequence of events is as follows.

(I) The principal decides whether or not to offer the AI twin.



(I) The principal decides whether or not to offer the AI twin.
(II) If the AI twin is offered, the agent decides on their training investment v.
(III) The principal either decides to use the AI twin without the agent, or announces the contract $\{\bar{t}(v),\underline{t}(v)\}$.
(IV) If a contract is announced (i.e. the agent is not fired), the agent chooses their effort level.
(V) Payoffs are realized.

Figure 1: Timing in the basic model

(II) If the AI twin is offered, the agent then decides on their training investment v.

(III) The principal then either decides to use the AI twin without the agent, or announces the contract $\bar{t}(v), \underline{t}(v)$.

(IV) If a contract is announced (that is, the agent is not fired), the agent chooses their effort level.

(V) Payoffs are realized. (See Figure 1.)

Given this sequence of events, assuming step (IV) is reached, the contract structure will follow Lemma 2:

$$\bar{t}(v) = \frac{c(v)}{\pi_1(v) - \pi_0(v)}, \underline{t}(v) = 0, \tag{9}$$

and following (7), the agent's expected surplus assuming step (IV) is reached is $U(v)$.

Now, assuming step (III) is reached, the principal will offer a contract only if, anticipating inducing a high effort level from the agent in step (IV), their expected surplus from doing so is higher than their expected surplus from using the AI twin imbued with the agent's training investment v but without the agent involved in exerting effort, or if:

$$\pi_1(v)(\bar{S} - \bar{t}(v)) + (1 - \pi_1(v))(\underline{S} - \underline{t}(v)) \geq \pi_0(v)\bar{S} + (1 - \pi_1(v))\underline{S}. \tag{10}$$

Note that (10) above also specifies a constraint on the agent's level of training investment. A strategic agent will not choose a training investment level that reverses the inequality in (10). This is because a strategic agent does not want to be displaced by their AI twin and lose the positive expected surplus associated with exerting high effort. For this reason, in what follows, we sometimes refer to (10) as the "displacement deterrent" condition.

Finally, in step (I), the principal offers the AI twin if doing so increases the principal's surplus, anticipating offering a contract and inducing a high level of effort from the agent:

$$\pi_1(v)(\bar{S} - \bar{t}(v)) + (1 - \pi_1(v))(\underline{S} - \underline{t}(v)) \geq \pi_1(0)(\bar{S} - \bar{t}(0)) + (1 - \pi_1(0))(\underline{S} - \underline{t}(0)). \tag{11}$$

3.3 AI twin: Agent's training investment

With these preliminaries out of the way, we are now ready to analyze the agent's choice of training investment in step (II). The agent chooses a training investment level that solves:

$$\max_{v\in[0,v_{\max}]} U(v) = \frac{\pi_0(v)}{\pi_1(v)-\pi_0(v)}c(v) \tag{12}$$

subject to (10). The objective function $U(v)$ in (12) is not necessarily strictly quasiconcave in v. Rather than imposing additional structure through assumptions that would ensure an interior solution, we instead focus on deriving and interpreting conditions under which (12) has a well-defined solution, which may either be an interior solution or one of the end points.

3.3.1 No displacement deterrent

We begin by ignoring the displacement deterrent condition, which we consider in the next subsection, and focus on the conditions that lead the agent to choose not to invest in their AI twin, to choose the maximum training investment, and to choose an intermediate level of investment. Define $Q_\pi(v)$, the *outcome separability* of the work in question, as:

$$Q_\pi(v) = \frac{\pi_1(v)}{\pi_0(v)}. \tag{13}$$

We call $Q_\pi(v)$ outcome separability because it measures the confidence that the principal has, after observing the realized outcome, in their assessment of the agent's effort level. Also note that

$$Q_\pi(v) - 1 = \frac{\pi_1(v)-\pi_0(v)}{\pi_0(v)}, \tag{14}$$

and thus,

$$U(v) = \frac{c(v)}{Q_\pi(v)-1}. \tag{15}$$

As (15) suggests, and the following results show, the optimal level of training investment v made by the agent will depend on the relative rate at which the training investment alters outcome separability and the agent's cost of effort.

Proposition 1 *If, for all* $v \in [0, v_{\max}]$,

$$\frac{c'(v)}{c(v)} < \frac{Q'_\pi(v)}{Q_\pi(v)-1} \tag{16}$$

then the agent's optimal training investment is $v = 0$.

Proofs are available in the Appendix.

Since $\pi_1(v) - \pi_0(v) > 0$, it follows that $Q_\pi(v) - 1 > 0$, and since $c'(v) \leq 0$, a natural corollary is that when $Q'_\pi(v) > 0$ for all v, the agent will choose $v = 0$.

Next, we characterize when the agent chooses the maximum training investment.

Proposition 2 *If, for all* $v \in [0, v_{\max}]$, $Q'_\pi(v) \leq 0$ *and*

$$\left|\frac{c'(v)}{c(v)}\right| < \left|\frac{Q'_\pi(v)}{Q_\pi(v) - 1}\right| \tag{17}$$

then the agent's optimal training investment is $v = v_{\max}$.

