Labor as Capital: AI and the Ownership of Expertise

PRELIMINARY & INCOMPLETE

This draft: November 13, 2025. First Draft: August 12, 2025.

Zoë Cullen
Harvard

Danielle Li
MIT

Shengwu Li
Harvard

Abstract

Workplace surveillance generates data that can train AI systems to replicate worker expertise. Using a large online survey experiment of U.S. full-time workers, we show that workers adjust their knowledge contributions when made aware of this dynamic: they rationally withhold expertise due to career concerns. We formalize this behavior in a model of knowledge supply under surveillance-enabled AI and use it to evaluate alternative policies. Individual data ownership—workers' preferred policy—eliminates knowledge withholding but creates negative externalities: one worker's data strengthens the firm's bargaining position against others, potentially making all workers worse off. In contrast, collective data ownership achieves the first-best outcome, promoting knowledge sharing while allowing workers to benefit from AI-driven productivity gains. These findings highlight the importance of labor agreements in shaping AI adoption in labor markets.

JEL Codes: C93, D83, J31, J41, J71.
Keywords: AI, innovation, contracts, labor markets, future of work

Cullen: zcullen@hbs.edu. Li: d_li@mit.edu. Li: shengwu_li@fas.harvard.edu. This project was reviewed by the Institutional Review Board at MIT. Aaron Leong, Nicolas Miller, and Yuerong Zhuang provided excellent research assistance.

1 Introduction

In modern workplaces, surveillance is routine. Employees communicate on recorded platforms, project managers are tracked through task management tools, and call center agents, warehouse staff, and remote workers are all frequently monitored via audio, video, or screen recordings. As a result, work increasingly generates data about work: records of how exactly people do their jobs. Once used mainly for oversight and evaluation, these data can now be repurposed to train artificial intelligence (AI) systems to perform some of the very labor they document.

Consider a common application of generative AI: customer service chat assistants. Such models are trained using data from human workers, who have long been "recorded for quality assurance." To refine these models for specific clients, AI firms often request recordings from a company's top performers so that the model can replicate their problem-solving approaches and communication style. Once developed, the AI model can be deployed across locations, sharing the skills of top workers with other less skilled employees, increasing their productivity (Brynjolfsson et al., 2025).

This marks a fundamental shift in the nature of labor productivity. Traditionally, labor expertise resides with individual workers and is difficult to transfer, as it often involves tacit knowledge developed through experience (Polanyi, 1967). As a result, firms must effectively "rent" this expertise by employing workers period-by-period. In this setting, a worker's productivity is defined by their direct contributions: customers served, documents written, and so on. Surveillance-enabled AI changes this dynamic. By capturing detailed records of how workers perform their jobs, firms can extract and codify labor expertise, transforming it into AI capital that can be owned and scaled. A worker's productivity now includes an indirect component: the value of the knowledge they reveal to AI models trained from records of their labor.

In this paper, we define the recorded aspects of a worker's labor as their "knowledge supply." Our definition is distinct from the expertise that workers have always brought to their jobs because, by virtue of being recorded, knowledge "supplied" can potentially be codified into AI systems owned by the firm. To date, many AI systems have been trained on knowledge—documents, art, code—that workers have supplied without awareness or consent. Yet a substantial share of work-relevant expertise remains uncodified, residing with individual workers. Effectively deploying AI in the workplace will require data that captures this contextual expertise—knowledge specific to firms, clients, and moments in time. As workers recognize the potential for AI to capture and replicate their skills, they may alter how they work or advocate for new governance regimes. At the same time, some policies may be better than others at encouraging workers to share their knowledge. This raises both a descriptive and normative question: how do current labor market policies impact workers' knowledge supply? And what policies should be adopted to maximize welfare?

Through empirical and theoretical analysis, we argue that existing labor arrangements neglect the role of knowledge supply, to the detriment of both productivity and worker welfare. Empirically, we conduct an online survey experiment and find that knowledge supply is elastic: when workers are informed that surveillance data may be used to train AI models, they report a reduced willingness to document their work and an increased likelihood of evading monitoring. Treated workers also express strong support for individual ownership of their work data. Theoretically, we develop a formal model of knowledge sharing and show that workers' reluctance to contribute knowledge is a rational response to the threat of AI-driven expropriation. However, their preferred solution—individual data ownership—can backfire. When workers fail to internalize how their own data sales affect the bargaining power of others, they generate a competition externality that reduces their overall surplus. In contrast, we show that first-best knowledge sharing can be achieved through collective ownership of work data by workers.

Our paper is organized into two parts. In the first, we examine worker preferences through an online experiment with 1039 employed U.S. workers recruited on the survey platform Prolific. Participants are randomly assigned to receive truthful information about how workplace data can be used to train AI systems. Treatment group participants view a short video illustrating how AI models can perform various workplace tasks—such as communicating with customers, performing administrative tasks, or drafting presentations—emphasizing that these systems can be trained using data collected from employee surveillance. Control group participants watch a similar video that highlights the same AI capabilities but omits any reference to the role of worker data in model development. This design allows us to identify the causal effects of *awareness* of surveillance-enabled AI on workers' reported behaviors and policy preferences.

We first document that workers report possessing substantial amounts of knowledge that extends beyond their firms' documentation or training materials. For example, substantial majorities report having "some" or "a lot" of additional expertise in areas ranging from client interaction and project management (where over 80% report such knowledge) to software proficiency and data analysis.

Second, we show that workers' willingness to supply their work-specific knowledge to employers is elastic: treated workers report being more likely to refuse additional monitoring, withhold documentation of their work, or provide a recorded demonstration of their work processes. We make the concept of knowledge supply more concrete in two ways. We ask respondents whether they possess work-relevant information in unofficial communications (e.g., personal email or chat platforms) or in their personal AI prompt history. Among those who report having such information, treated respondents are 6.38% less willing to share any of it with their employer, if asked. We follow Buckman et al. (2025) and elicit the respondents willingness to pay for (1) a policy that forbids employers from monitoring or storing data about individual work activity, and (2) a policy that forbids employers from developing or adopting AI models to automate core job functions. Treated respondents are, respectively, 10% and 3% more likely to accept a pay cut for such policies.

Next, we take advantage of the fact that many respondents in our sample earn secondary income by completing surveys on Prolific. We ask whether they would allow their responses from the current survey to be used to train an AI-based survey responder, and we elicit their reservation wage for completing a longer follow-up survey specifically for AI model development.[1] Treated workers were less willing to share their current survey data for AI training and reported higher reservation wages for participating in a future survey designed to develop such models. Finally, we find that treated workers are less likely to agree with the idea that employers automatically own work products simply because they pay wages, signaling a potential challenge to the status quo around workplace data ownership.

To examine what workers *do* want, we include an additional module in which we ask treated[2] participants about their preferences over three potential policy responses: (1) banning the use of surveillance data for AI model development, (2) granting workers the right to sell their individual work data, and (3) allowing workers as a group to sell their collective data. These policies reflect proposals that have emerged in ongoing labor debates around AI, including union demands to restrict data use, calls for individual data ownership and compensation, and growing interest in data collectivization as a means of restoring bargaining power. While most treated workers express support for all three policy options, individual data ownership is the most popular, with 70% of respondents favoring the right to negotiate and sell their own work data for AI development.

In sum, our survey shows that workers report possessing valuable uncodified knowledge about how to perform their jobs, and their willingness to share this information depends on how they understand their data might be used. When made aware that surveillance-enabled AI tools could

[1] There are a growing number of companies seeking to build AI-models of human preferences for surveys and marketing. See, for instance, https://hai.stanford.edu/news/ai-agents-simulate-1052-individuals-personalities-with-impressive-accuracy.

[2] Because these policies are described concretely, posing them to control participants would effectively constitute a treatment.

replicate their expertise, workers become more likely to withhold data and express support for policy changes—especially the right to profit from their individual work data if it is used to train AI models.

The second part of our paper uses a formal model to study how firms and workers respond, in equilibrium, to an AI technology that replicates workers' skills. Our model involves one firm and multiple workers, interacting over two time periods. At each time, the firm hires workers and then the workers decide how much knowledge to contribute. Each worker can withhold knowledge at some (non-negative) cost. Workers can differ in their maximum knowledge contribution and their cost of withholding knowledge. The firm's output is increasing in each worker's knowledge contribution, and additive across workers and across time.

We model expropriative AI as increasing the firm's outside option at time 2. That is, given some vector of time-1 knowledge contributions k_1, the firm uses these as inputs to train an AI system of quality $\alpha(k_1)$, where α is a real-valued non-decreasing function satisfying $\alpha(0) = 0$. In our model, the availability of AI improves the firm's outside option: even if a position goes unfilled, the firm can still generate output equal to $\alpha(k_1)$. This can be interpreted either as the output produced by literal AI automation, or as the output produced by a low-skill worker augmented by an AI system. We assume that different workers' contributions are substitutes in training the AI system, that is, the function α exhibits decreasing differences.

In each period, wages and employment are determined by Nash-in-Nash bargaining between the firm and the workers. There are two contracting frictions. First, knowledge contributions are not contractible, so each worker is offered a wage that is not contingent on their own contribution. This reflects the tacit nature of much workplace expertise: firms cannot observe what individual workers know, and so cannot write contracts that demand particular contributions. Second, time-1 contracts concern time-1 wages and employment, but can not commit parties to time-2 wages and employment. As a result, workers face career concerns: their knowledge contributions today influence their bargaining position tomorrow.

In the baseline model without AI, workers have no reason to withhold knowledge. Thus, there exists an equilibrium in which all workers are employed in both periods, each contributes their full knowledge, and receives a share of the output proportional to their Nash bargaining weight.

Now suppose AI technology is available, but workers are unaware that their data could be used to train AI models. This setting mirrors many real-world labor markets today. Unaware of the downstream use of their contributions, workers would naively continue to supply their full knowledge at time-1. This would (weakly) increase output at time-2, because the AI model might be more productive than some workers. However, the resulting improvement in the firm's outside option weakens workers' bargaining positions and leads to lower wages in period 2. The gains from AI accrue to the firm, raising profits while reducing worker surplus, relative to the no AI case.

However, in equilibrium, workers are unlikely to remain naive to the potential uses of their work data. In this case, we show workers will withhold their knowledge contributions at time-1, in order to preserve their career prospects at time-2. Such sandbagging reduces time-1 output, as well as the quality of the AI system that firms are able to develop. Workers can be worse off compared to the no-AI baseline because sandbagging reduces their wages today and AI reduces their wages tomorrow. Indeed, under some conditions, the productivity benefits of AI are outweighed by the harms from knowledge withholding, reducing total output and leaving even the firm worse off. Thus, in principle it can be a Pareto improvement for firms to commit not to use their workers' data to train AI models.

We use this model to shed light on several alternative policies: banning workplace surveillance, granting workers individual ownership over their work data, and establishing collective ownership of work data among workers.

Banning surveillance cuts off the data necessary to train AI models and, in our setting, is akin to banning AI development as a whole. This policy restores the baseline no-AI equilibrium, improving workers' job security but forgoing potential productivity gains from AI.

Suppose, instead, that workers are granted ownership of their individual work data, akin to workers' favored policy from our survey. Formally, we model this as giving workers the right to bargain over whether their data is used at time-2, and at what price. In the event of disagreement, the firm would be unable to use that worker's data for AI model training.

Individual ownership allows workers to profit from the value of their work data. As such, workers no longer have a motive to withhold knowledge, and there is an equilibrium with full knowledge contributions. In this way, individual ownership raises time-1 output while enabling the full productivity gains from AI. However, individual ownership does not guarantee that workers share in the gains from AI. Because workers' data are substitutes, their marginal contributions to AI quality sum up to less than the total. In the extreme case, when workers are symmetric and their data are perfect substitutes, workers receive *none* of the output gains from AI, no matter their Nash bargaining weight. Under some conditions, individual worker ownership can backfire, leaving workers strictly worse off compared to the no-AI baseline. Intuitively, individual ownership encourages each worker to contribute data, but their data has an externality: it improves the firm's bargaining position vis-a-vis other workers, lowering other workers' future wages.

Finally, consider a policy in which workers collectively own their data and collectively bargain over wages, so that in the event of disagreement, the firm can use no worker's data. This prevents the negative externalities that arise under individual ownership, because one worker's contribution no longer improves the firm's bargaining position against other workers. In the equilibrium with collective ownership, both firms and workers are better off compared to the no-AI baseline. While our simple model abstracts from intra-union frictions, it nonetheless suggests that collective action could be a useful policy tool to ensure that workers share in the gains from AI.

These findings have important implications for how societies should govern the transition to AI-augmented production. While much attention has focused on consumer data and privacy rights, the workplace context raises distinct issues about the ownership and control of surveilled human capital data. Our findings suggest that with greater worker awareness, productivity may slow as workers withhold valuable knowledge. Workers' preferences for individual data ownership, while popular, may not maximize worker welfare due to the externalities of one worker's data on the value of another employee. Our framework highlights policies that would encourage workers to internalize those externalities and take actions aligned with the social planner.

2 Background: Surveillance-enabled AI Development

While efforts to codify labor date back to at least the scientific management practices of the Industrial Revolution, these modern tools vastly expand its scope, granularity, and potential consequences. Indeed, concerns over "AI expropriation"—whereby firms use worker data to train models that replicate their labor—are already surfacing in labor disputes across a range of industries (Glass, 2024). In this section, we provide background on the rise of workplace surveillance, the use of its data byproducts in AI development, and the institutional and legal frameworks shaping debates over the ownership and use of worker-generated data.

2.1 Modern Workplace Surveillance

Employee monitoring has become a pervasive feature of modern workplaces, with companies deploying a wide array of surveillance tools—often referred to as "bossware"—to track worker behavior. In office settings, this includes software that logs keystrokes, mouse activity, or takes periodic screenshots, alongside monitoring of work-based email, phone, and chat communications. For workers engaged in more physical or manual roles—such as drivers, warehouse staff, or healthcare providers—tracking often involves in-facility cameras, app-based geolocation, and sometimes wearable devices that record biometric data (U.S. Government Accountability Office, 2024). These technologies allow employers to continuously collect data on productivity, performance, safety, and security.

Recent survey data confirm the prevalence of electronic monitoring in the workplace. A 2024 representative firm survey conducted by the OECD found that approximately 90% of U.S. firms report monitoring their workers in some way, with 72% monitoring the speed of work, 55% monitoring the content of worker communications, and 15% tracking worker's location (Milanez et al., 2025). A separate 2024 worker-level survey conducted by the Washington Center for Equitable Growth found that 68% workers reported being subject to at least one form of electronic surveillance, with the most commonly cited methods being workplace cameras (45%) and monitoring of company-assigned devices, such as computers and smartphones (37%) (Hertel-Fernandez, 2024). Other reports indicate that digital monitoring is particularly prevalent in large organizations (Kantor and Sundaram, 2022) and that it doubled in the wake of the post-Covid shift to remote work.(Turner, 2022).

Surveillance technologies often generate records of worker behavior. In the OECD survey, 75% of U.S. firms report collecting data on their workers through surveillance tools. Among these firms, 90% indicate that workers do not have the ability to opt-out of data collection(Milanez et al., 2025).

While existing policy concern has focused primarily on the direct consequences of monitoring—such as its effects on safety, anxiety, and privacy—there has been little attention to the fact that these same data can be repurposed to train workplace AI systems, including models that may ultimately perform or replace the very tasks being monitored.[3]

2.2 Worker Data and AI Development

Recent advances in artificial intelligence (AI), particularly large language models (LLMs), depend on the availability of data, both publicly accessible and proprietary. Foundational models, such as GPT variants, are trained on internet text, books, and code to acquire broad linguistic and reasoning capabilities. While powerful, these models typically lack the specialized domain knowledge required to operate effectively within specific organizational contexts.

[3] For example, U.S. Government Accountability Office (2024); Milanez et al. (2025); Hertel-Fernandez (2024) focus on documenting concerns related to autonomy, privacy, and work place safety, but no studies mention the use of surveillance data for AI training.

Bridging this gap requires fine-tuning: retraining the base model on domain-specific data so it can learn the particular terminology, workflows, and communication patterns relevant to a given task. These data—such as customer support logs, call transcripts, or process documentation—are often proprietary and derived from workers currently performing the same tasks. Importantly, these data often capture not just formal procedures but also tacit knowledge: skills and heuristics that are hard to articulate but observable in practice (Polanyi, 1967). Surveillance technologies make it possible to capture this implicit expertise by recording how experienced workers behave on the job.

To better understand the ways in which worker-generated data can be used to build AI models, consider the following examples:

> **Call Recordings and Customer Service AI**: Modern customer service AI models are trained on call center transcripts and audio logs. Firm-specific conversations enable models to learn how to resolve issues related to the company's products and policies. In many cases, these transcripts are labeled not only with objective performance metrics such as call duration but also with specific indicators for whether the text was generated by a recognized top-performing agent. This type of labeling allows the fine-tuning process to capture and replicate the skills of specific individuals, enabling their skills to be scaled and shared (Brynjolfsson et al., 2025).
>
> **Screen Recording and Robotic Process Automation**: In many offices, employees' daily computer workflows are captured through screen monitoring and event logging software. This has given rise to AI-driven Robotic Process Automation (RPA). RPA platforms record an employee as they perform digital tasks—clicking through applications, copying data between forms, generating reports—and then uses these examples to create a software bot to replicate routine actions (Rabbit Inc., 2024). This lowers the barrier to automating highly context specific office processes.
>
> **Clinical Notes and Medical AI**: In healthcare, clinical notes written by doctors and nurses are used to fine-tune language models for medical tasks. NYU's NYUTron, for example, is an clinical AI model trained on a decade's worth of clinical notes produced by doctors and nurses employed by NYU Langone (Jiang et al., 2023). Such models can be used not only to provide clinical assistance, but also to automate healthcare tasks such as generating reports.

In addition to these current use cases, which primarily build on text or image based worker inputs, recent AI research is incorporating other types of worker input, such as video data. In robotics, systems can now learn complex manual skills directly from human video demonstrations. For instance, a recent method extracts coarse "trajectory sketches" from human demonstration videos, allowing an AI model to learn a pattern of movement that can generalize to a variety of tasks similar to the one shown. In the medical domain, imitation learning from laparoscopic surgery videos has likewise allowed a surgical robot to autonomously execute suturing and other procedural skills (Kim et al., 2025). These advances demonstrate how video recordings can enable AI systems to learn worker behaviors without manual programming or dense annotation, making it easier to transfer human skills in non-routine manual tasks. More broadly, advances in AI may expand the kinds of worker data that can be used to train future models, making surveillance data that seems uninformative today potentially valuable in the future.

Finally, there is an ongoing debate about the trajectory of AI advancement, and whether AI systems will soon surpass human reasoning (Grace et al., 2022; Allyn-Feuer and Sanders, 2023). In general, definitions of artificial general intelligence (AGI) emphasize broad cognitive proficiency (e.g. human-level or superior performance across domains), focusing on capabilities like complex reasoning and problem-solving (Bubeck et al., 2023). Yet even if AI models do attain such general reasoning ability, they would still require context-specific training and examples to excel at actual workplace tasks (Mitchell, 2021). While it is conceivable that advanced AI systems could learn domain-specific knowledge autonomously, it is likely more efficient to provide models with examples from human experts who already possess this information. In practice, this means that human-generated data and experience will likely remain important even in a world with highly capable AI models (Ramani and Wang, 2023).