Before characterizing a set of sufficient conditions for an optimal interior training investment choice $v \in (0, v_{\max})$, it is helpful to interpret Proposition 1 and Proposition 2 based on the framework we introduced in Section 2.

Note that in our setting, the agent is the functional "expert" – when presented with output from the AI twin, they are able to assess its quality and, if needed, make changes. One can thus think of the rate at which the cost of effort goes down with training investment, or the absolute value $c'(v)$ as being related to the level of *editability* of the AI twin's output – a greater reduction in the cost of high effort is associated with AI output that the worker can modify to increase the likelihood of a good outcome.

In contrast, the rate at which outcome separability changes, or the value of $Q'_\pi(v)$ is related to the level of *verifiability* of the work in question. If the agent's training investment increases the rate at which standalone AI outcomes improve slower than it increases the rate at which to "human+AI" outcomes, or if $\pi'_1(v) > \pi'_0(v)$, which in turn implies that $Q'_\pi(v) > 0$, this corresponds to a situation in which, absent the human worker, the principal has a relatively lower level of confidence that the AI twin's output will lead to high quality. The verifiability associated with the work in question is therefore related to the rate of change in outcome separability: Outcome separability increasing with an increase in training investment corresponds to low verifiability, and correspondingly, a sufficiently negative value of $Q'_\pi(v)$, one that leads to $\pi'_1(v) < \pi'_0(v)$ corresponding to a higher level of verifiability.

Through this lens, it is clear that the agent's incentives to make training investments are shaped by them trading off their desire to increase the likelihood of a higher payoff (which incentivizes a higher v) against their desire to preserve their ability to command higher wages by maintaining a sufficiently high cost of effort and by lowering outcome separability (which goes up if $Q'_\pi(v) > 0$). The results of Proposition 2 show that the relative rate at which outcome separability decreases has to be sufficiently larger than the corresponding gains from a lower cost of effort for the agent to be willing to make a training investment. Thus, the incentives for the agent to make training investments are greatest for work in which both editability is sufficiently low and verifiability is sufficiently high. For example, an investment banking analyst, if given an AI system that can automate steps in their valuation or other financial modeling, is more likely to invest in improving this system if they are still needed to assess and vouch for whether the model is accurate.

One might intuitively think that if training investments lower an agent's cost of effort sufficiently, the agent would embrace the prospect of an AI twin and actively make these investments.

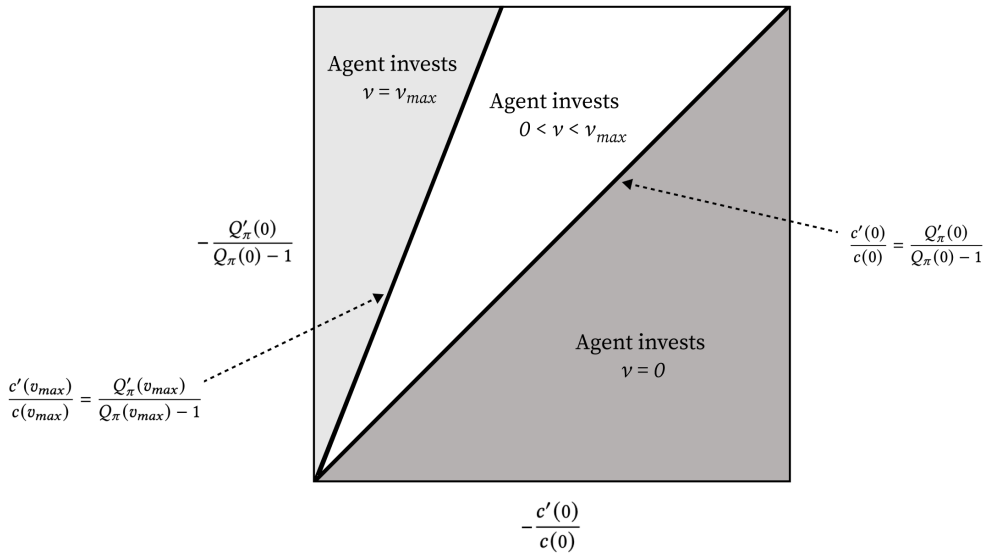


Figure 2: An illustration of optimal training investment regions as $c(v)$ and $Q_\pi(v)$ vary.

However, these investments simultaneously lower the agent's ability to command a higher wage, and as we show in Section 4, a non-strategic agent who makes this kind of investment finds their earning potential harmed in future periods even if they are not entirely displaced by the AI twin. Paradoxically, therefore, when the AI twin can ease the burden on the human worker more, the worker has less incentive to improve the AI twin, and this further bounds the competing trade-off of maintaining "unobservability" of sorts by keeping outcome separability relatively high.

Proposition 3 *A set of sufficient conditions for the agent to choose an interior value of* $v \in (0, v_{\max})$ *are:*

$$\frac{c'(0)}{c(0)} > \frac{Q'_\pi(0)}{Q_\pi(0) - 1}, \tag{18}$$

and

$$\frac{c'(v_{\max})}{c(v_{\max})} < \frac{Q'_\pi(v_{\max})}{Q_\pi(v_{\max}) - 1}. \tag{19}$$

A more restrictive sufficient condition is that for all $v \in [0, v_{\max}]$, $Q'_\pi(v) < 0$, $Q''_\pi(v) > 0$ and $Q''_\pi(v) > c''(v)$, which would imply the conditions of Proposition 3. An illustration of the different regions of the parameter space for which the conditions of our three propositions hold is provided in Figure 2. Note that since the agent always invests $v = 0$ when $Q'_\pi(v) > 0$, outcomes are illustrated for $Q'_\pi(v) \leq 0$.

3.3.2 Displacement as a deterrent to making a training investment

We now return to the possibility that the displacement deterrent condition (10) leads the agent to invest below the levels specified by Propositions 2 and 3.

Although $\bar{t}(v)$ is not a primitive of our model but a derived function, its properties (and, in particular, whether it is increasing or decreasing) are a convenient way to characterize the results about when an agent is threatened by or actually succumbs to displacement by their AI twin. In what follows, the issue of replacement depends directly on the difference in the rates at which $\pi_1(v)$ and $\pi_0(v)$ increase (rather than on their ratio $Q_\pi(v)$), and so we define

$$\Delta\pi(v) = \pi_1(v) - \pi_1(v). \tag{20}$$

From Lemma 2, it follows that

$$\bar{t}(v) = \frac{c(v)}{\Delta\pi(v)}. \tag{21}$$

The next lemma follows immediately from (21) and the fact that $c'(v) < 0$:

Lemma 3 *The following conditions partially characterize when $\bar{t}(v)$ is increasing or decreasing:*

- *If $\Delta\pi'(v) > 0$, then $\bar{t}'(v) < 0$.*
- *$\bar{t}'(v) > 0$ only if $\Delta\pi(v)$ is decreasing and decreases sufficiently faster than $c(v)$.*

Furthermore, $\bar{t}'(v) > 0$ is a sufficient condition for Proposition 2:

Lemma 4 *If $\bar{t}'(v) > 0$, then $Q'_\pi(v) < 0$ and $\left|\frac{c'(v)}{c(v)}\right| < \left|\frac{Q'_\pi(v)}{Q_\pi(v)-1}\right|$.*

Now, using (9) and (13), the displacement deterrent condition can be rearranged as follows:

$$(\bar{S} - \underline{S})\left(1 - \frac{1}{Q_\pi(v)}\right) \geq \bar{t}(v). \tag{22}$$

We refer to $(\bar{S} - \underline{S})$ as the *quality importance* of the work being done, since it measures the difference in payoff to the principal from high-quality and low-quality output.

Since $0 < 1 - \frac{1}{Q_\pi(v)} < 1$, the left-hand side of (22) can never be greater than $\bar{S} - \underline{S}$. Thus, if $\bar{t}'(v) \geq 0$ and $\bar{t}(v_{\max}) > (\bar{S} - \underline{S})$, then there will be some threshold value $v = v^*$ implicitly defined by

$$(\bar{S} - \underline{S})\left(1 - \frac{1}{Q_\pi(v^*)}\right) = \bar{t}(v^*) \tag{23}$$

such that the human will be displaced by their AI twin for any training investment $v > v^*$. We refer to v^* as the *displacement threshold*.

When is it likely that this displacement threshold will constrain an agent's training investment? Recall from the corollary to Proposition 1 that if $Q'_\pi(v) > 0$, the agent makes no training investment. It is easy to verify that $Q'_\pi(v) > 0$ implies $\Delta\pi'(v) > 0$, and $\Delta\pi'(v) < 0$ implies that $Q'_\pi(v) < 0$. Thus, the more relevant case for us to examine is the second setting of Lemma 3, when $\bar{t}'(v) > 0$, since the setting where $\bar{t}'(v) < 0$ is more likely to be associated with the agent making

no training investment anyway, which would occur, for example, when the work is verifiable but not editable.

When $\bar{t}'(v) > 0$, the left-hand side of (22) is more likely to cause the equation to bind when the quality importance $(\bar{S} - \underline{S})$ of the work is lower. For example, quality matters much less for a meme on a blog post than for a painting being auctioned at Christie's. Quality importance also grows with how harmful the bad outcome is—for example, relying entirely on a robotic surgeon that underperforms a human surgeon even slightly may only lower the likelihood of success a little, but the consequences of a bad outcome are substantial. Similarly, a digital programming twin that is empowered to autonomously edit the codebase of an entire application could introduce bugs that, if not detected, have severe damaging effects.

Put differently, the risk of displacement is inversely proportional to how important it is to perform the task well. Loosely speaking, when the stakes are lower, our measure of quality importance $(\bar{S} - \underline{S})$ is lower, making hitting the displacement threshold more likely. Current cases of generative AI-enabled automation seem to support this. Art, particularly as used in background designs, is one class of jobs in which replacement is occurring at a rapid pace (Demirci et al., 2025). Using current image generation tools, a marketing manager can quickly create an emotionally appealing image, or a blogger can use an LLM create a vaguely inspirational quote to break up walls of text. Such low-stakes tasks do not need the level of precision that a skilled worker can contribute, and thus are commonly delegated to AI. Many other types of work that are listed on freelance websites may be considered low-stakes by the lister, making the workers on these marketplaces especially susceptible to AI displacement. Conversely, this inequality also predicts that humans engaged in high-stakes tasks face a lower risk of displacement and, assuming it is otherwise incentive-aligned, can invest more readily in their AI twins.