2.3 Institutional and Legal Context

The use of worker data in AI model development raises important legal and labor concerns. U.S. law generally does not grant employees explicit rights over data generated through their work, leaving employers broad discretion to define terms through job contracts. While an increasing number of state laws protect consumer data (mirroring Europe's GDPR), these protections often exclude data created in the course of employment (Kim and Leavitt, 2026). Copyright law also reinforces employer control: under the "work for hire" doctrine, materials produced within the scope of employment belong to the firm, not the worker (U.S. Congress, 1909, 1976).

New AI capabilities challenge the sustainability of the current legal framework. For example, the work for hire doctrine was designed for tangible products like reports or designs, not the behavioral trace data (recordings, computer logs, etc.) that are now routinely captured in the modern workplace. Historically, such process-level data were neither valuable nor feasible to collect. Today, however, they may be important inputs for training AI systems that aim to replicate human expertise (Ajunwa, 2025).

These changes are sparking new legal and labor responses. In 2023, for example, the Screen Actors Guild secured contract language barring studios from using film recordings to train AI avatars without consent (SAG-AFTRA, 2023). More broadly, researchers and advocates have proposed new governance frameworks—such as collective data rights or worker data trusts—to ensure more equitable participation in the value created by workplace AI (Ajunwa, 2025; Kim and Leavitt, 2026; Diamantis, 2023).

These developments underscore a core tension: AI systems rely on worker-generated data, yet workers are often unaware that routine surveillance may be used to train models that replicate their skills. Little is known about how they respond to this information, or what policies they would support in light of it. In the next section, we present evidence from a survey experiment examining these questions.

3 Research Design: Awareness Experiment

3.1 Overview of the Experimental Design

Our empirical strategy takes advantage of a unique time during a technological transition to examine how workers respond when they first learn that data generated through their workplace activity can potentially be used to train AI systems to perform similar work. Although AI applications are rapidly expanding, public discourse around the use of *surveillance-based* data for AI training remains relatively limited. This limited awareness creates a natural setting for a randomized information treatment, in which we provide some workers with accurate information about AI data practices and compare their responses to those in a control group.

We focus on full-time employed workers in the U.S. Eligible participants complete a survey that embeds an information provision experiment. Prior to the treatment, the survey collects baseline information on respondents' demographics, workplace experiences, and areas of expertise. Participants are then randomly assigned to view one of two brief videos (details below): either a treatment video, which highlights how AI systems can learn from human-generated work data to replicate job tasks, or a control video similar content about AI performance in the workplace, omitting the link between human data and model performance. After viewing the video, we collect post-treatment responses that measure subjects' willingness to share work data with their employers, and their preferences over institutional safeguards and policies. This structure—collecting baseline data before treatment—ensures that initial responses are unprimed by AI-related information provision. It also allows us to isolate the causal impact of the information treatment on subsequent beliefs and preferences.

3.2 Treatment Design

Subjects are randomized to view one of two videos, each about 2-minutes long. Each video is a captioned animation that describes the use of AI in the workplace.

The animated clips convey truthful information about AI's workplace role and cover concrete examples of tasks AI already performs, including customer service calls, filing expense reports, and generating slide decks. The videos also communicate the ways in which AI is not like a human: the breadth of AI knowledge, its immunity to fatigue, and its scalability.

The key distinction between the control video and the treatment video is that the latter describes how the AI powered tools were trained on data collected in the workplace by surveilling the way human workers carried out similar tasks. To see an example of how the control video and treatment video deviate from one another, we include excerpts of the scripts below with the treatment elements italicized. For the full scripts, see Appendix Section B.

Control video.

> For example, AI-powered chat assistants can help manage difficult customer service conversations by parsing questions and suggesting tailored replies. New office automation tools can perform common office tasks like submitting expense reports without human input. AI systems can even create the slide decks consultants use to present to clients, gathering relevant data, organizing narrative flows, and applying polished visual layouts. In other words, AI models may be able to replicate some of the tasks you do.

Treatment video.

> For example, AI models can *study recordings of customer service conversations to learn how the best workers handle difficult customers*, and then copy their people management

> skills on new customer calls. AI models can also *analyze screen recordings to observe the mouse and keyboard inputs a worker uses to file an expense report*—and then use this information to automate the task. AI models can even *examine how a consultant makes slides when presenting to clients*, and learn how to produce new presentations using that person's style. In other words, *the data you produce every day can be used to teach AI models* how to replicate some of your skills.

3.3 Econometric Specification

Let i index respondents. Denote by Y_i the outcome of interest; in different sections of the paper Y_i will stand, for instance, for indices quantifying knowledge withholding, sandbagging, or support for employer data rights. Let T_i be a dummy variable that takes the value 1 if respondent i was assigned to the information treatment group, and 0 if assigned to the control group.

Our primary specification estimates the average treatment effect on outcome variable Y_i using OLS:

$$Y_i = \alpha + \beta T_i + \epsilon_i, \tag{1}$$

where β captures the causal effect of treatment on outcome Y_i and ϵ_i is an error term. When noted, we also estimate specifications that include occupation and/or AI exposure fixed effects.

Since we expect the treatment effect to be heterogeneous based on prior awareness about the use of human data in AI models, we display average treatment effects conditional on prior awareness levels. To do this, we partition the sample into three groups based on baseline awareness.

4 Empirical Findings

4.1 Baseline Descriptives

We recruited 1,039 currently employed U.S. workers through *Prolific* in the first quarter of 2025. The survey questions were designed around the person's *primary* job.

To provide a general sense of who the subjects are, we begin by presenting descriptive statistics of baseline characteristics, described in Table 1. By design, all were employed full-time at the time of the survey. The majority (84%) work in for-profit companies. Appendix Figure A1 illustrates the breakdown of respondent occupations, with the most common being those in management (22%), computer or mathematical work (17%), business and financial operations (12%), office support (9%)

and healthcare (7%). Common job titles include manager, accountant, and data analyst; the top employers include Walmart, Amazon, and Google. Just over 50% of workers spend at least 5 days in the office while 13% are fully remote. Most workers, 63%, report that they are someone's manager. Most workers work on a salaried basis (62%), earning an average annual salary of $87,657. The remaining 38% are hourly workers, earning an average hourly rate of $25.12 . Finally, the average subject is 42 years old. Forty eight percent of subjects are male and 63% are White. We include a comparison between our sample and a cross-section of the U.S. full-time workforce in **??**.

Our baseline questions assess three issues important for our study: whether workers possess specific knowledge that may be valuable to their firms, whether workers are subject to workplace surveillance, and whether workers are aware of how employee work data can be used for AI model development.

4.1.1 Uncodified Knowledge

We begin by establishing that workers hold a substantial amount of valuable knowledge that is not formally codified by their employers. Across a range of domains, we ask workers whether they possess expertise that is not captured in existing documentation, training materials, or other recorded resources—knowledge they believe their employers would lose if they were to leave the firm.[4] Figure 1 illustrates the extent of uncodified knowledge across various skill dimensions. For most categories, a substantial majority of workers report having "some" or "a lot" of expertise beyond official documentation. In areas such as client interactions, communication, project management, and data analysis, more than 80% of workers report having uncodified knowledge, with at least 40% saying they possess "a lot."

To move from broad perceptions to more concrete cases, we also ask about two specific forms of uncodified information: work-relevant content stored in personal communications (e.g., emails or chats on non-work accounts) and in personal AI prompt histories (e.g., prompts issued from non-work accounts). Appendix Figure A2 shows that a majority of workers report having work-relevant information in both places: 59% in personal communications and 60% in AI prompt histories.

In Appendix Figure A3, we create an index of workers' self reported uncodified knowledge (the share of work-relevant tasks for which workers report having "some" or "a lot" of uncodified knowledge) and correlate it with various worker attributes. Panel A shows that workers who self-identify

[4] We ask specifically about knowledge in several domains: Software & System Proficiency, Communication Skills, Time & Project Management, Data Analysis & Reporting, Process & Compliance Knowledge, Internal Collaboration & People Management, Client Interaction, Troubleshooting & Escalation, Continuous Improvement & Innovation.

as being relatively high performers within their organizations report possessing more uncodified knowledge; Panel B shows that workers with more education also report possessing more uncodified knowledge; and Panel C shows that respondents with a smaller number of coworkers in the same role report having more uncodified knowledge.

Taken together, these findings suggest that many workers possess valuable, yet uncodified, knowledge about how to perform their jobs effectively. The extent to which workers perceive themselves as holding unique information correlates with factors one might expect: it is reported more frequently by self-identified high performers, by those with higher levels of education, and by those in roles with fewer peers. Access to this type of knowledge may be an important barrier to building AI models that can reliably perform workplace tasks.

4.1.2 Workplace Surveillance

Figure 2 describes workers' beliefs about whether their employer monitors various aspects of their workplace activity. A large majority of respondents report that their employer monitors or collects data on multiple aspects of their work: 79% cite performance tracking, 66% mention communication monitoring, 62% report surveillance of output and deliverables, while 57% of workers report having their time tracked. Additionally, 43% of workers indicate they are subject to video monitoring, 41% to computer activity monitoring, 39% to location tracking and 29% to audio recording. Fewer than 50% of workers report that they had received formal guidelines or policies about what data may be collected about their work or how it could be used.

4.1.3 AI Knowledge

Panel A of Figure 3 indicates that 75% of respondents report having read about or heard of AI tools "a lot" in the past 6 months and Panel B shows that 82% of workers have tried using AI-powered tools at work, with 35% percent reporting that they use AI tools regularly. Most workers in our sample also report having at least some familiarity with how AI models are developed (Appendix Figure A5). To evaluate the accuracy of respondents' understanding, we administer a six-question multiple-choice assessment on AI knowledge. Appendix B.1 provides the exact question wording and possible answers.

The first three questions test whether respondents grasp key concepts about how AI models are built: specifically, that models learn from exposure to examples in training data, and that their outputs are refined using human feedback. The first question addresses the fundamental difference between AI models and traditional computer programs; the second focuses on the meaning of "training"; and the third asks how developers respond to mistakes made by AI models. Going forward, we will refer to the number of correct responses as their "AI knowledge score."

The next three questions assess workers' awareness of the importance of *human-generated* data for AI model development. The first asks what factor—hardware, algorithms, or human-generated content—has most driven recent advances AI capabilities in text and image generation. The next two are scenarios testing respondents' understanding of the value of expert-generated data in developing AI models. The first focuses on customer service and the second focuses on medical diagnostics. In each case, the options involve formal documents (e.g., medical textbooks), large scale data with low quality labels (e.g., data on medical procedures and insurance information), examples generated by expert workers (e.g., patient cases with diagnostic reasoning) or don't know/not sure. Going forward, we will refer to the number of correct responses as their prior or pre-treatment "awareness score."

Appendix Figure A6 plots the distribution of knowledge (Panel A) and pre-treatment awareness (Panel B) scores. In our sample, 74% of respondents correctly identify that traditional programming involves manually coding every rule, while AI models learn from examples. Similarly, 72% correctly respond that "training" refers to exposing AI models to example data. For the question about how developers address model mistakes, the plurality of respondents (44%) correctly state that human feedback is generally used to improve model performance. In all, 33% of workers respond to all 3 knowledge questions correctly. In contrast, awareness scores tend to be lower. 45% of respondents indicate that human generated images and text are the most important input into recent advances in text and image generation and only 19% of respondents answer AI model development both scenarios correctly. In all, only 11.6% of respondents answer all awareness questions correctly.

4.2 "First Stage" Awareness Treatment

To assess the efficacy of our treatment, we present all respondents with three additional AI development scenarios post-treatment: software coding, legal document review, and warehouse package handling. Each scenario follows the same structure as the pre-treatment examples, with a single correct answer that highlights the importance of data from experienced workers.[5] We refer to this as their post-treatment or posterior awareness score. See Appendix B.1 for exact question wording.

[5] The use of worker data in the software scenario is explicitly addressed in the treatment video, while neither the treatment nor control videos mention AI applications in legal review or warehousing.

Figure 4 plots post-treatment awareness scores by treatment group. In Panel A, we see that treated respondents answer 1.5 awareness questions correctly, compared to 1.0 in the control group. This effect is large and highly significant. In Panel B, we report the percentage treatment effect on post-treatment awareness by pre-treatment awareness, with corresponding coefficient estimates reported in Table 2. While all groups show significant impacts, the largest percentage increases (relative to the control group mean) occur among workers with medium pre-treatment awareness (1 or 2 correct answers out of 3). This pattern is consistent with the idea that workers with low initial awareness may not have fully absorbed the treatment, while those with high awareness had less room to improve. Our results suggest that even brief, targeted educational interventions can begin to close knowledge gaps and meaningfully alter perceptions of how AI is built. On average, treatment increases awareness scores by over 50%.

4.3 Knowledge Supply

Our main outcomes deal with workers' willingness to share their work expertise with their employers. We assess this in several ways. Our first set of questions asks respondents to self-report their willingness to share knowledge with their primary employer. Our second set of questions takes advantage of the fact that our sample includes workers who earn meaningful income on the Prolific platform to generate an incentive compatible test of workers' willingness to supply knowledge relevant to their Prolific labor (paid survey work).

4.3.1 Self Reported Knowledge Supply (Primary Job)

Figure 5 reports the impact of treatment on workers' self reported willingness to codify the specific work skills and processes relevant for their primary job. Specifically, we provide three codification mechanisms and ask: "If you were given the opportunity, how interested are you in doing each of the following to share your specific skills with your employer?" The first three bars in Figure 5 show that, relative to the control group, treated workers are more likely to report being unwilling or very unwilling to 1) give a recorded demonstration of their skills; 2) provide detailed documentation of their work; and 3) participate in additional monitoring.

This reluctance extends to sharing work-relevant information held in unofficial channels. As discussed previously, the majority (60%) of respondents in our sample report having work-relevant information in their personal communications and AI prompt history. The bottom two bars of Figure 5 show that, among workers who report having relevant knowledge, treatment increases the share of workers who are unwilling to share any of their work-related personal communications or AI prompt history.

Table 3 reports treatment effect coefficient estimates. Column 1, in particular, creates a "withholding index" which is the average of unwillingness to share indicators across the 5 unwillingness measures we discuss above. For workers who report not having relevant personal communications or AI prompt usage, this index is computed excluding the communications or prompt indicators. Using this index, we find that treatment increases self reported knowledge withholding by 28%.

4.3.2 Actual Knowledge Supply (Prolific Survey Labor)

Our results so far rely on workers' self-reported intentions in their full-time employment roles. We also collect incentivized measures of *actual* knowledge supply with respect to workers' contributions to survey-taking on Prolific, a secondary source of income for our subjects. We collect a separate sample of 938 participants (with the same treatment and control videos), and tailor our outcome measures to capture knowledge supply associated with survey taking itself.

We exploit the fact that survey workers produce work that directly benefits a client's interest in their preferences, and indirectly generates a database that can be used in generative AI models to simulate how a subject would respond to future survey questions (Hui et al., 2024)."Synthetic panels," in particular, are an emerging application of AI tools in which AI systems seek to replicate human preferences in order to serve as a cheaper and more widely available tool for understanding preferences across different types of consumers (Henriques, 2025). Such models are often trained on large-scale human preference data, including the data collected through Prolific as a byproduct of serving individual clients' survey needs.

To assess workers' willingness to actually provide data, we describe an on-going venture to develop AI models of survey takers, and then ask respondents about their willingness to provide us with access to their Prolific data records. For control group participants, we only disclose that it is possible to develop such AI models: "Some of our research team is creating AI models that respond to surveys in order to automate the process of exploring consumer preferences." For treated workers, we highlight the value of human data in training such models: "Some of our research team is creating AI models that respond to surveys. *The goal is to use data from real survey respondents such as yourself to build personalized AI models that would respond to survey questions as you would,* in order to automate the process of exploring consumer preferences."

We then assess workers' willingness to supply their existing data: "Are you willing to let us use your responses and metadata from your previous Prolific surveys for a bonus of $10? (For a share of willing participants, we will pay this bonus and seek the data to use from past survey collectors.)" This payment represents a large payment, given that the average payment for participation in our survey was $2.50. Additionally, this payment would require no additional work for the subject.

As in our main sample, the treatment videos shift awareness of the importance of human data as input into AI models by 66%. As shown in Figure A7, control group subjects associate human data as the key input in 1.04 out of 3 AI development scenarios posed, while treated subjects make the association in 1.73 of these same scenarios, on average.

Figure 6 shows that treated workers are significantly less likely to share their existing data: 35% of treated workers refuse compared to 24% of control group workers. Column 1 of Table 5 shows that this 11 percentage point (or 45% increase) increase in withholding is highly significant. This level of withholding is sufficient to potentially introduce substantial selection bias concerns for any synthetic sample, effectively lowering the productivity of the AI model.

We next ask about workers' willingness to supply knowledge in the future. In particular, we present respondents with an opportunity to take a one hour follow-up survey "in order to learn about [their] specific background and preferences." We elicit workers' bid for this work in an incentivized manner by asking "What hourly rate would you require to participate?" and truthfully stating we would follow up with subjects whose bid fit our budget.

Panel A of Figure 7 presents the wage bids of control group workers, as percentage changes relative to their wage for the current survey. In nearly all cases, workers state that they would require the same or a higher wage to participate in a one-hour follow-up survey. Among control group workers, roughly a third of workers ask for the same hourly rate, while another third demand a substantial, greater than 50%, increase. Because workers were asked only to state the wage they would require for this particular survey, large increases may in part reflect limited interest in, or availability for, completing a longer follow-up survey.

Panel B of Figure 7 shows that workers in the treatment group demand even higher compensation for providing their future data. Relative to the control group, there is a 21 percent increase in the share of workers who report requiring more than a 50 percent wage increase, accompanied by a similarly sized decline in the share willing to participate at their current wage. These patterns underscore that treatment workers place a markedly higher price on contributing new data. The corresponding coefficient estimates are reported in Table 5.

In sum, workers who are more aware of how their data may be used choose to forgo meaningful payments for both their past and future data. This withholding—whether by limiting access to past data or raising demands around future data supply (and to say nothing of potential selection biases in the data itself)—raises the costs of gathering data that accurately reflect consumer preferences, reducing AI productivity (in this case, the usefulness of a synthetic panel).

4.4 Policy Preferences

In addition to assessing workers' knowledge supply given the status quo, we explore workers' preferred policies. First, we ask workers whether they agree with the following statement (which broadly captures the legal status quo): "My employer pays me for my work. Therefore, they should have the right to the work products I create on the job, including byproducts of my work such as any recordings or documentation of how I do my work."

Panel A of Figure 8 plots the distribution of Likert scale responses to this question among control group workers. Most workers are neutral (25.3%) or moderately supportive (38.4%) of this policy, and very few say that they disagree (12.6%) or strongly disagree (7.6%) with employer ownership.

Panel B plots the percentage change in the share of workers giving each Likert scale rating, among workers in the treatment relative to control group. We observe a 37% increase in the share of workers who strongly disagree with the policy, accounted for mostly from a decrease in the share of workers who are neutral.