Of course, our analysis thus far assumes that agents are sufficiently strategic to anticipate the consequences of their training investments. However, it is possible that today's workers are making training investments to reduce their cost of effort or improve outcomes without fully considering the future consequences of their present-day choices. We explore some aspects of this scenario in a two-period extension of our model.

4 Two-period model with a myopic agent

In this section, we consider a two-period extension, contrasting the behavior of a myopic agent who does not internalize the future consequences of their training investment choices with that of the single-period strategic agent. We also explore extensions which assess the effects of how time-persistent the learning of the AI twin is, and the possibility that the use of the AI twin following each instance of training is for multiple periods.

The specific setting we model is summarized in the timeline of Figure 3. In the first period, the principal offers a contract. After this contract is offered, the principal can offer the AI twin to the agent, but cannot change the contract. The agent then chooses their training investment level v.

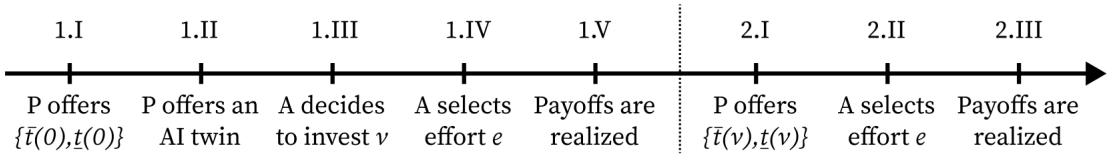


Figure 3: Timing of two-period model

After that, the agent chooses their effort level, and first-period payoffs are realized.

In the second period, the principal, taking into account the training investment made by the agent, decides whether to offer a new contract to the agent, or to simply use the AI twin instead. If the contract is offered, the agent cannot alter their training investment (the AI twin has already been trained), and only decides whether to accept or reject the contract. If the contract is accepted, the agent chooses an effort level, and payoffs are realized. We begin with a baseline two-period model, and then briefly examine the implications of the AI "degrading" over time, and of the second period by itself comprising an infinite horizon of periods.

Denote the training investment made by the myopic agent as v^M. Since the agent is myopic, they only consider their first period payoff when deciding their investment level. Recall as well that the agent now chooses their training investment after the principal proposes a contract but before they decide their effort. Thus, it is possible for the training investment made by the agent following the contract being offered by the principal to cause the agent to choose to "shirk," or to choose the lower rather than higher effort level the principal intended to induce.

We first establish that independent of their subsequent period 1 choice of effort, the agent will always choose the maximum training investment. Consider any contract $\{\bar{t}, \underline{t}\}$ offered by the principal in period 1. If the agent intends to exert the higher effort level, then:

$$v^M = \arg \max_{v \in [0, v_{\max}]} \pi_1(v)\bar{t} + (1 - \pi_1(v))\underline{t} - c(v). \tag{24}$$

Since $\pi_1'(v) > 0$ and $c'(v) < 0$, the objective function in (24) is increasing in v and thus, $v^M = v_{\max}$. Similarly, if the agent intends to exert the lower effort level, then

$$v^M = \arg \max_{v \in [0, v_{\max}]} \pi_0(v)\bar{t} + (1 - \pi_0(v))\underline{t}, \tag{25}$$

and since $\pi_0'(v) > 0$, the objective function in (25) is increasing in v and thus, $v^M = v_{\max}$. Since the principal's offer of a contract is fixed before the agent chooses to invest, the myopic agent's training investment is independent of the principal's offer.

When will the agent's training investment in their AI twin cause them to "shirk"? Comparing (24) and (25), this occurs when

$$\pi_1(v_{\max})(\bar{t} - \underline{t}) - c(v_{\max}) < \pi_0(v_{\max})(\bar{t} - \underline{t}). \tag{26}$$

Rearranging (26), this condition reduces to:

$$\bar{t} - \underline{t} < \frac{c(v_{\max})}{\pi_1(v_{\max}) - \pi_0(v_{\max})}. \tag{27}$$

Following Lemma 2, it follows that $\underline{t} = 0$. Thus (27) implies that we have proved the following:

Lemma 5 *The myopic agent will "shirk" in period 1 after making the maximum training investment if the principal's offer is less than the optimal contract for the current level of investment, or if*

$$\bar{t} < \bar{t}(v_{\max}). \tag{28}$$

Next, we characterize when the myopic agent will get displaced by their AI twin. This follows directly from the definition of the displacement threshold defined in (23).