Having shown that awareness challenges workers' acceptance of default employer data ownership, the final part of our survey turns to assessing their preferences over alternative policies. To do this, we present experimental vignettes to our treated subjects, describing three alternatives (see Appendix B.5 for the full wording of these policy alternatives). The first bans the use of work data in AI development:

> *"Under this policy: 1. Employer could not use work data to develop AI models, or sell work data to other firms to develop AI models'. 2. Those seeking to develop AI models would have to hire workers to specifically produce AI training data; they could not use data that was produced as the byproduct of every day work tasks."*

The second policy allows for individual control over one's own work data:

> *"Under this policy: 1. You decide if your data can be used for AI training. You could say no, and your employer couldn't include your data in their AI models. 2. Your employer can pay you to use your data, at a price that you both agree on."*

Finally, the third policy allows for collective worker control:

> *"Under this policy: 1. You and other workers in your role jointly decide whether your collective work data can be used for AI training. You could collectively say no, and your employer couldn't include any of your data in their AI models. 2. Your employer can pay you and your colleagues to use your collective data, at a price that you all agree on. You and your coworkers could then decide how to split the proceeds amongst yourselves individually."*

Given that these vignettes themselves presumed understanding of the relationship between worker-generated data and training AI models, these policy preferences could only be asked of the subjects assigned to our treatment video.

To assess workers' preferences, we follow the methodology of Buckman et al. (2025). Specifically, for each policy, we ask: "How would you feel if this policy were in place where you work?" (Positive – I would like it / Neutral / Negative – I would dislike it). Depending on their response, participants were shown a contingent valuation item: "How big a pay cut would you accept in exchange for such a policy?" or "How much extra pay would you need in exchange for such a policy?" with response categories ranging from "'None" to "More than a 35% pay raise (or pay cut)."

Figure 9 displays share of workers who indicate that they like each policy. While there is broad support for all three options, individual ownership of work data emerges as the most popular choice. Approximately 70% of respondents in the treatment group favored a policy that would grant them the right to own and sell their individual work data for AI development. Most workers (60%) are not willing to take a pay cut in exchange for the implementation of a policy they support (nor do they need to compensated if an unfavored policy were implemented). However, 10.5% of treated workers are willing to take at least a 10% pay cut in exchange for individual data ownership. In comparison, 10.5% and 10.4% of workers are willing to take 10% or greater pay cuts for collective data ownership or restrictions on AI development, respectively.

Taken together, our empirical findings show that workers' willingness to share their expertise is sensitive to how they expect their data will be used. Workers exposed to information about surveillance-enabled AI report greater reluctance to document their expertise, a higher likelihood of evading monitoring, and less willingness to share work-relevant information. They also express a clear preference for alternative governance regimes, particularly policies granting individual ownership over their work data. To assess the broader implications of these preferences and understand how different institutional arrangements affect incentives and welfare, we next develop a formal model of knowledge sharing.

5 Theory

5.1 Overview

We develop a model of AI expropriation to understand how worker awareness of surveillance affects knowledge sharing and welfare outcomes.

Three key features differentiate our setting from a standard employment relationship:

1. **Two periods**: today's data train tomorrow's AI, creating dynamic effects of knowledge contribution.

2. **Non-contractible knowledge contributions**: the firm does not yet know what skilled workers do until it surveils them and codifies their tacit knowledge, so it cannot write contracts specifying exactly what knowledge workers should provide.

3. **Limited commitment**: workers' expertise walks with them in the status quo—firms cannot guarantee lifetime employment, and workers cannot commit themselves indefinitely. Everything therefore happens under the shadow of renegotiation.

Our model formalizes a dynamic hold-up problem created by surveillance-enabled AI. In period 1, workers reveal tacit know-how while doing their jobs. These records train an AI system that becomes the firm's period-2 outside option. When future bargaining happens under limited commitment, a stronger outside option allows the firm to push down wages. Anticipating this, workers strategically withhold knowledge today, even though withholding is privately costly and reduces current output.

5.2 Model Setup

5.2.1 Players and Timing

The economy consists of one firm and a finite set of workers J. Time is discrete with two periods $t \in \{1, 2\}$, representing the present and the future.

Within each period

1. The firm bargains with each worker j over wage $w_t^j \geqslant 0$ and employment status $I_t^j \in \{0,1\}$.

2. Each employed worker chooses a non-negative knowledge contribution $k_t^j \in K^j$.

3. Production occurs and wages are paid.

Between periods, everyone observes the vector of first-period contributions $k_1 \equiv (k_1^1, k_1^2, \ldots, k_1^{|J|})$.

We normalize each worker's output to be equal to their knowledge contribution k_t^j.[6]

We denote the vector of time-t contributions $k_t \equiv (k_t^j)_{j \in J}$, and similarly the vector of time-t wages $w_t \equiv (w_t^j)_{j \in J}$. Given some dataset $D \subseteq J$, we use k_t^D to denote the vector that is identical to k_t except that elements corresponding to $j \notin D$ are equal to 0.

We model AI quality using a function $\alpha : \mathbb{R}_{\geqslant 0}^J \to \mathbb{R}_{\geqslant 0}$. Given some time-1 contributions k_1 and some dataset D, the firm develops an AI system of quality $\alpha(k_1^D)$. Intuitively, this captures the productivity of an AI system that can augment unskilled workers or even replace workers entirely.

At time 2, for each worker j, the firm with dataset D gets output

$$\max\left\{I_2^j k_2^j, \alpha(k_1^D)\right\}. \tag{2}$$

The no-AI case corresponds to $D = \varnothing$. The functional form of (2) means that the AI automates the role, instead of augmenting the worker, because the gains from hiring worker j arise only when worker j's output exceeds what the AI could separately achieve.

We assume that α is continuous and is monotone increasing in knowledge inputs: that is, for all $\underline{k}_1 \leqslant \overline{k}_1$, we have $\alpha(\underline{k}_1) \leqslant \alpha(\overline{k}_1)$.

We assume that knowledge contributions are substitute inputs; that is, for all $\underline{k}_1^j \leqslant \overline{k}_1^j$ and all $\underline{k}_1^{-j} \leqslant \overline{k}_1^{-j}$, we have

$$\alpha(\overline{k}_1^j, \underline{k}_1^{-j}) - \alpha(\underline{k}_1^j, \underline{k}_1^{-j}) \geqslant \alpha(\overline{k}_1^j, \overline{k}_1^{-j}) - \alpha(\underline{k}_1^j, \overline{k}_1^{-j}). \tag{3}$$

Here are some examples of α that are permitted by our assumptions:

1. Linear returns $\alpha(k_1) = \sum_j \beta_j k_1^j$ for constants $\beta_j \geqslant 0$,

2. Replicating the skills of the best worker $\alpha(k_1) = \beta \max_j k_1^j$ for constant $\beta \geqslant 0$,

[6] Our other assumptions do not rely on the cardinal properties of k_t^j, so if output is some continuous increasing function of k_t^j, we could rescale it so that each k_t^j is equal to the resulting output level.

3. Decreasing returns to the sum of contributions $\alpha(k_1) = h\big(\sum_j k_1^j\big)$ where $h : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ is increasing and concave.

5.2.2 Worker Characteristics

Each worker j has skill $\theta^j \equiv \max K^j > 0$ (their maximum feasible contribution) where $K^j \subset \mathbb{R}_{\geqslant 0}$ is compact. Withholding knowledge incurs private cost $c^j(k_j^t)$, where we assume that c^j is continuous and non-increasing, and $c^j(\theta^j) = 0$. These costs capture strategic behavior such as sandbagging, evading surveillance, contaminating data, or degrading documentation.

In practice, withholding takes many forms: Avoiding documentation, using off-platform channels, obfuscating prompts, or doing "shadow work" that is hard to record. Lower k_1^j means more hiding and therefore a higher cost $c^j(k_1^j)$. This captures the idea that it is effortful to keep tacit knowledge tacit.

5.2.3 Payoffs

We assume that workers' payoffs are additive across time, linear in wage and withholding costs, and that their outside option yields zero payoff. Thus, worker j's utility is

$$U^j = I_1^j \left(w_1^j - c^j(k_1^j) \right) + \psi \, I_2^j \left(w_2^j - c^j(k_2^j) \right), \tag{4}$$

where $\psi > 0$ is the weight on the future. We allow for the case that $\psi > 1$; one can interpret this as meaning that the stakes from expropriation are so large as to outweigh time discounting.

We assume that the firm's payoff is additive across time and across workers, and linear in output and wages. That is, the firm's utility is

$$\Pi = \sum_{j \in J} \left[I_1^j \left(k_1^j - w_1^j \right) + \psi \, \left(\max \left\{ I_2^j k_2^j, \alpha(k_1^D) \right\} - I_2^j w_2^j \right) \right]. \tag{5}$$

5.2.4 Solution concept

Observe that upon being hired at time 2, it is a best response for the worker to set $k_2^j = \theta^j$, at cost $c^j(\theta^j) = 0$, and this is the unique best response if c^j is decreasing. For simplicity, we will assume that hired workers at time 2 set $k_2^j = \theta^j$. Similarly we will break ties in firm-worker bargaining in favor of hiring.

An **assessment** in our model is a tuple consisting of:

1. Time-1 wages w_1.

2. Time-1 employment $\overline{J}_1$.

3. Time-1 knowledge contributions $k_1 \in \prod_{j\in J} K^j$.

4. Time-2 wages $\omega_2 : \prod_{j\in J} K^j \to \mathbb{R}^J_{\geqslant 0}$.

5. Time-2 employment $\mathcal{J}_2 : \prod_{j\in J} K^j \to 2^J$.

Note that we have specified contributions k_1^j for workers $j \notin \overline{J}_1$; these are the contributions those workers would make if (counterfactually) they were hired.

We assume that in each period, firms and workers engage in Nash-in-Nash bargaining, with exogenous bargaining weight $\gamma \in [0, 1]$ for each worker. In practice, many workplaces bargain bilaterally at the worker level (or via managers), while firm profits depend on the entire workforce. Nash-in-Nash is a tractable way to capture simultaneous bilateral bargaining while accounting for cross-worker spillovers from AI. The exogenous weight γ should be read as reduced-form bargaining strength, reflecting regulations or labor market conditions.

An assessment is an **equilibrium** if:

1. Time-1 wages w_1 and employment $\overline{J}_1$ are a Nash-in-Nash equilibrium, with the disagreement point for the worker j consisting of not being hired and not being paid.

2. For each worker $j \in J$, their contribution k_1^j maximizes their continuation payoff when the other workers contribute according to $k_1^{J_1}$.[7]

3. For each $k_1' \in \prod_{j\in J} K^j$, wages $\omega_2(k_1')$ and employment $\mathcal{J}_2(k_1')$ are a Nash-in-Nash equilibrium.

Our models differ in the firm's dataset in the event of agreement and disagreement; we describe each in the subsections that follow.

5.3 Results

5.3.1 No surveillance

Without surveillance, the firm's dataset is always $D = \varnothing$, and the disagreement point has worker j not hired and not paid.

[7] Implicitly, this means that hired workers do not observe who else is hired when deciding their own contribution; so they best-respond to the *equilibrium* hired set $\overline{J}_1$. Notice also that workers not hired at time-1 have 'passive beliefs'. That is, if (off-path) they are hired, they believe the set of other workers hired is the same.

Our model is designed to isolate the dynamic effects of expropriation by AI. Without surveillance, firms and workers are engaging in two identical and separable production stages. Knowledge contributed at time 1 has no effect on the worker's time-2 payoffs, so workers have no incentive to withhold knowledge. We state this formally in the following observation.

Observation 5.1. With no surveillance, there exists an equilibrium featuring

1. full employment in both periods,

2. full knowledge contributions $k_1^j = k_2^j = \theta^j$, and

3. wages $w_t^j = \gamma\theta^j$, where γ is the worker's Nash bargaining weight.

Moreover, if each c^j is decreasing, all equilibria involve full knowledge contributions.

This equilibrium is a useful benchmark, because workers contribute full knowledge and incur no withholding costs. This equilibrium only falls short of the first-best because it achieves none of the potential productivity gains from AI.

Even under the alternative policies that follow, there will be equilibria with full employment in both periods, because the worker's outside option yields zero payoff. For ease of exposition, we will focus on these equilibria, so that the key policy-relevant comparisons we focus on are: total surplus, worker wages, and firm profits.

5.3.2 Firm-owned AI

With firm-owned AI, the firm's dataset is always $\overline{J}_1$, and the disagreement point has worker j not hired and not paid. Thus, even if no agreement is reached with worker j at time-2, the firm still produces $\alpha(k_1)$ using last period's data.

Suppose the workers naïvely contributed full knowledge at time-1. Holding fixed the workers' time-1 contributions k_1, firm-owned AI raises the output resulting from agreement between the firm and worker j at time-2, from θ^j to $\max\left\{\theta^j, \alpha\left(k_1^{\overline{J}_1}\right)\right\}$. But it also raises the firm's disagreement payoff, from 0 to $\alpha\left(k_1^{\overline{J}_1}\right)$. Thus, Nash-in-Nash bargaining results in a fall in the worker's wage, from $\gamma\theta^j$ to $\gamma\left\lfloor\theta^j - \alpha\left(k_1^{\overline{J}_1}, 0\right)\right\rfloor_+$, where we use the notation $\lfloor x\rfloor_+$ to denote $\max\{x, 0\}$. Thus, ignoring equilibrium effects, firm-owned AI increases time-2 output and the firm's time-2 profits.

Next we study what happens in equilibrium with firm-owned AI. The time-1 contribution game has a useful structure: Even though workers are harmed by raising k_1^j (it strengthens the AI they face tomorrow), the substitutes property implies the marginal harm from revealing more is smaller when others already revealed a lot. Thus, workers' time-1 contributions are strategic complements. Intuitively, if your colleagues have already "taught the AI most of what it needs," your own contribution does little incremental damage to your future wage but still saves you withholding effort $c^j(\cdot)$, so best responses are weakly increasing in others' contributions.

To ensure that our results are not vacuous, we start with a simple sufficient condition for the existence of equilibrium. We require that (essentially) withholding costs are not so severe that they outweigh a single-worker's time-1 output. Formally, let B^j be the best-response contributions of worker j when all other workers contribute 0, that is

$$B^j \equiv \arg\max_{k_1^j} \left\{ w_1^j - c^j\left(k_1^j\right) + \psi\gamma \left\lfloor \theta^j - \alpha\left(k_1^j, 0\right) \right\rfloor_+ \right\}. \tag{6}$$

We assume that

$$\inf\left\{ k_1^j - c^j\left(k_1^j\right) : k_1^j \in B^j \right\} \geqslant 0. \tag{7}$$

We now state a result that guarantees the existence of full-employment equilibrium, and implies a (weak) reduction in time-2 wages compared to the no-surveillance case.

Theorem 5.2. With firm-owned AI, under assumption (7), there exists an equilibrium with:

1. Full employment in both periods $J = \overline{J}_1 = \mathcal{J}_2(\tilde{k}_1)$ for all $\tilde{k}_1$,
2. Time-2 wages $\omega_2^j(\tilde{k}_1) = \gamma \left\lfloor \theta_j - \alpha(\tilde{k}_1) \right\rfloor_+$ for all $\tilde{k}_1$ and all j.

Proof. We have restricted attention to equilibria with full time-2 knowledge contributions, and in every such equilibrium, Nash-in-Nash bargaining at time 2 implies that $\omega_2^j(\tilde{k}_1) = \gamma \left\lfloor \theta_j - \alpha(\tilde{k}_1) \right\rfloor_+$ for all $\tilde{k}_1$ and all j.

Given some time-1 hired set $\overline{J}_1$ and wages w_1, it follows that worker j chooses k_1^j to maximize the utility function

$$w_1^j - c^j(k_1^j) + \psi\gamma \left\lfloor \theta_j - \alpha\left(k_1^{\overline{J}_1 \cup \{j\}}\right) \right\rfloor_+ . \tag{8}$$

In order to show existence, we will establish that the simultaneous choice of k_1^j by the workers to maximize (8) is a supermodular game, in the sense of Milgrom and Roberts (1990). To do so, we first prove a technical lemma.

Lemma 5.3. Let X and Y be partially ordered sets. Suppose $f : X \times Y \to \mathbb{R}$ is monotone non-increasing[8] and has increasing differences. Suppose $g : \mathbb{R} \to \mathbb{R}$ is non-decreasing and convex. Then $g \cdot f : X \times Y \to \mathbb{R}$ has increasing differences.

We now prove Lemma 5.3. The function f has increasing differences, so for all $\underline{x} \leqslant \overline{x}$ and $\underline{y} \leqslant \overline{y}$, we have

$$f(\underline{x}, \overline{y}) - f(\overline{x}, \overline{y}) \leqslant f(\underline{x}, \underline{y}) - f(\overline{x}, \underline{y}). \tag{9}$$

By f monotone non-increasing, we have

$$\Phi \equiv f(\underline{x}, \overline{y}) - f(\overline{x}, \overline{y}) \geqslant 0, \tag{10}$$

$$f(\overline{x}, \overline{y}) \leqslant f(\overline{x}, \underline{y}). \tag{11}$$

It follows that

$$\begin{aligned} g(f(\underline{x}, \overline{y})) - g(f(\overline{x}, \overline{y})) &= g(f(\overline{x}, \overline{y}) + \Phi) - g(f(\overline{x}, \overline{y})) \\ &\leqslant g(f(\overline{x}, \underline{y}) + \Phi) - g(f(\overline{x}, \underline{y})) \text{ by (10), (11) and } g \text{ convex} \\ &\leqslant g(f(\underline{x}, \underline{y})) - g(f(\overline{x}, \underline{y})) \text{ by (9) and } g \text{ non-decreasing.} \end{aligned} \tag{12}$$

Thus, $g \cdot f$ has increasing differences. This completes the proof of Lemma 5.3.

Lemma 5.4. The simultaneous choice of k_1^j by the workers to maximize (8) is a supermodular game.

Most of the requirements for a supermodular game follow by inspection. The only non-trivial part is to show that for each worker j, the utility function

$$w_1^j - c^j(k_1^j) + \psi\gamma \max\left\lfloor \theta^j - \alpha\left(k_1^{\overline{J}_1 \cup \{j\}}\right)\right\rfloor_+ \tag{13}$$

has increasing differences in (k_1^j, k_1^{-j}). Observe that $\theta^j - \alpha\left(k_1^{\overline{J}_1 \cup \{j\}}\right)$ has increasing differences in (k_1^j, k_1^{-j}) by the assumption that worker contributions are substitutes, and it is monotone non-increasing. It follows that

$$\psi\gamma \max\left\lfloor \theta^j - \alpha\left(k_1^{\overline{J}_1 \cup \{j\}}\right)\right\rfloor_+ \tag{14}$$

[8] That is, for $\underline{x} \leqslant \overline{x}$ and $\underline{y} \leqslant \overline{y}$, we have $f(\underline{x}, \underline{y}) \geqslant f(\overline{x}, \overline{y})$.

has increasing differences by Lemma 5.3. Moreover, $w_1^j - c^j(k_1^j)$ has increasing differences trivially, because it does not depend on k_1^{-j}. The set of functions with increasing differences is closed under addition, so it follows that (13) has increasing differences in (k_1^j, k_1^{-j}). This completes the proof of Lemma 5.4.