Proposition 4 *If the displacement threshold $v^* < v_{\max}$, then the myopic agent gets displaced by their AI twin in period 2.*

With this shirking condition in place, the principal must now decide between incentivizing the agent to not shirk by offering a contract at least as good as $\bar{t}(v_{\max})$, or choosing to induce the lower effort level. The principal chooses the former if:

$$\pi_1(v_{\max})(\bar{S} - \bar{t}(v_{\max})) + (1 - \pi_1(v_{\max}))(\underline{S} - \underline{t}(v_{\max})) > \pi_0(v_{\max})\bar{S} - (1 - \pi_0(v_{\max}))\underline{S}, \tag{29}$$

which is exactly the displacement deterrent condition (22). Thus, it follows that both the first and second period contracts offered by the principal are determined by the contract derived in Lemma 2, under certain conditions:

Lemma 6 *In the two-period model outlined above*

- *If the displacement threshold $v^* < v_{\max}$, then the optimal contract for the principal in both periods is $\{0, 0\}$.*
- *If the displacement threshold does not exist or $v^* \geq v_{\max}$, the optimal contract for the principal in both periods is $\{\bar{t}(v_{\max}), \underline{t}(v_{\max})\}$.*

The intuition for this result is simple: if the principal anticipates replacing the myopic agent after the first period, the same condition that makes this displacement optimal for the principal also makes it optimal to not induce a higher effort level in period 1.

Notice that the condition that specifies when the agent will "shirk" in period 1 in Lemma 6 is satisfied if $\bar{t}'(v) > 0$, and recall from Section 3 that $\bar{t}'(v) > 0$ was also associated with a higher likelihood of a displacement threshold $v^* < v_{\max}$. This connection seems natural—if an agent decides that it is easier to use the AI twin and put in low effort in period 1, it stands to reason that a principal would reach the same conclusion in period 2. When tasks that are more verifiable, less editable and have a lower quality importance, this risk is especially significant. Ironically, therefore,

the settings in which training investments yield the highest immediate personal productivity gains also put the agent at the greatest risk of mistakenly creating an AI twin that will displace them.

4.1 Extensions: AI time persistence and a longer post-training horizon

We briefly discuss a couple of simple extensions. The first one considers the possibility that he AI twin's performance will degrade over time without additional training inputs, for example, because of new knowledge, shifting market conditions, changes in customer needs or the need for the AI to know about new products. Define $0 < \alpha < 1$ to be the time persistence of the AI twin. If the principal chooses to displace the agent with the AI twin in the second period, then absent the continued human input (or owing to, for example, changes in the business environment) the performance of the AI twin degrades, and the AI twin performs at the level associated with a training investment αv. Consequently, the displacement deterrent condition in the second period changes accordingly:

$$\pi_1(v_{\max})(\bar{S} - \bar{t}(v_{\max})) - (1 - \pi_1(v_{\max}))(\underline{S} - \underline{t}(v_{\max})) > \pi_0(\alpha v_{\max})\bar{S} - (1 - \pi_0(\alpha v_{\max}))\underline{S}. \quad (30)$$

The new displacement deterrent condition implies a new displacement level $v^{\dagger}(\alpha) > v^*$ under degradation. The principal will choose to employ the agent in period two if $v^{\dagger}(\alpha) > v_{\max}$. Although the agent may have chosen to shirk in period 1, in the scenario with AI twin degradation, if the degradation is large enough, condition (30) may be satisfied, and the principal may continue to employ the myopic agent in the second period. A lower time persistence thus leads to fragility in the AI twin that mitigates the risk faced by myopic agents who make poor training investment choices. This lends a note of caution to organizations that may opt hastily to fire their human workers only to regret it later when they face a degradation in the quality of their model outputs.

Finally, we consider a setting where the second period is not a single period, but is a sequence of periods with future payoffs discounted at $\delta < 1$. In each of these periods, the principal can choose to offer a contract or not. As discussed before, if $v^* > v_{\max}$, the principal will always hire the agent. Assuming this condition is not met, it is easy to see that if this second period consists of sequence of n periods, the displacement deterrence condition becomes:

$$\pi_1(v_{\max})(\bar{S} - \bar{t}(v_{\max})) - (1 - \pi_1(v_{\max}))(\underline{S} - \underline{t}(v_{\max})) < \pi_0(\alpha^n v_{\max})\bar{S} - (1 - \pi_0(\alpha^n v_{\max}))\underline{S}. \quad (31)$$

As $n \to \infty$, $\alpha^n v_{\max} \to 0$. Thus, as the number of periods grows, at some point the performance of the AI twin would have degraded sufficiently to warrant rehiring the agent. Thus, one would expect a cyclical pattern where the principal hires the agent for one period to re-train the model, and then uses the AI twin for m periods, and the cycle repeats.

5 Ongoing work and conclusions

The results and framework presented in this paper offer a new lens through which individual investments into improving generative AI systems might be viewed. While our model has been chosen to be as parsimonious as is possible while still demonstrating the key economic tradeoffs, our framework does not make assumptions about the nature of the AI implementation or work it supports. As generative AI continues to evolve beyond current large language models and towards more sophisticated multimodal systems and advanced reasoning capabilities, we believe our simple framework will remain relevant.

We show that the tradeoffs associated with a strategic human's incentives to actively contribute to the creation of better generative AI systems are real—even when one's AI twin enables the production or higher-quality output or a reduction in effort, incentives to improve this twin are often constrained by the risk of replacement or the potential lowering of earnings. While counterintuitive at first glance, the finding—that training an AI twin is incentive-aligned only when the stand-alone gains grow faster rather than slower than those when the human and AI working in conjunction—makes sense when one considers a human's desire to preserve their wage bargaining power.