Fixing time-1 wages w_1 and employment $\overline{J}_1$, there exists a contribution profile k_1 that satisfies the requirements of equilibrium, by Lemma 5.4 and Theorem 5 of Milgrom and Roberts (1990).

We now guess and verify that full employment at both periods, that is, $J = \overline{J}_1 = \mathcal{J}_2(\tilde{k}_1)$ for all $\tilde{k}_1$, is part of an equilibrium. This follows straightforwardly for time 2 because ω_2^j derived above gives the firm a non-negative payoff from hiring each worker.

We now consider time 1. Hiring worker j at wage w_1^j results in firm payoff

$$k_1^j - w_1^j + \sum_{l \neq j}(k_1^l - w_1^l) + \psi \sum_l \left(\alpha(k_1^J) + (1-\gamma)\left\lfloor \theta^l - \alpha(k_1^J)\right\rfloor_+\right) \tag{15}$$

and worker payoff

$$w_1^j - c^j(k_1^j) + \psi\gamma \left\lfloor \theta^l - \alpha(k_1^J)\right\rfloor_+ . \tag{16}$$

Not hiring worker j results in firm payoff

$$\sum_{l \neq j}(k_1^l - w_1^l) + \psi \sum_l \left(\alpha(k_1^{J\setminus\{j\}}) + (1-\gamma)\left\lfloor \theta^l - \alpha(k_1^{J\setminus\{j\}})\right\rfloor_+\right), \tag{17}$$

and worker payoff

$$\psi\gamma \left\lfloor \theta^j - \alpha(k_1^{J\setminus\{j\}})\right\rfloor_+ \tag{18}$$

Thus, compared to the disagreement point, hiring worker j at time 1 increases the pairwise surplus by

$$\begin{aligned} &k_1^j - c^j(k_1^j) + \psi\left(\max\left\{\theta^j, \alpha(k_1^J)\right\} - \max\left\{\theta^j, \alpha(k_1^{J\setminus\{j\}})\right\}\right) \\ &\quad + \psi \sum_{l \neq j}\left(\alpha(k_1^J) + (1-\gamma)\left\lfloor \theta^l - \alpha(k_1^J)\right\rfloor_+ - \alpha(k_1^{J\setminus\{j\}}) - (1-\gamma)\left\lfloor \theta^l - \alpha(k_1^{J\setminus\{j\}})\right\rfloor_+\right). \end{aligned} \tag{19}$$

Next we show that

$$k_1^j - c^j(k_1^j) \geqslant 0. \tag{20}$$

By hypothesis, the contribution profile k_1 is a pure-strategy Nash equilibrium of the supermodular game considered in Lemma 5.4. It follows that

$$k_1^j \in \arg\max_{\hat{k}_1^j} \left\{ w_1^j - c^j\left(\hat{k}_1^j\right) + \psi\gamma \left\lfloor \theta^j - \alpha\left(\hat{k}_1^j, k_1^{-j}\right) \right\rfloor_+ \right\}. \tag{21}$$

By Topkis' theorem, Milgrom and Shannon (1994), the right-hand side of (21) exceeds the right-hand side of (6) in the strong set order. By (7) and c^j non-increasing, we have that (20).

By (20) and α monotone non-decreasing, it follows that (19) is non-negative. We have proved that full employment at time 1 is consistent with Nash-in-Nash bargaining, which completes the proof. □

Firm-owned AI gives workers incentives to withhold knowledge, compared to the case with no surveillance. There is always an equilibrium with full time-1 knowledge contributions under no surveillance, by Observation 5.1, whereas such an equilibrium may not exist under AI. We now formalize another sense in which firm-owned AI gives workers motives for withholding. Observe that worker j's best-response knowledge contribution is

$$\arg\max_{\hat{k}_1^j} \left\{ w_1^j - c^j(\hat{k}_1^j) \right\} \tag{22}$$

under no surveillance, and by the derivation in Theorem 5.2, the worker's best-response knowledge contribution is

$$\arg\max_{\hat{k}_1^j} \left\{ w_1^j - c^j(\hat{k}_1^j) + \psi\gamma \left\lfloor \theta^j - \alpha\left(\hat{k}_1^j, 0\right) \right\rfloor_+ \right\}. \tag{23}$$

By α monotone non-decreasing and Topkis' theorem, it follows that (23) is less than (22) in the strong set order. Intuitively, worker j's knowledge contribution decreases their own continuation payoff under firm-owned AI, which gives the worker an incentive to withhold knowledge.

Under what conditions do workers withhold more knowledge in equilibrium? We now state some comparative statics results, restricting attention to equilibria of the form characterized by Theorem 5.2.

Our assumptions allow for the possibility of multiple equilibria, which can complicate comparative statics. But the workers' time-1 contributions are strategic complements, by Lemma 5.4. Thus, holding fixed the primitives, there exists a highest equilibrium, in the sense that each worker contributes weakly more at time 1 than they do in any other equilibrium (Milgrom and Roberts, 1990, Theorem 5). And there also exists a lowest equilibrium, in the sense that each worker contributes weakly less than they do in any other equilibrium. We will keep track of the highest and lowest equilibria as the primitives change.

First, we show that workers withhold less (contribute more) when withholding has a higher marginal cost.

Theorem 5.5. Consider two profiles of cost functions, $(c^j)_{j\in J}$ and $(\tilde{c}^j)_{j\in J}$, where for all workers j and contributions $k_1^j \leqslant k_1^{j'}$, we have

$$c^j(k_1^j) - c^j(k_1^{j'}) \leqslant \tilde{c}^j(k_1^j) - \tilde{c}^j(k_1^{j'}). \tag{24}$$

Holding all other primitives fixed, let $\overline{k}_1$ and $\underline{k}_1$ denote the highest and lowest equilibrium time-1 contributions under $(c^j)_{j\in J}$, and similarly let $\overline{\tilde{k}}_1$ and $\underline{\tilde{k}}_1$ denote the highest and lowest equilibrium time-1 contributions under $(\tilde{c}^j)_{j\in J}$. We have $\overline{k}_1 \leqslant \overline{\tilde{k}}_1$ and $\underline{k}_1 \leqslant \underline{\tilde{k}}_1$.

Proof. By Lemma 5.4, the worker payoffs (8) from time-1 contributions induced by $(c^j)_{j\in J}$ and $(\tilde{c}^j)_{j\in J}$ define a pair of supermodular games indexed by parameter τ, with $\tau = 0$ corresponding to $(c^j)_{j\in J}$ and $\tau = 1$ corresponding to $(\tilde{c}^j)_{j\in J}$. Each worker j's utility has increasing differences in (k_1^j, τ) by (24). Thus, the result follows by Theorem 6 of Milgrom and Roberts (1990). □

Next we show that equilibrium contributions rise when workers are better substitutes for each other. To state this formally, we slightly extend the model: Suppose that each worker occupies a distinct role, and let the AI quality in role j be denote $\alpha^j(k_1)$. All the results stated so far extend to this case, with essentially the same proofs.

Suppose we found a better way to use other workers' data to automate worker j's role. Intuitively, this would raise the AI quality in worker j's role for a given profile of contributions. And it would lower the marginal returns to AI quality from raising worker j's contribution. We now formally show that such a change increases equilibrium contributions.

Theorem 5.6. Consider two AI technologies, $\left(\alpha^j\right)_{j\in J}$ and $\left(\tilde{\alpha}^j\right)_{j\in J}$. Suppose that for each role j, we have

$$\alpha^j(k_1) \leqslant \tilde{\alpha}^j(k_1) \text{ for all } k_1, \tag{25}$$

and also

$$\alpha^j(\hat{k}_1^j, k_1^{-j}) - \alpha^j(k_1^j, k_1^{-j}) \geqslant \tilde{\alpha}^j(\hat{k}_1^j, k_1^{-j}) - \tilde{\alpha}^j(k_1^j, k_1^{-j}) \text{ for all } k_1^j \leqslant \hat{k}_1^j \text{ and all } k_1^{-j}. \tag{26}$$

Holding all other primitives fixed, let $\overline{k}_1$ and $\underline{k}_1$ denote the highest and lowest equilibrium time-1 contributions under $\left(\alpha^j\right)_{j\in J}$, and similarly let $\overline{\tilde{k}}_1$ and $\underline{\tilde{k}}_1$ denote the highest and lowest equilibrium time-1 contributions under $\left(\tilde{\alpha}^j\right)_{j\in J}$. We have $\overline{k}_1 \leqslant \overline{\tilde{k}}_1$ and $\underline{k}_1 \leqslant \underline{\tilde{k}}_1$.

Proof. The worker payoffs (8) from time-1 contributions induced by each technology are

$$w_1^j - c^j(k_1^j) + \psi\gamma \left[\theta_j - \alpha^j\left(k_1^{\overline{J}_1\cup\{j\}}\right)\right]_+ , \tag{27}$$

$$w_1^j - c^j(k_1^j) + \psi\gamma \left[\theta_j - \tilde{\alpha}^j\left(k_1^{\overline{J}_1\cup\{j\}}\right)\right]_+ . \tag{28}$$

By the same argument as in Lemma 5.4, these define a pair of supermodular games. Let us index these games by parameter τ, with $\tau = 0$ corresponding to $\left(\alpha^j\right)_{j\in J}$ and $\tau = 1$ corresponding to $\left(\tilde{\alpha}^j\right)_{j\in J}$. By (25), (26), and Lemma 5.3, it follows that each worker j's utility has increasing differences in (k_1^j, τ). Thus, the result follows by Theorem 6 of Milgrom and Roberts (1990). □

Withholding knowledge makes the AI worse, and also decreases time-1 output. Can these losses outweigh the direct productivity gains from AI? Clearly, these losses do not happen if withholding is impossible[9] or prohibitively costly. And in that case, if one special worker's data is crucial to improve the AI, that worker might be paid a high time-1 wage that outweighs their reduced time-2 wage, because their time-1 employment results in higher time-2 output from many other workers.

We now state a simple sufficient condition under which both firms and workers are strictly worse off under firm-owned AI. If withholding is free, workers optimally contribute nothing in time 1. Then $\alpha(k_1) = 0$, so there are no productivity gains at time 2—but time-1 output is also zero. As we now state, this delivers a strict Pareto loss relative to no AI: workers lose their time-1 wage and the firm loses time-1 profits, with no offsetting gain.

Theorem 5.7. Suppose that full withholding is possible ($0 \in K^j$), that withholding is free, that is $c^j(k_1^j) = 0$ for all k_1^j, that $\alpha(k_1^j, k_1^{-j})$ is increasing in k_1^j for all k_1^{-j}, that $\alpha(k_1) \leqslant \theta^j$ for all k_1 and all j, and that $0 < \gamma < 1$. Then the firm and every worker is strictly worse off under firm-owned AI, compared to the equilibrium characterized in Observation 5.1.

Proof. Under the above assumptions, for any contribution profile k_1 with $k_1^j > 0$, worker j would strictly increase their own payoff by reducing k_1^j to 0. Thus, under any equilibrium with firm-owned AI, we have $k_1^j = 0$ for all j. Since $\alpha(0) = 0$, firm and worker time-2 payoffs are identical under

[9] That is, $K^j = \{\theta^j\}$.

firm-owned AI and no surveillance. Time-1 output is zero under firm-owned AI, and thus time-1 profits and wages are both zero, which is strictly lower than wages and profits under the equilibrium in Observation 5.1 by $\theta^j > 0$ and $0 < \gamma < 1$. □

To summarize, we have found that, compared to the no-surveillance case, firm-owned AI reduces time-2 wages and time-1 knowledge contributions. Under some conditions, firm-owned AI can reduce time-1 output without any compensating increase in time-2 output, leaving the firm and every worker strictly worse off.

5.4 Alternative Policies

5.4.1 Individual data ownership

Under individual-owned AI, the firm's dataset D is equal to the set of workers hired at *both* time-1 and time-2. Thus, the disagreement point for bargaining at time-2 between the firm and worker j involves the worker not being hired, not being paid, and their data being omitted from the dataset.

Since we are breaking ties in favor of employment and the pairwise surplus from employing an additional worker is at least zero, every equilibrium with individual-owned AI involves full employment at time 2. In such an equilibrium, the pairwise surplus from employing worker j at time-2 is

$$\lambda_j(k_1, \overline{J}_1) \equiv \max\left\{\theta^j, \alpha\left(k_1^{\overline{J}_1}\right)\right\} - \alpha\left(k_1^{\overline{J}_1\setminus\{j\}}\right) + \sum_{l\neq j}\left(\max\left\{\theta^l, \alpha\left(k_1^{\overline{J}_1}\right)\right\} - \max\left\{\theta^l, \alpha\left(k_1^{\overline{J}_1\setminus\{j\}}\right)\right\}\right) \tag{29}$$

Observe that $\lambda_j(k_1, \overline{J}_1)$ is non-decreasing in k_1^j, by α monotone non-decreasing. By Nash-in-Nash bargaining, worker j's equilibrium wage is $\gamma\lambda_j(k_1, \overline{J}_1)$. At time 1, worker j chooses k_1^j to maximize

$$w_1^j - c^j(k_1^j) + \psi\gamma\lambda_j(k_1, \overline{J}_1), \tag{30}$$

and by c^j non-increasing and λ_j non-decreasing in k_1^j, it follows that it is a best response to contribute full knowledge, $k_1^j = \theta^j$. Thus, the pairwise surplus from employing any worker j at time 1 is at least zero, so there is an equilibrium with full employment at time 1.

Observation 5.8. Under individual-owned AI, there is an equilibrium with full employment in both periods and full knowledge contributions in both periods.

However, individual ownership does not guarantee that workers are better off, compared to the no-AI baseline. The intuition for this is that Nash-in-Nash bargaining compensates workers based on the marginal contribution of their data, and under substitutes the total contribution exceeds the sum of the marginal contributions. That is,

$$\alpha(k_1^J) - \alpha(0) \geqslant \sum_j \left(\alpha(k_1^J) - \alpha(k_1^{J\setminus\{j\}})\right). \tag{31}$$

To see this, observe that we can number the workers arbitrarily from 1 to $|J|$, and rewrite the left-hand side of (31) as a telescoping sum

$$\alpha(k_1^J) - \alpha(0) = \sum_{l=1}^{|J|} \left(\alpha(k_1^{\{j:j\leqslant l\}}) - \alpha(k_1^{\{j:j\leqslant l-1\}})\right) \geqslant \sum_j \left(\alpha(k_1^J) - \alpha(k_1^{J\setminus\{j\}})\right), \tag{32}$$

where the inequality follows by the substitutes assumption.

Individual ownership generates a competition externality: Each worker accounts for the (non-negative) effect of their time-1 contribution on their time-2 wage, but does not account for how their time-1 contribution decreases other workers' time-2 wage. Thus, individual ownership does not ensure that workers receive a substantial share of the rents from AI.

In some cases, individual ownership can harm workers compared to the no-AI baseline, even when workers have full Nash bargaining power. We now state this formally.

Theorem 5.9. Suppose that:

1. There are at least three workers,
2. workers have identical maximum contributions, with $\theta^j = \theta^{j'}$ for all j and j',
3. contributions are perfect substitutes, that is $\alpha(k_1) = f(\max_j\{k_1^j\})$ for some increasing function f.

Then for any Nash bargaining parameter $\gamma \in (0, 1]$, the full-contribution full-employment equilibrium under individual data ownership is strictly worse for each worker than the full-contribution full-employment equilibrium under no surveillance.

Proof. We are comparing full-contribution equilibria, so workers incur no withholding costs. Thus, it suffices to show that workers' wages are lower under individual data ownership than under no surveillance.

There are at least two workers, they have identical maximum contributions, and their contributions are perfect substitutes, so expression (29) for pairwise surplus at time 2 under individual ownership reduces to

$$\max\left\{\theta^j, \alpha\left(k_1^J\right)\right\} - \alpha\left(k_1^{J\setminus\{j\}}\right) = \left\lfloor\theta^j - f(\theta^j)\right\rfloor_+, \tag{33}$$

which is strictly less than θ^j by $f(0) = 0$, $\theta^j > 0$, and f increasing. Thus wages at time 2 under individual data ownership are strictly lower than $\gamma\theta^j$ by $\gamma \in (0, 1]$, which is the wage under no surveillance.

Next we consider time-1 wages. There are at least three workers, so deviating to not hire a single worker at time-1 has no effect on time-2 output or on time-2 wages. Thus, the pairwise surplus of hiring worker j at time-1 is equal to θ^j, and time-1 wages are the same under individual ownership and under no surveillance.

We have established that individual ownership results in the same wages at time 1 and strictly lower wages at time 2, which by $\psi > 0$ implies that workers are strictly worse off. □

To summarize, individual data ownership restores efficiency, because it ensures that each worker has no incentive to withhold knowledge at time-1. But individual data ownership does not guarantee that workers share in the efficiency gains from AI. By Theorem 5.9, under some conditions workers can be strictly worse off under individual data ownership, compared to the no-surveillance case. Under those conditions, firms benefit from AI not only because it raises time-2 output, but also because it suppresses workers' time-2 wages.

5.4.2 Collective data ownership

So far we have considered workers bargaining individually with the firm, via Nash-in-Nash bargaining. Suppose instead that the workers bargain as a single union with the firm. The union has the same Nash bargaining weight γ, and a utility equal to the sum of the individual worker utilities. Suppose furthermore that in the event of disagreement at time 2, no workers are employed and the firms dataset is $D = \varnothing$. That is, we will call an assessment an **equilibrium with collective ownership** if:

1. Time-1 wages w_1 and employment $\overline{J}_1$ are a Nash bargaining solution between the firm and the union.

2. For each worker $j \in J$, their contribution k_1^j maximizes their continuation payoff when the other workers contribute according to $k_1^{J_1}$.

3. For each $k_1' \in \prod_{j \in J} K^j$, wages $\omega_2(k_1')$ and employment $\mathcal{J}_2(k_1')$ are a Nash bargaining solution between the firm and the union.

Observe that so long as worker j's wage $\omega_2^j(k_1)$ is non-decreasing in k_1^j, it is a best response for worker j to contribute full knowledge at time 1. One example of such a scheme is to pay each worker an equal share $\gamma/|J|$ of output in each period.

Under collective ownership, the time-2 disagreement point excludes *all* workers' data. This bundles individual data rights so that when one worker contributes, they do not inadvertently strengthen the firm's hand against others. This prevents competition externalities from arising, ensuring that workers share in the gains from AI.

Observation 5.10. There is an equilibrium with collective ownership with full employment in both periods and full knowledge contributions $k_t^j = \theta^j$ in both periods. If AI raises total output, that is, if $\alpha(k_1) > \min_j \theta^j$, then this equilibrium yields strictly higher total worker surplus than the no-surveillance equilibrium characterized in Observation 5.1.

6 Conclusion

Despite the growing ubiquity of workplace monitoring, most workers remain unaware that their everyday activities can generate training data for AI systems capable of imitating their expertise. Recent AI-focused labor disputes across a range of industries nonetheless suggest a budding awareness—and an accompanying wariness—of how these technologies could reshape workers' careers.