Further, our finding that the strategic imperative a human might have to contribute to the performance of the AI twin are lower when one's individual productivity gains are higher highlights the importance of careful incentive alignment when aiming to induce human workers to contribute to the improvement and customization of generative AI systems. As our two-period model shows, in the short run, workers may mistakenly embrace and overinvest in these training efforts, but once it becomes clear that those investments that yield the highest immediate personal productivity gains also put one at the greatest risk of mistakenly creating an AI twin that will displace them in the long run, this behavior will change.

Our ongoing work is focused on extending and generalizing our model while also refining its underlying framework. Our core results generalize naturally to a setting involving continuous outcomes where more training leads to an outcome distribution that is better in the sense of first-order stochastic dominance. However, this extension also allows us to define the nature of the work and the technology in a more nuanced manner by expanding the ways in which we can differentiate the distribution of outcomes of the stand-alone AI twin from the human-AI combination, considering complementarity versus substitution and also allowing a richer specification of verifiability that separates the stand-alone AI twin's outcome distribution from that of the human-AI combination even when the human exerts low effort. **These extensions are in progress and we expect to have them completed by June 2025.**

Reinterpreted appropriately, our results could also suggest that effective AI tools will incorporate features that make outputs more transparent and assessable by non-experts, while also providing robust editing capabilities that allow for meaningful human intervention, and thus a shift away from black-box solutions toward more interpretable and malleable AI systems. Strengthening this kind of interpretation of our analytical results will require an approach that unpacks the connection between the tasks comprising the work being done and the capabilities of the supporting AI, and

models them separately.

The empirical literature connecting generative AI to productivity changes is growing rapidly. As we encounter new results, especially those from the social sciences that deepen our understanding of the nature of interaction between generative AI and human workers, it is possible that other aspects of work beyond editability and verifiability could be relevant.

As the returns from improvements to general-purpose generative AI systems are increasingly constrained by the plateauing of "scaling laws," by the unavailability of increasingly larger training data sets, and by the specter of "model collapse," engaging human workers in the improvement of AI systems will be increasingly important for organizations to continue to realize productivity gains from generative AI. The realization that the recent performance improvements of DeepSeek over OpenAI's systems stem from not just the use of more sophisticted reasoning but from a heavier reliance on RLHF only underscore this importance.

Preserving and sustaining improvements in company-specific generative AI will likely require nuanced design of human incentives. The viability of contracting that align a human's returns from contributing to improvements in AI through, for example, fractional IP ownership, may be countered by the associated transaction costs and challenges in attribution, but this nevertheless represents a direction for future inquiry. Policy interventions that better align an organization's incentives to balance worker returns with generative AI improvements could yield more reliable training inputs and more sustained long-run productivity gains for a country.

A Proofs

A.1 Proof of Proposition 1

PROPOSITION. If, for all $v \in [0, v_{\max}]$,

$$\frac{c'(v)}{c(v)} < \frac{Q'_\pi(v)}{Q_\pi(v) - 1} \tag{32}$$

then the agent's optimal training investment is $v = 0$. **Proof.** If $U'(v) < 0$ for all $v \in [0, v_{\max}]$, then $U(v)$ is maximized at $v = 0$. Thus, we want to find conditions under which $U'(v) < 0$. Recall that we defined $U(v) = \frac{c(v)}{Q_\pi(v)-1}$. Therefore,

$$U'(v) = \frac{c'(v)(Q_\pi(v) - 1) - c(v)Q'_\pi(v)}{(Q_\pi(v) - 1)^2}.$$

Now, recall that $Q_\pi(v) = \frac{\pi_1(v)}{\pi_0(v)}$, and by assumption, $\pi_1(v) > \pi_0(v)$ for all $v \in [0, v_{\max}]$. Thus we know that if $Q_\pi(v) > 1$, and it follows that $Q_\pi(v) - 1 > 0$.

Thus, the denominator of $U'(v)$ is always strictly positive, and thus, the inequality $U'(v) < 0$ is equivalent to

$$c'(v)(Q_\pi(v) - 1) - c(v)Q'_\pi(v) < 0.$$

By assumption, $c(v) > 0$ for all v, so we can arrange the inequality to

$$\frac{c'(v)}{c(v)} < \frac{Q'_\pi(v)}{Q_\pi(v) - 1}.$$

The result follows. ■

A.2 Proof of Proposition 2

PROPOSITION. If, for all $v \in [0, v_{\max}]$, $Q'_\pi(v) \leq 0$ and

$$\left|\frac{c'(v)}{c(v)}\right| < \left|\frac{Q'_\pi(v)}{Q_\pi(v) - 1}\right| \tag{33}$$

then the agent's optimal training investment is $v = v_{\max}$.

Proof. If $U'(v) > 0$ for all $v \in [0, v_{\max}]$, then $U(v)$ is maximized a $v = v_{\max}$. Thus, we want to find conditions in which $U'(v) > 0$ for all v. As above, we take the derivative of $U(v)$ with respect to v,

$$U'(v) = \frac{c'(v)(Q_\pi(v) - 1) - c(v)Q'_\pi(v)}{(Q_\pi(v) - 1)^2}.$$

Since the denominator of $U'(v)$ is strictly positive as established in the previous proof, the inequality $U'(v) > 0$ is equivalent to

$$c'(v)(Q_\pi(v) - 1) - c(v)Q'_\pi(v) > 0.$$

By assumption, $c(v) > 0$ for all v, so we can arrange the inequality to

$$\frac{c'(v)}{c(v)} > \frac{Q'_\pi(v)}{Q_\pi(v) - 1}.$$

We have also assumed that $c'(v) \leq 0$ and since $c(v) > 0$, we also know that $0 \geq \frac{c'(v)}{c(v)} > \frac{Q'_\pi(v)}{Q_\pi(v)-1}$. Thus $\frac{Q'_\pi(v)}{Q_\pi(v)-1}$ is negative, which is equivalent to $Q'_\pi(v) < 0$.

Both sides of the inequality are negative, so we can rewrite it as

$$-\left|\frac{c'(v)}{c(v)}\right| > -\left|\frac{Q'_\pi(v)}{Q_\pi(v) - 1}\right|$$

and negate the entire inequality to obtain

$$\left|\frac{c'(v)}{c(v)}\right| < \left|\frac{Q'_\pi(v)}{Q_\pi(v) - 1}\right|.$$

The result follows. ■

A.3 Proof of Proposition 3

PROPOSITION. A set of sufficient conditions for the agent to choose an interior value of $v \in (0, v_{\max})$ are that:

$$\frac{c'(0)}{c(0)} > \frac{Q'_\pi(0)}{Q_\pi(0) - 1}, \tag{34}$$

and

$$\frac{c'(v_{\max})}{c(v_{\max})} < \frac{Q'_\pi(v_{\max})}{Q_\pi(v_{\max}) - 1}. \tag{35}$$

Proof.

To show an interior maximum exists, we first evaluate the derivative $U'(v)$ at the endpoints $v = 0$ and $v = v_{\max}$.

At $v = 0$, using the proof from Proposition 2, the first condition $\frac{c'(0)}{c(0)} > \frac{Q'_\pi(0)}{Q_\pi(0)-1}$ implies the numerator of $U'(0)$, $c'(0)(Q_\pi(0)-1)-c(0)Q'_\pi(0)$, is positive. Since the denominator $(Q_\pi(0)-1)^2 > 0$, it follows that $U'(0) > 0$.

At $v = v_{\max}$, the condition $\frac{c'(v_{\max})}{c(v_{\max})} < \frac{Q'_\pi(v_{\max})}{Q_\pi(v_{\max})-1}$ implies the numerator of $U'(v_{\max})$, $c'(v_{\max})(Q_\pi(v_{\max})-1) - c(v_{\max})Q'_\pi(v_{\max})$, is negative. Thus, $U'(v_{\max}) < 0$.

Since $U'(0) > 0$, we know that $U(v)$ is increasing at 0, which means $v = 0$ cannot be optimal. Similarly, since $U'(v_{max}) < 0$, we know $U(v)$ is decreasing at $v_{\max}$, which means $v = v_{\max}$ cannot be optimal.

We assume $U'(v)$ is continuous on $[0, v_{\max}]$. Then there exists some $\hat{v} \in (0, v_{\max})$ where $U'(\hat{v}) = 0$. Furthermore, $U'(v)$ transitions from positive to negative as v increases from 0 to $v_{\max}$, which means $\hat{v}$ is a local maximum. Therefore, $U'(\hat{v}) = 0$ must yield higher utility than either endpoint, and thus the optimal solution is interior.

■

A.4 Proof of Lemma 4

LEMMA. If $\bar{t}'(v) > 0$, then $Q'_\pi(v) > 0$ and $U'(v) > 0$.

Proof. First, we show that $t'(v) > 0$ implies $Q'_\pi(v) < 0$:

Using the definition of $t(v) = \frac{c(v)}{\Delta\pi(v)}$, we obtain

$$t'(v) = \frac{c'(v)\Delta\pi(v) - c(v)\Delta\pi'(v)}{[\Delta\pi(v)]^2}.$$

If the above expression is positive, since $\Delta\pi(v) > 0$, its numerator must also be positive:

$$c'(v)\Delta\pi(v) - c(v)\Delta\pi'(v) > 0.$$

Since $c'(v) \leq 0$, it follows that $\Delta\pi'(v) < 0$, i.e., $\Delta\pi(v)$ is strictly decreasing.

Now, re-writing $Q_\pi(v) = 1 + \frac{\Delta\pi(v)}{\pi_0(v)}$, its derivative is

$$Q'_\pi(v) = \frac{\Delta\pi'(v)\pi_0(v) - \Delta\pi(v)\pi'_0(v)}{[\pi_0(v)]^2}.$$

Because $\Delta\pi'(v) < 0$ and $\pi'_0(v) \geq 0$ is increasing, it follows that $Q'_\pi(v) < 0$.