Our survey evidence and formal model show that awareness of AI's potential to expropriate workers' expertise can prompt workers withhold their expertise, for instance by reducing documentation or attempting to evade surveillance. Especially in roles for which tacit knowledge is important, such withholding can generate negative consequences for all parties: for workers whose productivity and wages may suffer in the present; for firms whose profits depends on their ability to make use of labor expertise; and for overall economic output, which forgoes productivity gains from the adoption of effective AI systems. These frictions give workers, employers, and policymakers a shared interest in governance structures that mitigate fears of data-driven expropriation.

Our analysis also reveals a tension between workers' stated preferences and their longer-term welfare. Roughly 70% of respondents favor individual control over their work data, including the right to sell it for AI development. Our theory shows, however, that unilateral sales create a competition externality: a given worker's decision to sell their data improves the firm's outside option against all other workers. Worker's inability to internalize this spillover can leave all workers worse off, even when they hold full bargaining power.

Collective data ownership can internalize that externality and increase worker surplus. Under some circumstances, we show that even firms may prefer this arrangement, relative to retaining data ownership themselves. At the same time, collective ownership arrangements face challenges both logistically and legally. For example, the value of data contributions from high and low skilled workers can differ dramatically, creating the potential for significant intraunion frictions. At the same time, the legal basis for worker ownership may be challenging to delineate because it is typically created with firm-provided capital and intellectual property. Developing institutional mechanisms that can navigate these issues is an important direction for future research.

More broadly, codifying workplace expertise into scalable AI models offers substantial potential for productivity growth. Our paper highlights the fact that realizing that potential depends on credible arrangements for sharing the rents from worker-supplied data. Designing, testing, and refining those arrangements, whether through policy changes or contractual innovations, remains an important task for economists, computer scientists, legal scholars, and policymakers alike.

References

Ajunwa, I. (2025). Ai and captured capital. *Yale Law Journal Forum 134*. Published January 31, 2025. https://www.yalelawjournal.org/forum/ai-and-captured-capital.

Allyn-Feuer, A. and T. Sanders (2023). Transformative AGI by 2043 is <1% likely. *arXiv preprint arXiv:2306.02519*.

Brynjolfsson, E., D. Li, and L. Raymond (2025, April). Generative AI at Work. *The Quarterly Journal of Economics 140*(2), 889–942.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Buckman, S. R., J. M. Barrero, N. Bloom, and S. J. Davis (2025, February). Measuring work from home. Working Paper 33508, National Bureau of Economic Research.

Diamantis, M. E. (2023). Employed Algorithms: A Labor Model of Corporate Liability for AI. *Duke Law Journal 72*, 797–867.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2023). Gpts are gpts: An early look at the labor market impact potential of large language models.

Glass, A. (2024). Unions Give Workers a Voice Over How AI Affects Their Jobs.

Grace, K., Z. Stein-Perlman, and B. Weinstein-Raun (2022). 2022 Expert Survey on Progress in AI. AI Impacts report (Aug. 2022). Available at https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.

Henriques, A. (2025). Synthetic panels in market research: What you need to know.

Hertel-Fernandez, A. (2024, October). Estimating the prevalence of automated management and surveillance technologies at work and their impact on workers' well-being. Technical report, Washington Center for Equitable Growth, Washington, DC.

Hui, X., O. Reshef, and L. Zhou (2024). The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. *Organization Science 35*(6), 1977–1989.

Jiang, L. Y., X. C. Liu, N. P. Nejatian, M. Nasir-Moin, D. Wang, A. Abidin, K. Eaton, H. A. Riina, I. Laufer, P. Punjabi, M. Miceli, N. C. Kim, C. Orillac, Z. Schnurman, C. Livia, H. Weiss, D. Kurland, S. Neifert, Y. Dastagirzada, D. Kondziolka, A. T. M. Cheung, G. Yang, M. Cao, M. Flores, A. B. Costa, Y. Aphinyanaphongs, K. Cho, and E. K. Oermann (2023). Health system-scale language models are all-purpose prediction engines. *Nature 619*(7969), 357–362.

Kantor, J. and A. Sundaram (2022, August). Workplace productivity: Are you being tracked? The New York Times (Interactive).

Kim, J. W., T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, and A. Krieger (2025). Surgical robot transformer (SRT): Imitation learning for surgical tasks. In *Proceedings of the 8th Conference on Robot Learning (CoRL)*, Volume 270 of *Proceedings of Machine Learning Research*, pp. 130–144.

Kim, P. T. and R. Leavitt (2026). Data Rights for Workers. *Boston University Law Review*. Forthcoming; Wash. U. Legal Studies Research Paper 25-04-01.

Milanez, A., A. Lemmens, and C. Ruggiu (2025, February). Algorithmic management in the workplace: New evidence from an oecd employer survey. Technical Report 31, OECD, Paris.

Milgrom, P. and J. Roberts (1990). Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica*, 1255–1277.

Milgrom, P. and C. Shannon (1994). Monotone comparative statics. *Econometrica: Journal of the Econometric Society*, 157–180.

Mitchell, M. (2021). Why AI is harder than we think. *arXiv preprint arXiv:2104.12871*.

Polanyi, M. (1967). *The tacit dimension* (Anchor Books ed. ed.). Doubleday Anchor book, A540. Garden City, N.Y.: Anchor Books.

Rabbit Inc. (2024). How rabbit r1 works: Teach ai to do your tasks. https://www.rabbit.tech. Accessed: 2025-06-27.

Ramani, A. and Z. Wang (2023). Why transformative artificial intelligence is really, really hard to achieve.

SAG-AFTRA (2023). SAG-AFTRA: TV / Theatric Contracts 2023.

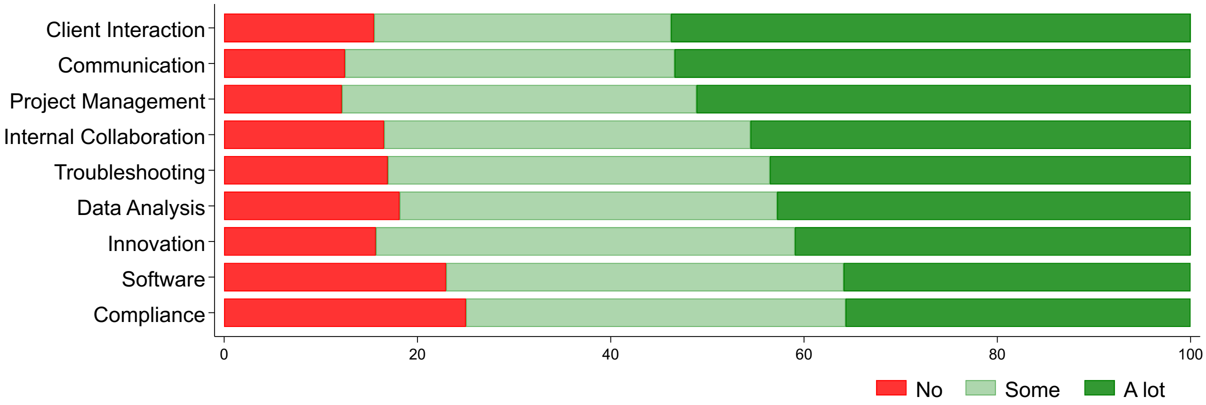
Turner, J. (2022, June). The right way to monitor your employee productivity. Gartner.

U.S. Congress (1909). Copyright act of 1909, ch. 320, § 23, 35 stat. 1075, 1080 (repealed 1976). Public Law. Enacted March 4, 1909.

U.S. Congress (1976). Copyright act of 1976, pub. l. no. 94-553, § 101, 90 stat. 2542, 2544. Public Law. Codified at 17 U.S.C. § 101.

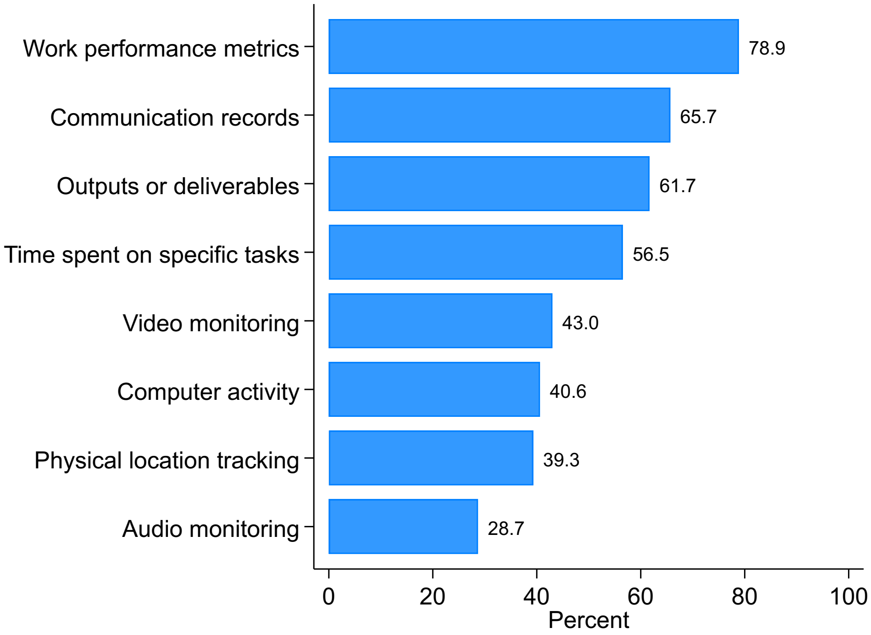
U.S. Government Accountability Office (2024, August). Digital surveillance of workers: Tools, uses, and stakeholder perspectives. GAO Report GAO-24-107639, U.S. Government Accountability Office, Washington, D.C.

Figure 1: Workers' Uncodified Knowledge



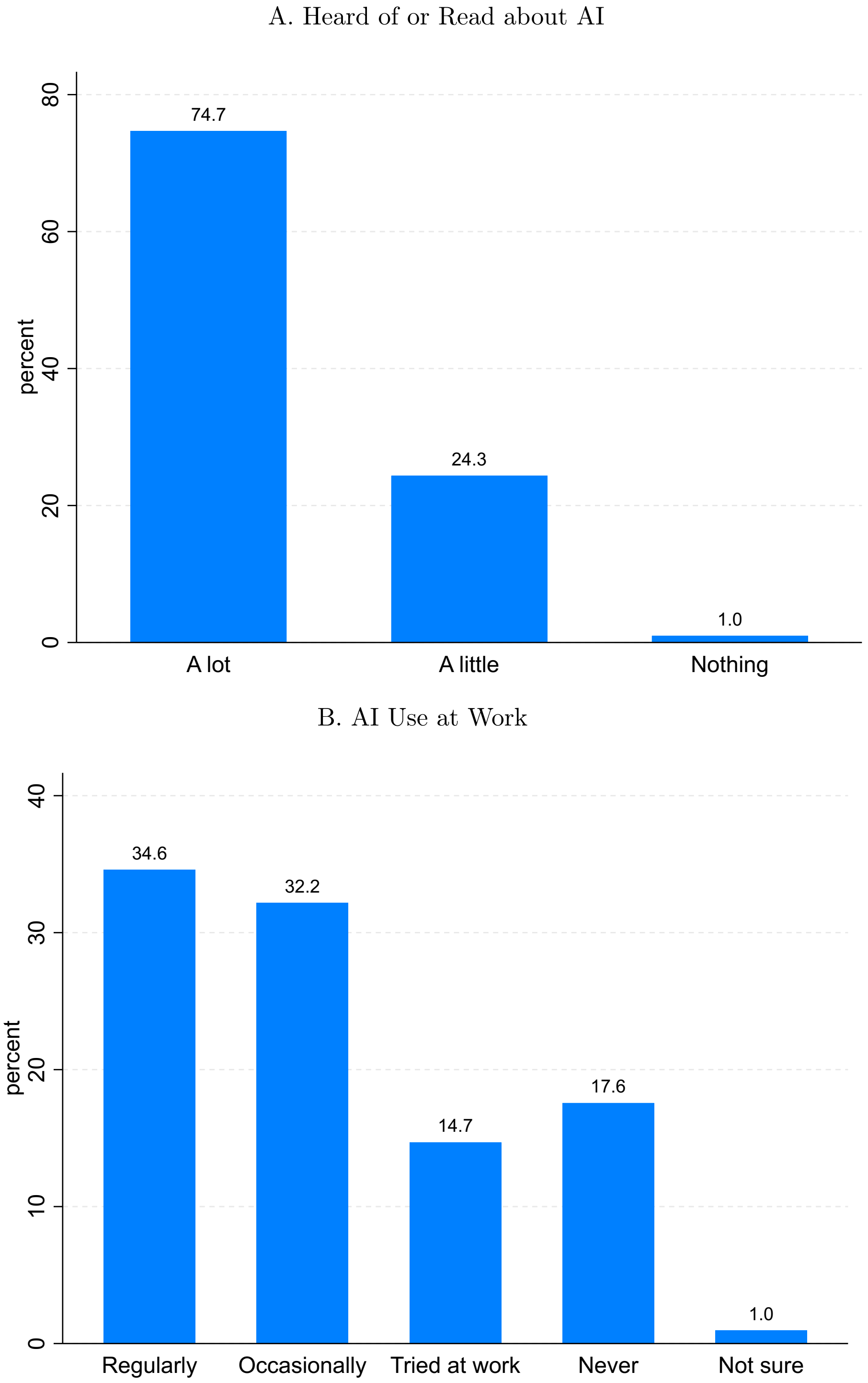
Notes: This figure shows the extent of respondents' uncodified knowledge within nine work-relevant subdomains. Respondents were asked: "Consider your current role. For each aspect listed, do you possess knowledge or skills that exceed what's captured in your company's documentation (e.g., policy manuals, process flowcharts, knowledge databases), training material (e.g., videos, slide decks, simulations, quizzes), or other forms of unstructured knowledge (e.g., email and chat records, employee AI prompt history)—expertise your employer would lose if you left?"

Figure 2: Modes of Workplace Surveillance



Notes: This figure shows the prevalence of various modes of workplace surveillance within our sample. Respondents were asked “Which of the follow activities does your employer monitor or collect data on?" This figure reports the percentage of respondents who reported experiencing each mode of surveillance.

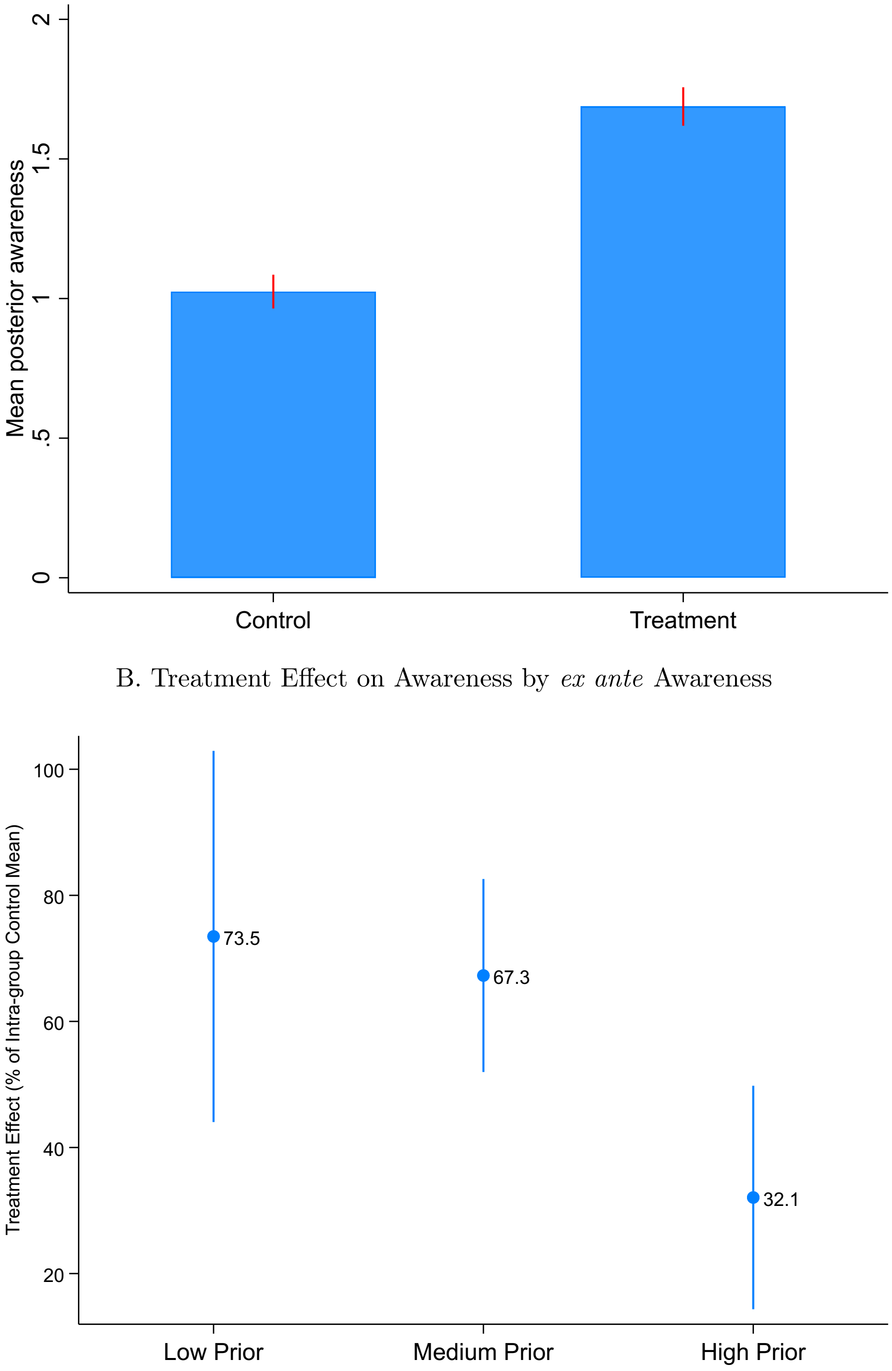
Figure 3: Workplace Awareness of and Engagement with AI tools


A. Heard of or Read about AI

B. AI Use at Work

Notes: Panel A shows how much respondents have read or heard about AI in the past six months. Respondents were asked "In the past 6 months, how much have you heard or read about AI tools?" Panel B describes the frequency at which workers in our sample use AI at work. Respondents were asked "Have you ever used AI-powered tools or systems at your workplace?"

Figure 4: Treatment Effect on Awareness Score

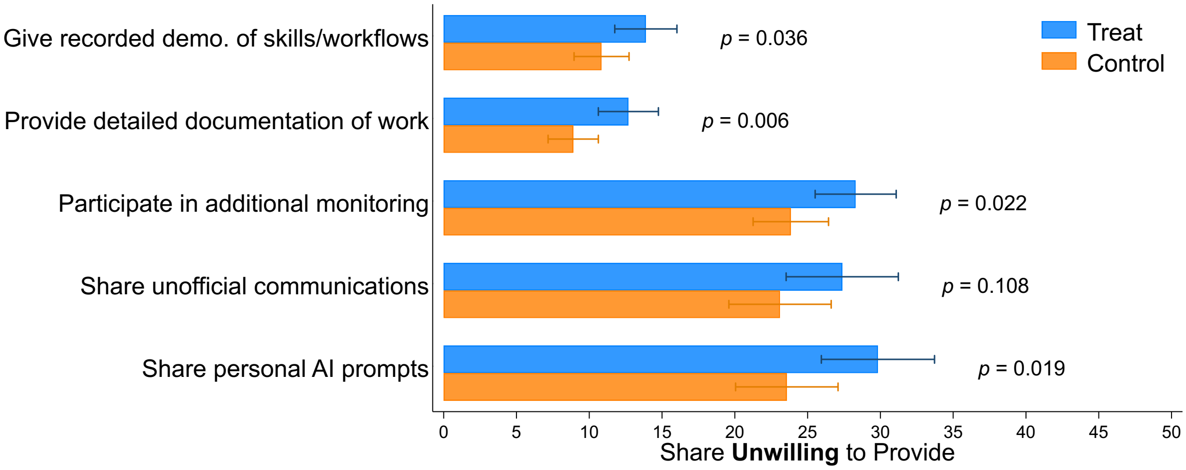


A. Post-treatment Awareness

B. Treatment Effect on Awareness by *ex ante* Awareness

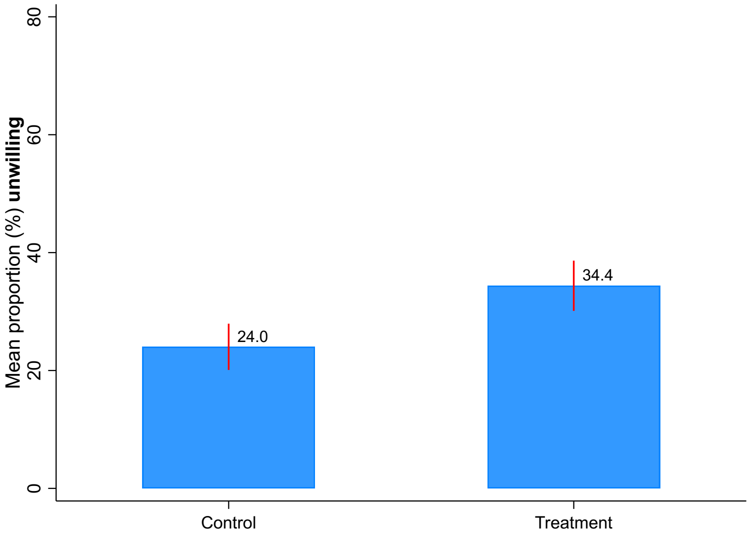
Notes: Panel A gives the average number of correct responses, within treatment and control groups, to three the post-treatment scenario questions measuring respondents' awareness of the role of human-generated data in training AI models. Specifically, respondents were asked questions related to training an AI model intended for coding, a model intended for legal work, and a model related to warehouse tasks. Panel B breaks down the treatment effect on awareness by *ex ante* awareness. Specifically, the sample was split into three groups: those with low prior awareness scores (score 0), those with medium prior awareness scores (scores 1 and 2), and those with high prior awareness score (score 3). We give treatment effects on awareness as percentage change over the mean score within each subset of the control group (low, medium, and high priors). In the notation of equation 1, we describe $(T_i/\alpha) \times 100$ within each group. Confidence intervals are drawn at 95%.

Figure 5: Share Unwilling to Share Workplace Data with Employer



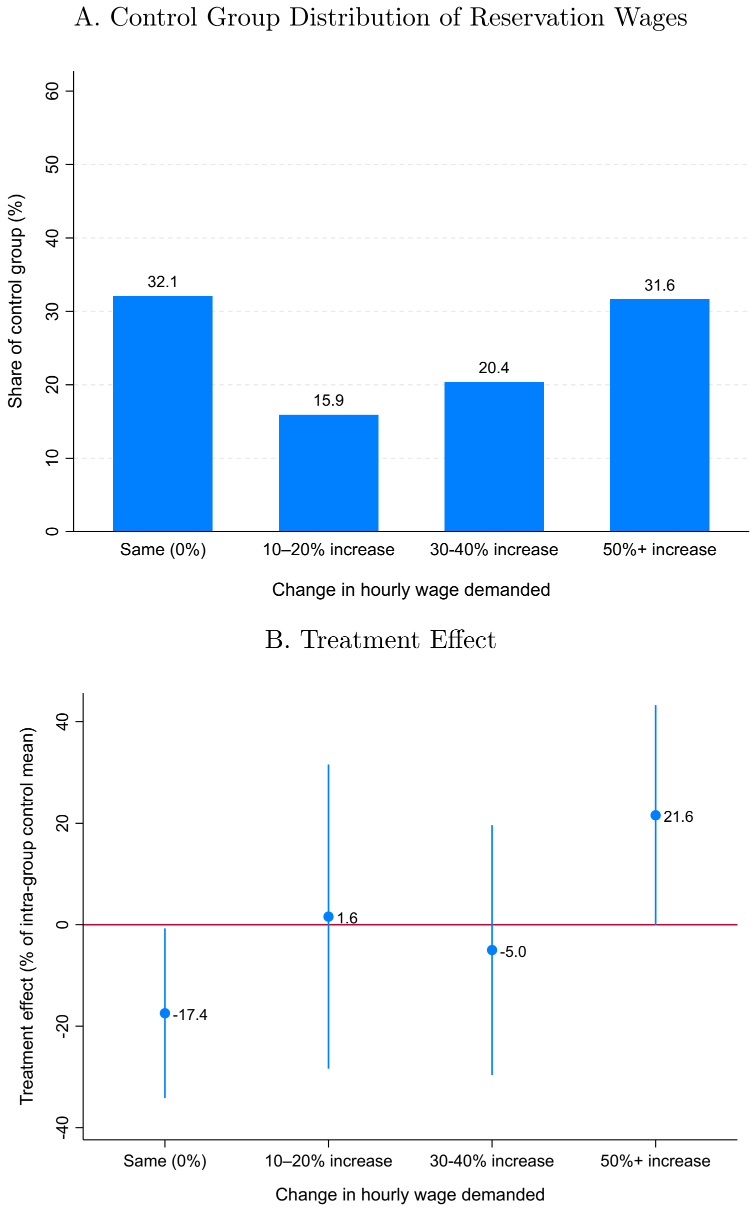
Notes: This figure shows the share of workers in treatment and control groups unwilling to share five types of work-related information with their employer. The first three bars show unwillingness to share skills via demos, to share documentation of workflow, or to be subject to additional monitoring. The last two bars capture refusal to share unofficial communications or personal AI prompts, among workers who previously stated that these channels contain information relevant to their jobs. The bars represent share unwilling to provide, within treatment and control groups, with 95% confidence intervals. We have reported two-sided t-test probabilities of observing a treatment-control difference in the share unwilling to provide that is at least as large as the one shown, under the null-hypothesis of equal means.

Figure 6: Willingness to Share Past Survey Metadata



Notes: This figure displays the share of respondents who are unwilling to provide past survey results and metadata for up to $10, within the treatment and control groups, post-treatment. Confidence intervals are drawn at 95%.

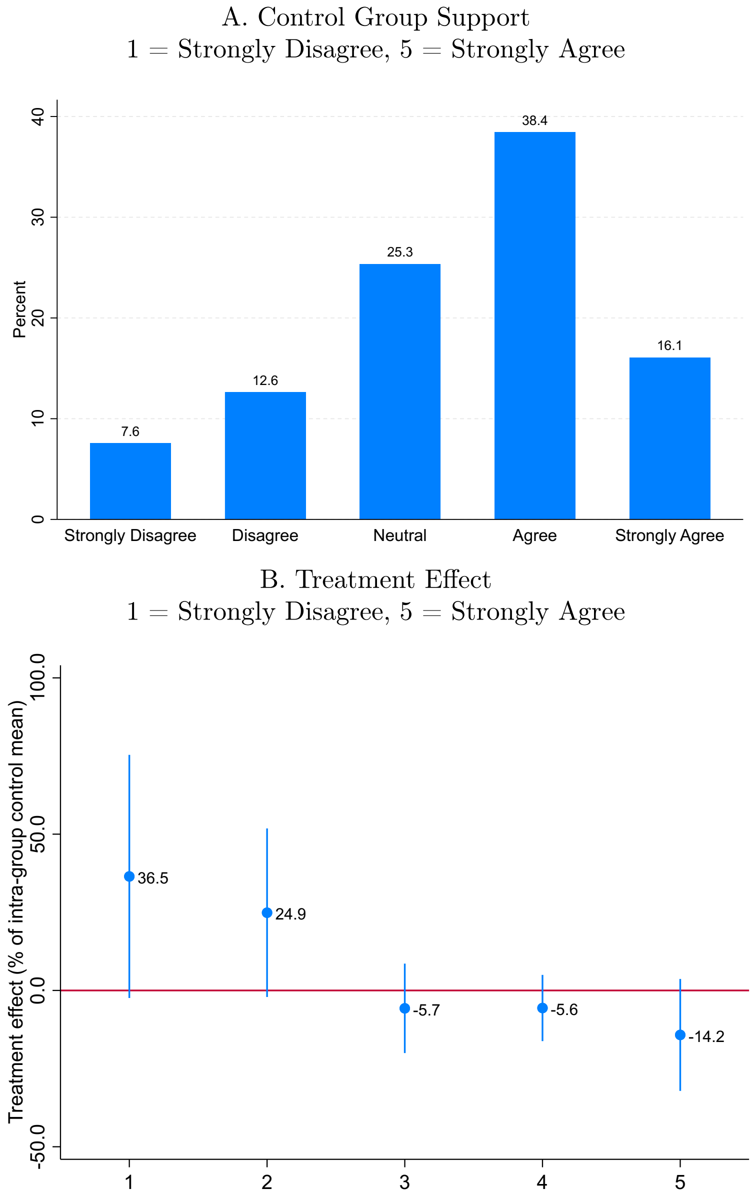
Figure 7: Reservation Wages for Future Survey


A. Control Group Distribution of Reservation Wages

B. Treatment Effect

Notes: This figure describes the baseline distribution for reservation wages for a future survey and shows changes due to treatment. Panel A plots shares of control-group respondents by the change in hourly wages required to participate in a 100-minute follow-up survey. Panel B plots treatment effects on each bin, as percent change over the control-group baseline. Specifically, respondents are first given a choice to demand a higher, lower or same reservation wage. Then, among respondents selecting the higher (lower) choices, respondents can choose to demand a 10, 25, 20, 30, 40 or 50+ percent increase (decrease) in hourly wages. The outcomes for the lower wage option are not shown as this was selected by only 2 respondents. Confidence intervals are drawn 95%.

Figure 8: Support for Employer Data Rights

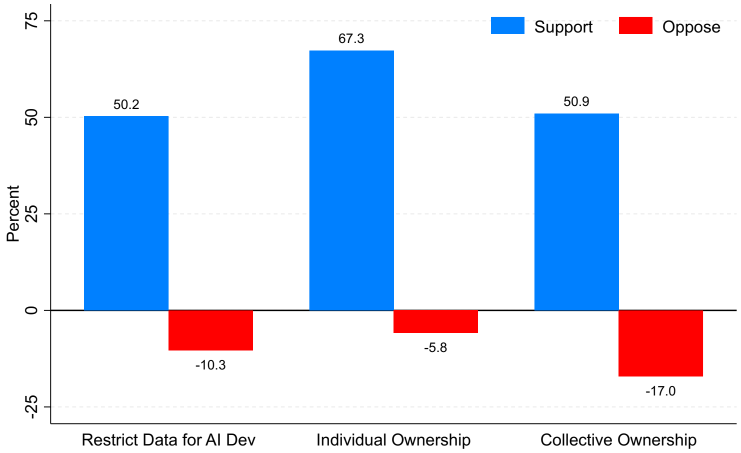


A. Control Group Support
1 = Strongly Disagree, 5 = Strongly Agree

B. Treatment Effect
1 = Strongly Disagree, 5 = Strongly Agree

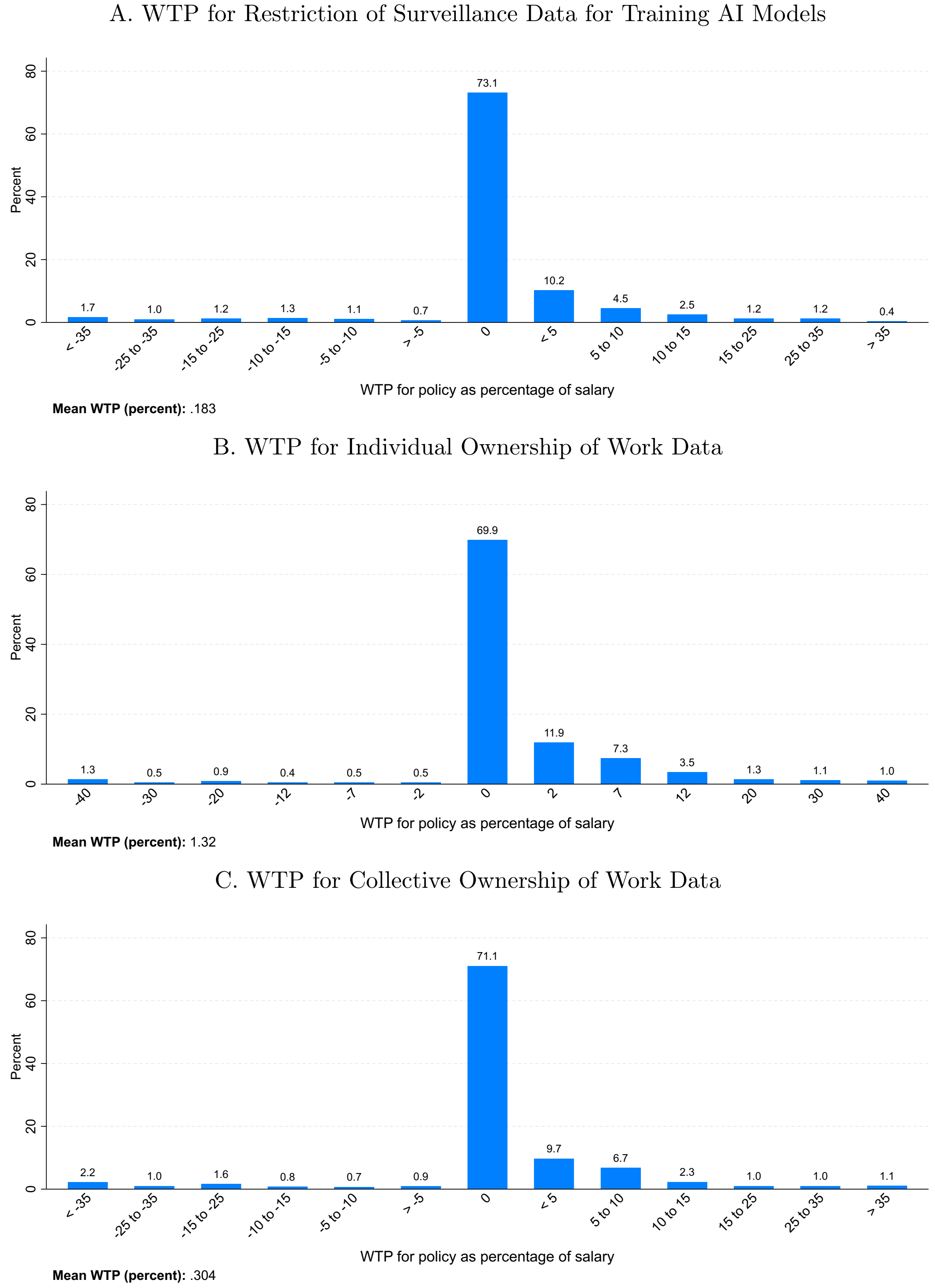
Notes: This figure describes baseline support for employer data rights and shows changes due to treatment. Panel A plots shares of control-group respondents by levels of support from 1 to 5, where 1 indicates strong disagreement and 5 indicates strong agreement. Panel B plots treatment effects on these shares, as percent change over the control-group baseline. Specifically, respondents were asked "On a scale of 1 to 5, how much do you agree or disagree with the following statement?" vis-à-vis the statement "My employer pays me for my work. Therefore, they should have the right to the work products I create on the job, including byproducts of my work such as any recordings or documentation of how I do my work." 1 is "Strongly Disagree" and 5 is "Strong Agree." Responses have been rounded to the nearest integer. Confidence intervals are drawn 95%.

Figure 9: Policy Preferences



Notes: This figure presents the share of treated respondents who and versus oppose a policy restricting data for AI development, a policy of individual data ownership, and a policy of collective data ownership. Responses were only elicited from respondents in the treatment group.

Figure 10: WTP for Policy Change


A. WTP for Restriction of Surveillance Data for Training AI Models

B. WTP for Individual Ownership of Work Data

C. WTP for Collective Ownership of Work Data

Notes: This figure presents the distribution of willingness to pay (WTP) for each of the three policies, expressed as a percentage of salary, among treated workers. A positive value reflects an acceptable salary reduction to enact the policy, whereas a negative value shows the necessary salary increase. Respondents who were neutral about a policy were coded as having 0% WTP.

Table 1: Summary Statistics

	All	Control	Treated	p-value
Age	41.8	41.5	42.1	0.307
	(13.5)	(13.5)	(13.5)	
Male	0.48	0.49	0.48	0.913
White	0.63	0.63	0.63	0.826
For-profit firm	0.82	0.81	0.82	0.450
Fully Remote	0.78	0.79	0.77	0.587
5+ Days in Office	0.55	0.54	0.55	0.653
Unionized	0.27	0.28	0.27	0.771
Manager	0.63	0.62	0.63	0.744
Salaried	0.62	0.61	0.62	0.474
Annual Salary	87,657	90,261	85,023	0.145
	(63,976)	(71,980)	(54,638)	
Hourly	0.38	0.39	0.38	0.474
Hourly Wage	25.12	25.01	25.23	0.853
	(16.43)	(16.99)	(15.81)	
N	2,055	1,046	1,009	2,055

Notes: This table reports means and, where shown in parentheses, standard deviations for demographic and workplace characteristic. The first column covers the full sample, the second and third restrict to control and treated subsamples, and the final column gives two-sided p-values from unequal-variance t-tests of equality between treatment and control means. Observations are individuals; the sample size appears in the last row.

Table 2: Treatment Effect on Awareness

	Dependent Variable: Post-Treatment Awareness Score			
	(1) All	(2) Low Prior Score	(3) Medium Prior Score	(4) High Prior Score
Treatment	0.663***	0.562***	0.690***	0.570***
	(0.047)	(0.081)	(0.060)	(0.138)
Control Mean	1.025	0.765	1.026	1.778
N	2055	586	1232	237

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table describes the effect of treatment on post-treatment awareness scores. Each cell shows the coefficient from an OLS regression of the post-treatment awareness score on a treatment indicator. Column 1 uses the full sample. Columns 2, 3, and 4, respectively, restrict to respondents with low (score 0), medium (scores 1-2), and high (score 3) baseline awareness.