Next, we can show that $U'(v) > 0$ by rewriting $U(v) = \pi_0(v)t(v)$, and differentiating to yield

$$U'(v) = \pi'_0(v)t(v) + \pi_0(v)t'(v).$$

Because $\pi_0(v) > 0$, $t(v) > 0$, $\pi'_0(v) \geq 0$, and by hypothesis $t'(v) > 0$, both terms are nonnegative (with the second term strictly positive). Hence,

$$U'(v) > 0.$$

The rest of the lemma follows from Proposition 2.

■

A.5 Proof of Lemma ??

LEMMA. If $\bar{t}'(v) > 0$, then the strategic agent will invest either v^* or $v_{\max}$.

Proof. In the two-period model, the strategic agent's objective function is the sum of the shirking first period's utility $\pi_0(v)\bar{t}(0)$ and the second period's utility $\pi_1(v)\bar{t}(v) - c(v) = \pi_0(v)\bar{t}(v)$, assuming the agent works in the second period.

If there is no displacement, the objective function is

$$\max_{v\in[0,v^*]} \pi_0(v)\bar{t}(0) + \pi_0(v)\bar{t}(v) \tag{36}$$

The optimum is achieved at the point $v^S = v^*$, because by assumption, both $\pi'_0(v) > 0$ and $\bar{t}'(v) > 0$ and thus, the objective function is always increasing.

If there is displacement, the objective function is

$$\max_{v\in(v^*,v_{\max}]} \pi_0(v)\bar{t}(0) \tag{37}$$

which is optimized at $v^S = v_{\max}$ since $\pi'_0(v) > 0$, so the objective is always increasing.

Since these two cases cover all possible values of v, the result follows. ■

References

Daron Acemoglu. 2002. Technical Change, Inequality, and the Labor Market. *Journal of Economic Literature* 40, 1 (March 2002), 7–72. https://doi.org/10.1257/0022051026976

Daron Acemoglu. 2021. Harms of AI. https://doi.org/10.3386/w29247 arXiv:29247

Ajay K. Agrawal, Joshua S. Gans, and Avi Goldfarb. 2023. The Turing Transformation: Artificial Intelligence, Intelligence Augmentation, and Skill Premiums. https://doi.org/10.3386/w31767 arXiv:31767

David H. Autor, Frank Levy, and Richard J. Murnane. 2003. The Skill Content of Recent Technological Change: An Empirical Exploration*. *The Quarterly Journal of Economics* 118, 4 (Nov. 2003), 1279–1333. https://doi.org/10.1162/003355303322552801

Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. 2023. Generative AI at Work. https://doi.org/10.3386/w31161 arXiv:31161

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, J. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Y. Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *ArXiv* (March 2023).

Michael Chui, Roberts Roger, Lareina Yee, Eric Hazan, Alex Singla, Kate Smaje, Alex Sukharevsky, and Rodney Zemmel. 2024. Economic Potential of Generative AI | McKinsey. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#key-insights.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4573321

Ozge Demirci, Jonas Hannane, and Xinrong Zhu. 2025. Who Is AI Replacing? The Impact of Generative AI on Online Freelancing Platforms. *Management Science* (Jan. 2025). https://doi.org/10.1287/mnsc.2024.05420

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. https://doi.org/10.48550/arXiv.2303.10130 arXiv:2303.10130

Edward W. Felten, Manav Raj, and Robert Seamans. 2023. Occupational Heterogeneity in Exposure to Generative AI. https://doi.org/10.2139/ssrn.4414065 arXiv:4414065

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. https://doi.org/10.48550/arXiv.2001.08361 arXiv:2001.08361 [cs]

Kelly Monahan and Gabby Burlacu. 2024. From Burnout to Balance: AI-Enhanced Work Models for the Future. https://www.upwork.com/research/ai-enhanced-work-models.

Shakked Noy and Whitney Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. (2023).

Nicholas Otis, Rowan Clarke, Solène Delecourt, David Holtz, and Rembrand Koning. 2024. The Uneven Impact of Generative AI on Entrepreneurial Performance. https://doi.org/10.2139/ssrn.4671369 arXiv:4671369

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. https://doi.org/10.48550/arXiv.2302.06590 arXiv:2302.06590 [cs]

J. Francis Reintjes. 1991. *Numerical control: making a new technology.* Oxford University Press.

Arun Sundararajan. 2024. Artificial Intelligence, Human Intellectual Autonomy and the Future of Work. https://www.aspeninstitute.org/blog-posts/artificial-intelligence-human-intellectual-autonomy-and-the-future-of-work/.

Catherine Tucker. 2024. How Does Competition Policy Need to Change in a World of Artificial Intelligence? *Oxford Review of Economic Policy* 40, 4 (Dec. 2024), 834–842. https://doi.org/10.1093/oxrep/grae043

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv* (Jan. 2022).

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. (2023). https://doi.org/10.48550/ARXIV.2305.10601

Doron Yeverechyahu, Raveesh Mayya, and Gal Oestreicher-Singer. 2024. The Impact of Large Language Models on Open-source Innovation: Evidence from GitHub Copilot. https://doi.org/10.2139/ssrn.4684662 arXiv:4684662