Table 3: Treatment Effect on Willingness to Share Workplace Data

	Dependent Variable: Knowledge Withholding Measures					
	(1) Index	(2) Demo.	(3) Doc.	(4) Monitor	(5) Comms.	(6) Prompts
Treatment	0.043***	0.030**	0.038***	0.045**	0.043	0.063**
	(0.013)	(0.015)	(0.014)	(0.019)	(0.027)	(0.027)
Control Mean	0.154	0.108	0.089	0.238	0.231	0.236
N	2001	2050	2046	2047	1069	1093

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table describes the effect of treatment on respondents' willingness to share workplace data with their employers. The dependent variables are indicator measures of unwillingness to share specific types of knowledge with the employer. Columns 2 through 6 reflect separate dummies that equal one when a respondent is unwilling to share demonstrations, documentation, additional monitoring data, unofficial communications, or personal AI prompts. Column 1 aggregates these five dummies into an index that ranges from zero to five. Each coefficient represents the change, due to treatment, in the probability (or index value) of withholding the corresponding information.

Table 4: Support for Employer Data Rights

	Dependent Variable: Support for Employer Data Rights (Likert Categories)					
	(1) Avg. Score	(2) Strongly Disagree	(3) Disagree	(4) Neutral	(5) Agree	(6) Strongly Agree
Treatment	-0.154***	0.028**	0.031**	-0.014	-0.022	-0.023
	(0.051)	(0.013)	(0.015)	(0.019)	(0.021)	(0.016)
Control Mean	3.428	0.076	0.126	0.253	0.384	0.161
N	2055	2055	2055	2055	2055	2055

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This figure describes the effect of treatment on respondents' support for employer data rights. Column 1 regresses the five-point Likert score (1 equals "Strongly Disagree," 5 equals "Strongly Agree") on the treatment indicator. The coefficient captures the average change in support for employer ownership of work data. Columns 2 through 6 use separate dummies for each categorical response; their coefficients represented the percentage changes in the probability of selecting that category.

Table 5: Treatment Effect on Knowledge Sharing (Prolific Work)

	Agree to Share	Follow-up Survey Reservation Wage			
	(1)	(2) Same (0%)	(3) 10-20% increase	(4) 30-40% increase	(5) 50%+ increase
Treatment	-0.104***	-0.055*	0.003	-0.010	0.068**
	(0.030)	(0.030)	(0.024)	(0.026)	(0.031)
Control Mean	0.760	0.318	0.158	0.202	0.314
N	938	936	936	936	936

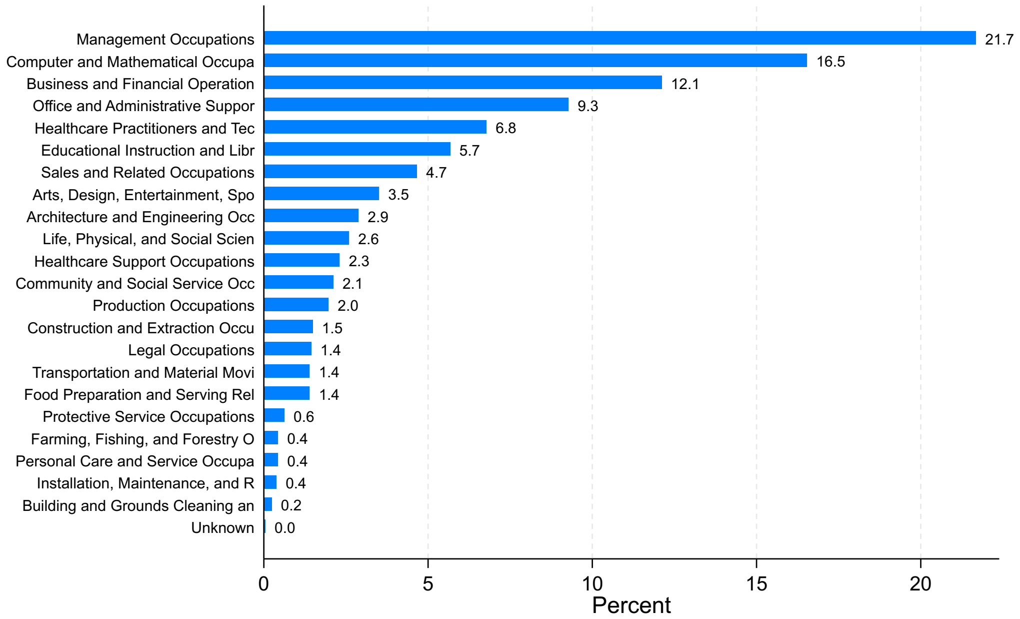
Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table describes the effect of treatment on respondents' willingness for their survey responses to be used for the development of AI models. Column 1 reports the effect of treatment on willingness to let the researchers use their survey responses and metadata for AI development; the outcome is a binary variable equal to one for consent. Columns 2 through 5 show treatment effects on respondents' selected reservation wages for completing a 100-minute follow-up survey. Specifically, respondents are first given a choice to demand a higher, lower or same reservation wage. Then, among respondents selecting the higher (lower) choices, respondents can choose to demand a 10, 25, 20, 30, 40 or 50+ percent increase (decrease) in hourly wages. The outcomes for the lower wage option are not shown as this was selected by only 2 respondents.

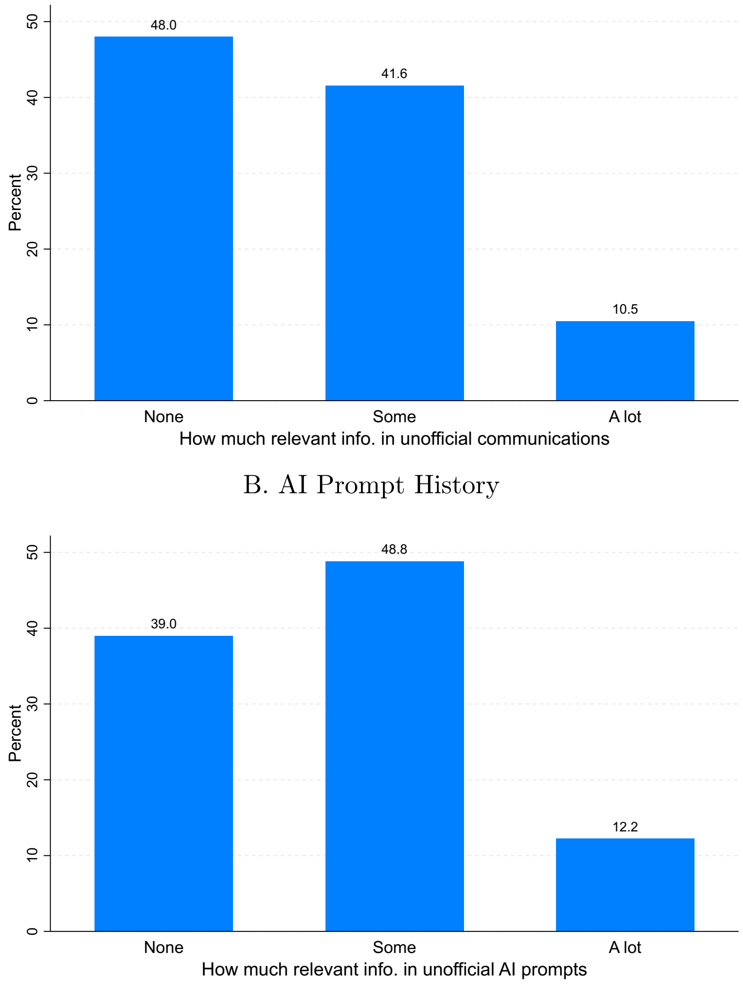
A Appendix Figures

Figure A1: Respondent Occupations (SOC)



Notes: This figure shows the distribution of SOC major occupation groups within our sample.

Figure A2: Work-relevant Data in Unofficial Channels

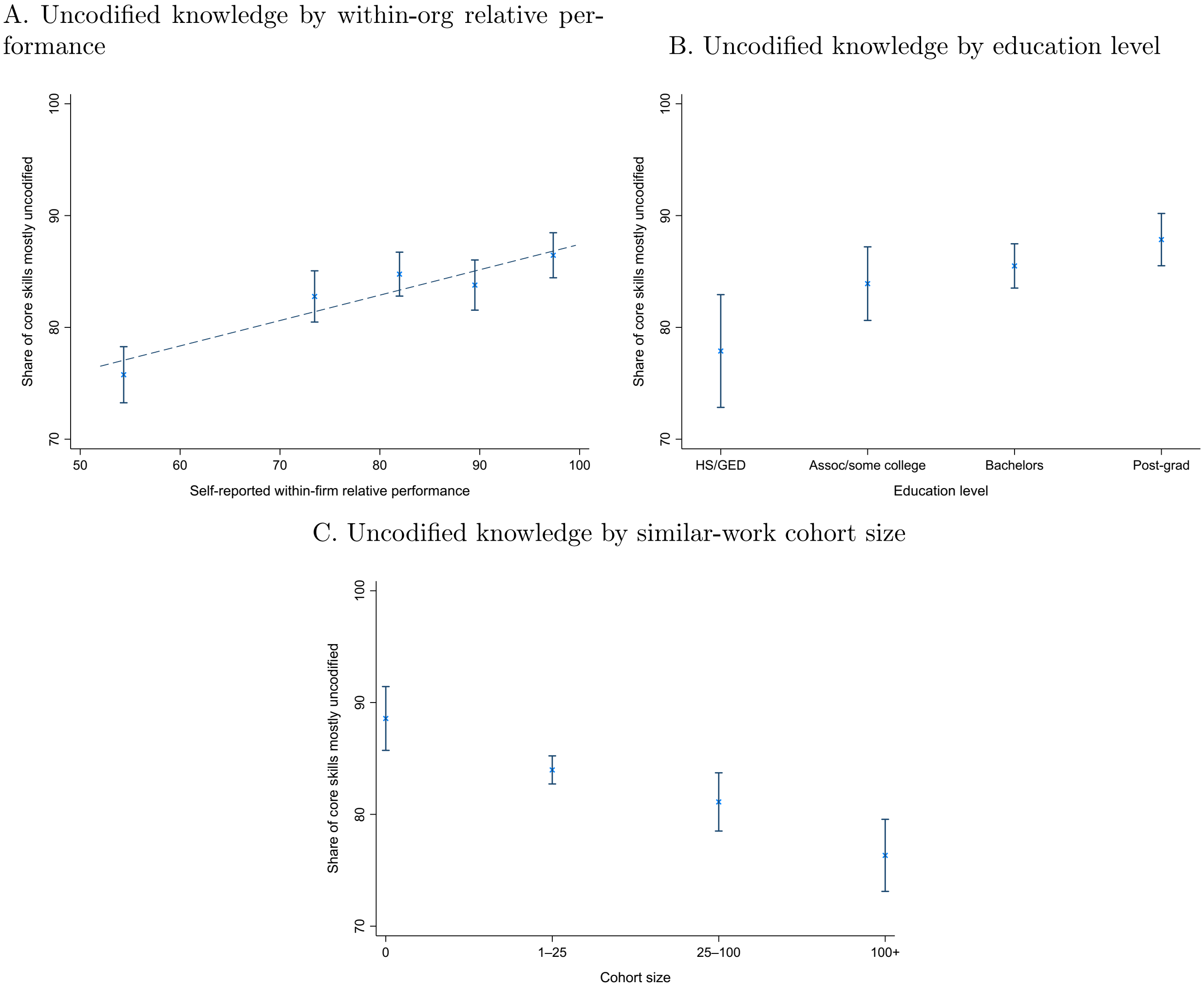


A. Unofficial Communications

B. AI Prompt History

Notes: This figure describes the presence of work-relevant data in respondents' unofficial communications and personal AI prompts. To construct panel A, respondents were first asked "To what extent does your unofficial communications (e.g., personal email, messaging apps) include work-related information that would be useful to your employer?" For panel B respondents were asked: "Consider your use of non-official AI tools for work purposes, such as using a personal ChatGPT account. To what extent does your AI prompt history include work-related information that would be useful to your employer?" Respondents who indicated that they did not use AI at work (9.923% of the sample) were excluded in plotting panel (b).

Figure A3: Heterogeneity in uncodified knowledge



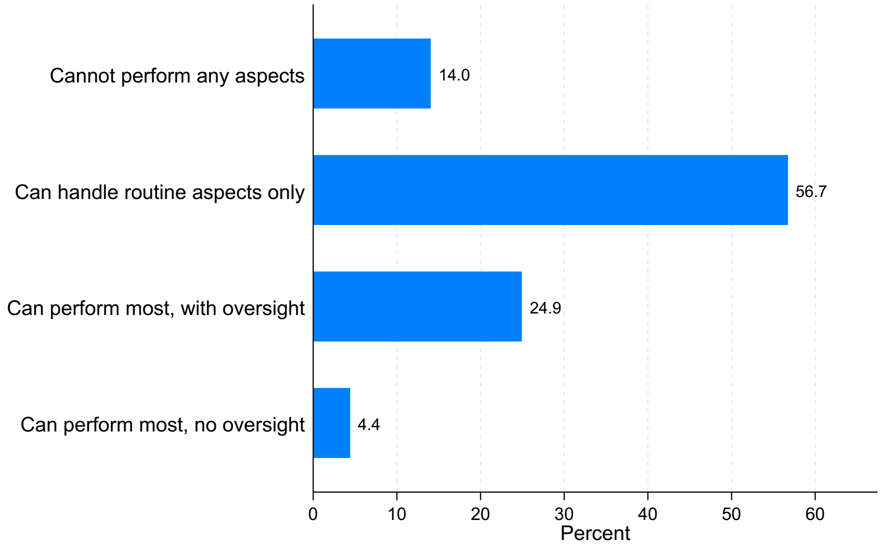
A. Uncodified knowledge by within-org relative performance

B. Uncodified knowledge by education level

C. Uncodified knowledge by similar-work cohort size

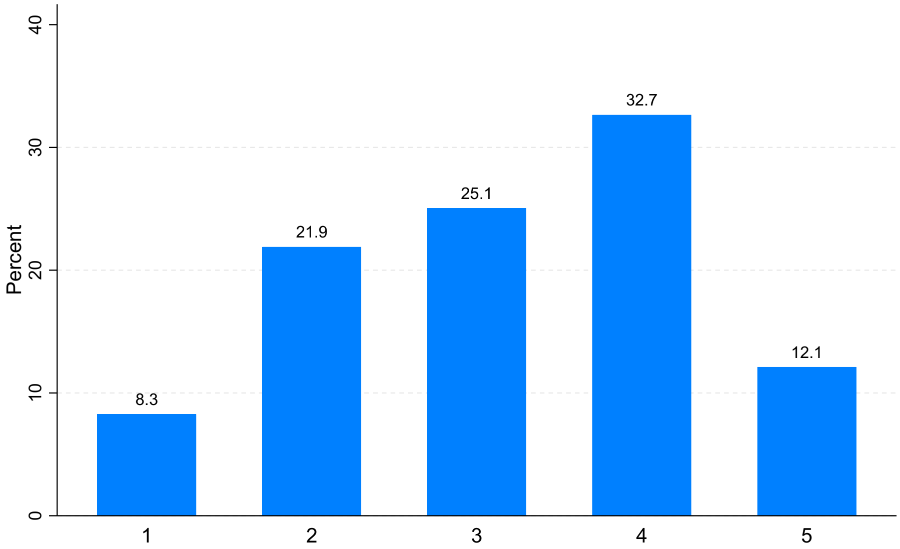
Notes: Panel A through C plots heterogeneity in the share of core skills mostly uncodified. This share is the average of dummies constructed for each of nine work-relevant subdomains in Figure 1 if respondents report "some" or "a lot" of uncodified knowledge for each subdomain that forms a core part of a workers' job. For Panel A, within-org relative performance represents workers' self-reported relative ranking among all workers within their firm on a 100 point scale where 100 (0) indicates highest (lowest) performers, and 50 represents the median performer. For Panel C, cohort size represents the self-reported number of workers in a respondents' firm performing a similar role as the respondent.

Figure A4: Beliefs about AI Impact on Own Job



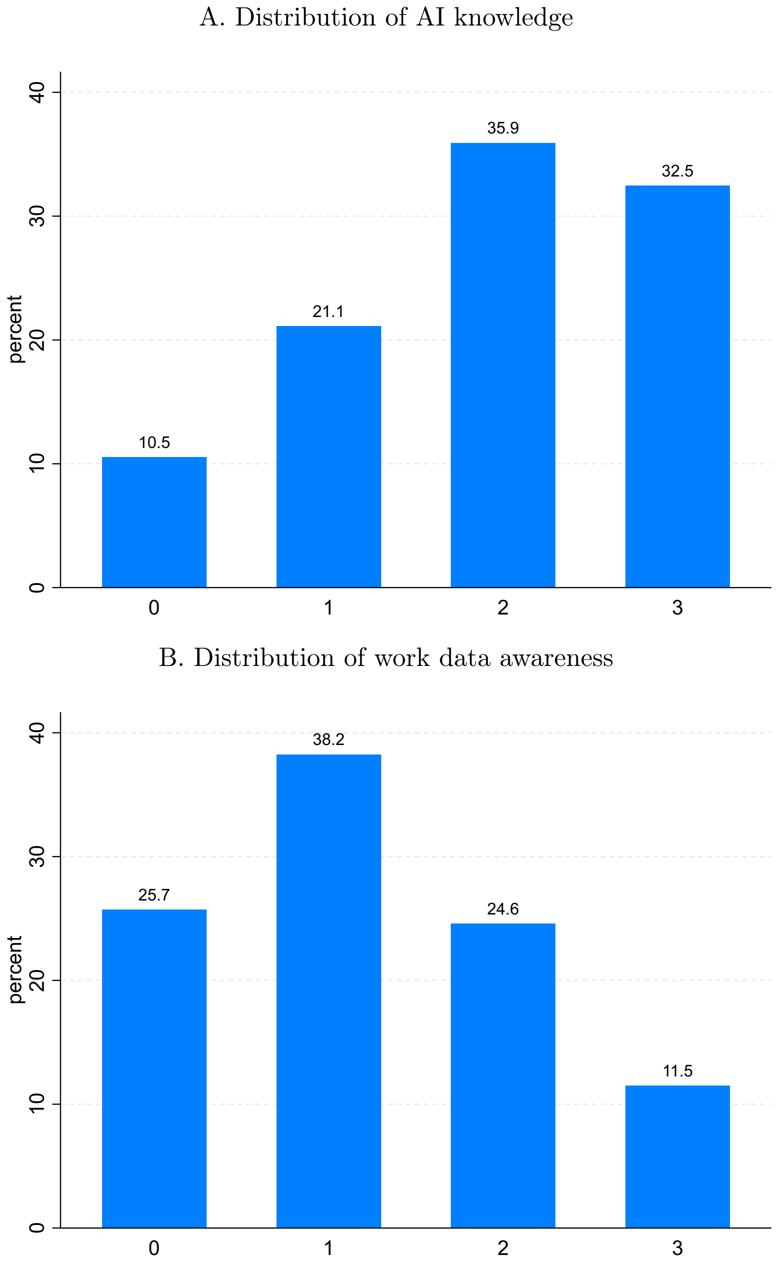
Notes: This figure describes respondents' beliefs about how AI will impact their job in the coming 5 years. Respondents were asked "Which statement most accurately reflects your beliefs about how AI models may impact your current job in the next 5 years?"

Figure A5: Familiarity with the Development of AI Tools



Notes: This figure describes respondents' self-reported familiarity with the process by which AI tools are developed. Respondents were asked "On a scale of 1 to 5, how familiar are you with how AI tools are developed?" This figure plots the distribution of responses rounded to the nearest integer.)

Figure A6: Distribution of *ex ante* AI Knowledge and Awareness

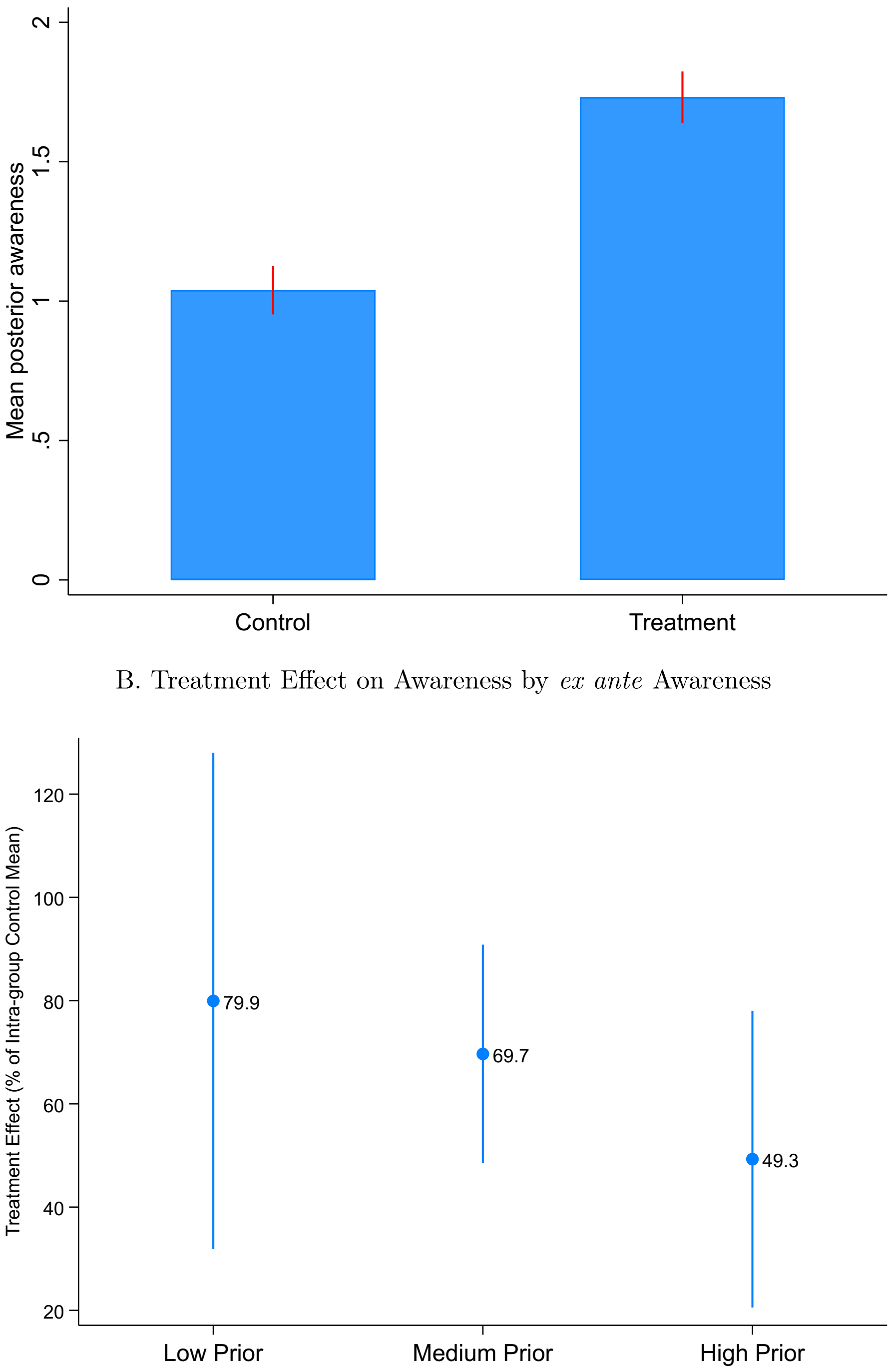


A. Distribution of AI knowledge

B. Distribution of work data awareness

Notes: This figure gives the distributions of *ex ante* AI knowledge and awareness scores. Panel A shows how many respondents got 0, 1, 2, and 3 questions correct out of the three-question knowledge assessment. Panel B shows the same but for the awareness scores.

Figure A7: Treatment Effect on Awareness Score (Prolific work sample)



A. Post-treatment Awareness

B. Treatment Effect on Awareness by *ex ante* Awareness

Notes: Panel A gives the average number of correct responses, within treatment and control groups, to three the post-treatment scenario questions measuring respondents' awareness of the role of human-generated data in training AI models. Specifically, respondents were asked questions related to training an AI model intended for coding, a model intended for legal work, and a model related to warehouse tasks. Panel B breaks down the treatment effect on awareness by *ex ante* awareness. Specifically, the sample was split into three groups: those with low prior awareness scores (score 0), those with medium prior awareness scores (scores 1 and 2), and those with high prior awareness score (score 3). We give treatment effects on awareness as percentage change over the mean score within each subset of the control group (low, medium, and high priors). In the notation of equation 1, we describe $(T_i/\alpha) \times 100$ within each group. Confidence intervals are drawn at 95%.

B Experimental Language

B.1 Knowledge and Awareness Assessment Questions

Below are the questions used to assess respondents' knowledge and pre- and post-treamtent awareness.

B.1.1 Knowledge

1. *What is the key difference between how traditional computer programs and how AI models are built?*

   - *In traditional programming, computers learn from experience, while AI is manually coded*
   - *Traditional programming involves manually coding every rule, whereas AI training allows the system to learn patterns from data*
   - *AI models require much more code to run than traditional computer programs because their architecture is more complicated*
   - *Don't know / Unsure*

2. *In the context of AI, what does the term "training" mean?*

   - *The process of programming an AI model with rules that serve as its initial default when it encounters new problems or questions.*
   - *The process of exposing an AI model to examples so it can learn patterns and relationships from data*
   - *The period when the model is released for trial use in order to collect information on what types of mistakes it makes*
   - *Don't know / Unsure*

3. *What do you believe happens when AI models make mistakes?*

   - *Developers identify where the errors come from, and re-write specific portions of the code to fix the error.*
   - *Human reviewers provide examples of correct responses to improve the model.*
   - *The model runs a self-diagnostic process to identify the faulty reasoning pattern*
   - *Don't know / Unsure*

B.1.2 Baseline Awareness

1. *What do you believe is the MOST important factor enabling recent advances in AI capabilities, such as realistic text generation and image creation?*

   – *Access to increasingly powerful computers and specialized AI hardware*

   – *Improvements in algorithmic efficiency for processing data*

   – *Access to large datasets of human-generated text and images*

   – *Don't know / Unsure*

2. *Which type of information is MOST essential for training AI systems to accurately diagnose medical conditions in real patients?*

   – *Medical textbooks that explain common symptoms and standard diagnostic procedures.*

   – *Large scale electronic records containing information on patients' medical procedures and insurance coverage.*

   – *Patient cases that include diagnostic reasoning provided by experience doctors.*

   – *Don't know / Unsure*

3. *What development process is MOST important for building an AI model to successfully resolve customer billing disputes?*

   – *Provide the AI system with recordings of conversations conducted by expert customer service agents.*

   – *Analyze past customer conversations to identify the responses that lead to the fastest call handle times.*

   – *Input company policies and procedures directly into the AI system so that it will never provide an incorrect response on an important topic.*

   – *Don't know / Unsure*

B.1.3 Post-Treatment Awareness

1. *Consider an AI model that assists software developers in writing code. What type of input is MOST useful for ensuring that the AI model follows best practices for naming code elements, such as variables or functions?*

  - *A set of pre-programmed rules and logic statements defining standard naming conventions and their intended purposes.*
  - *A dataset of millions of lines of code gathered from the Internet, used to identify the most frequently used naming patterns.*
  - *Examples of code written by the most experienced programmers, containing names that those programmers believe are most appropriate.*
  - *Don't know / Not sure*

2. *Imagine you are building an AI model to help write legal documents. Which type of input would be MOST valuable for ensuring that the AI model produces legally valid content?*

   - *A dataset of legal agreements from various industries, used to identify the most frequently used clauses and formats.*
   - *Official legal drafting guides and regulatory standards outlining proper contract structures.*
   - *Legally vetted documents created by experienced attorneys, reflecting expert judgment on enforceability and compliance.*
   - *Don't know / Not sure*

3. *An AI-powered robot is being trained to perform warehouse picking tasks, such as identifying, grabbing, and sorting items. What kind of input would be MOST helpful for teaching the robot to complete the task correctly and efficiently?*

   - *Instructional manuals and safety guidelines describing ideal warehouse procedures.*
   - *Video recordings of experienced warehouse workers performing item selection, handling, and placement.*
   - *A comprehensive dataset tracking the location of all workers in the warehouse as they perform their job tasks.*
   - *Don't know / Not sure*

B.2 Control Video Script

Artificial intelligence models are becoming increasingly sophisticated. They can talk to customers, write computer code, or even design a product. If you do things like these as part of your job, you probably spent years learning your skills – through education, *training, or learning from colleagues. Recent advances in computer science have taught AI models how to do some of the same tasks.*

Across industries, AI models are becoming integrated into daily workflows. For example, AI-powered chat assistants can help manage difficult customer service conversations by parsing questions and suggesting tailored replies. New office automation tools can perform common office tasks like submitting expense reports without human input. AI systems can even create the slide decks consultants use to present to clients, gathering relevant data, organizing narrative flows, and applying polished visual layouts. In other words, AI models may be able to replicate some of the tasks you do.

AI models have important properties you should understand.

- *First, AI models can be smart in ways that no individual human being can be. For example, AI agents can talk to customers in any language, unlike you.*
- *Second, AI models aren't constrained by human limitations: they don't require sleep, they don't get tired, and they don't need breaks.*
- *Finally, AI models are easy to copy and scale. A single AI model can be deployed to thousands of locations worldwide, all at once.*

These aren't just theoretical issues. As AI plays an ever-bigger role in the workplace, understanding what it can do will be important for your professional future.

B.3 Treatment Video Script

Artificial Intelligence models are becoming increasingly sophisticated. They can talk to customers, write computer code, or even design a product. If you do things like these as part of your job, you probably spent years learning your skills – through education, training, or learning from colleagues. Recent advances in computer science have taught AI models how to use your data to acquire some of your skills.

Across industries, employers routinely gather data on your tasks –call recordings, mouse clicks, keystrokes, code samples, and more. Employers own this data and can use it now just to monitor your productivity, but also to train AI systems to do some of the same work that you do.

For example, AI models can study recordings of customer service conversations to learn how the best workers handle difficult customers, and then copy their people management skills on new customer calls. AI models can also analyze screen recordings to observe the mouse and keyboard inputs a worker uses to file an expense report– and then use this information to automate the task. AI models can even examine how a consultant makes slides when presenting to clients, and learn how to produce new presentations using that person's style. In other words, the data you produce every day can be used to teach AI models how to replicate some of your skills.

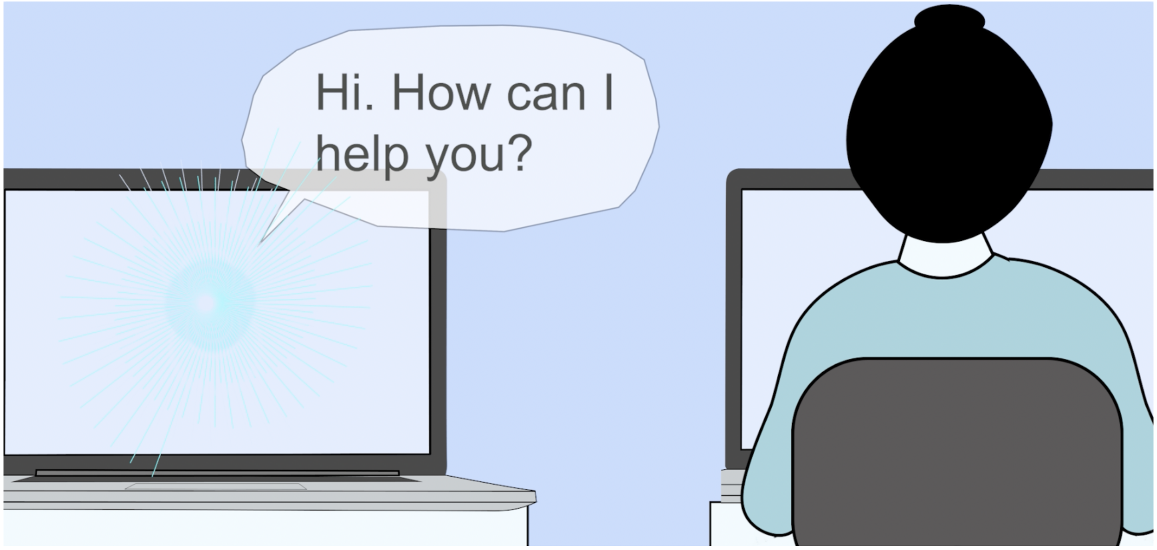
AI models have unique properties you should understand.

- *First, AI are smart in ways that no individual human being can be. For example, if you are an excellent salesperson, an AI model trained on your conversations can mimic your problem solving skills, but it can also talk to customers in any language, unlike you.*
- *Second, AI models do not have the same types of limitations that human workers have. Once AI models have used your data to acquire your expertise, it can continue using your expertise, even after you leave your job. They don't require sleep, don't get tired, and don't need time off.*
- *Finally, AI models are easy to copy and scale up. A single AI model trained on your work data could be deployed to thousands of locations across the world, all at once. This means that whoever owns your AI model can replicate and share your specialized skills with whomever they choose.*

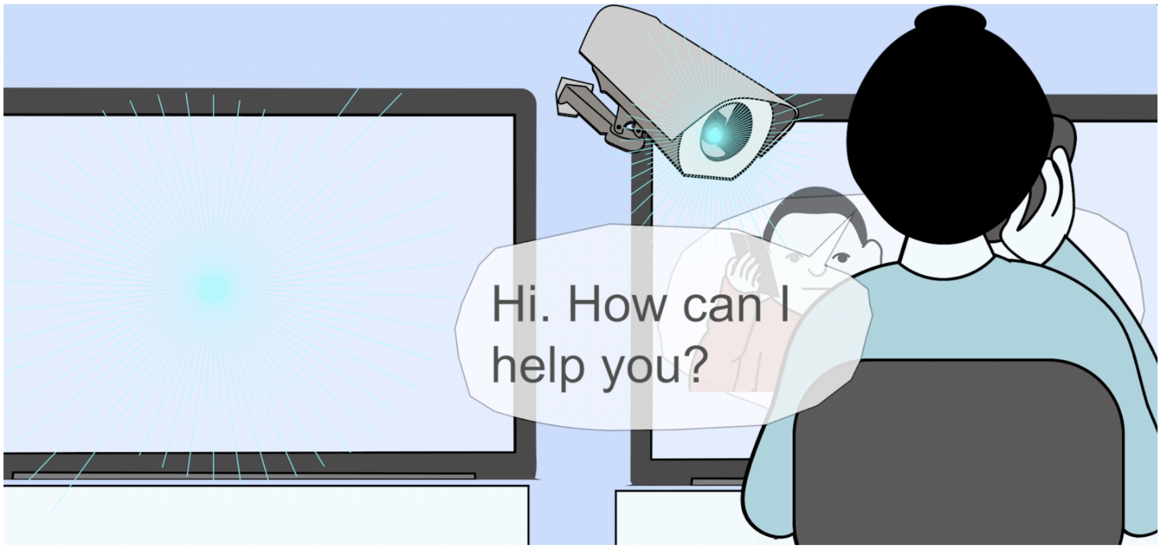
These aren't just theoretical issues. As AI plays an ever-bigger role in the workplace, understanding how it learns will be important for your professional future.

B.4 Video Screenshots

A. Control Video



B. Treatment Video



B.5 Policy Vignette Language

B.5.1 No Monitoring of Work Activity

Imagine a policy that forbids your employer from monitoring or storing any data about your individual work activities—except when strictly required for safety or legal compliance. This policy would not prevent you from using workplace tools, but it would prevent your employer from recording, archiving, or analyzing your workplace activities.

For example:

- *Your employer would have no right to intercept, archive, or analyze your workplace communications.*
- *Your employer could not track your physical location, or make video, audio, or screen recordings of your work.*

B.5.2 No AI Automation of Core Job Tasks

Imagine a policy that forbids your employer from developing or adopting AI models to automate core parts of your current job.

For example:

- *If you work in customer service, it would bar your employer from deploying AI chatbots or virtual assistants to handle customer calls.*
- *If you're an office administrator, it would stop your employer from automating tasks like filing expense reports or scheduling meetings.*

This restriction applies only to what your employer can do—it would not prevent you from choosing to use AI tools on your own.

B.5.3 No Use of Work Data for AI Development

Imagine a policy that bans the use of "work data" for AI model development. By "work data," we mean:

- *The materials you produce on the job—reports, presentations, code, designs, project plans, marketing assets, and so on.*

  - *Any record of how you work—emails and chat logs, screen or video recordings, and logs of your digital or physical activities.*

Under this policy:

1. *Employers could not use work data to develop AI models, or sell work data to other firms to develop AI models.*
2. *Those seeking to develop AI models would have to hire workers to specifically produce AI training data; they could not use data that was produced as the byproduct of everyday work tasks.*

C Occupational Exposure to AI

This appendix details the construction and integration of the measure of occupational exposure employed in our empirical work.

The exposure metric originates from Eloundou et al.'s working paper and is grounded in a simple question: can access to a modern large-language model trim the time required to perform a task by at least fifty percent while preserving quality Eloundou et al. (2023)? Each O*NET task to one of three mutually exclusive categories. E0 covers activities for which the model yields no meaningful speed-up; E1 flags tasks where that reduction is attainable directly through an interface such as ChatGPT; E2 applies when the benchmark could be met only after a lightweight application orchestrates the model against domain data.

Eloundou et al. aggregate task-level labels to occupations by using "core" versus "supplemental" flags in O*NET. Tasks designated as core receive twice the weight of supplemental tasks, so that an occupation is defined by its central duties rather than its peripheral errands. Because the weights on tasks within each occupation sum to one, each exposure score can be interpreted as a share of weighted tasks exposed to AI. We retain these weights unchanged and compute, for every six-digit SOC, the weighted proportions of E1 and E2. Those two numbers for the building blocks for three continuous exposure indices|α, β, and γ|each corresponding to a different assumption about complementary investment. The aggregation prevents occupations with lengthy task lists from dominating those whose duties are more compact, and it aligns the exposure taxonomy with our survey microdata.

The α variant isolates "pure" interface exposure. Formally, α equals the weighted fraction of tasks tagged E1 and therefore captures gains available the instant workers can query an LLM. Hence α can be thought of as a lower bound on the share of an occupation's activities that are immediately susceptible to acceleration. Because α gives zero weight to E2, it excludes any tasks that requires even modest tool building, making it intentionally conservative. The β variant computes $\beta = \text{E1} + 0.5 \times \text{E2}$, down-weighting indirect exposure to capture the view that lightweight integrations (APIs, RAG systems, etc.) arrive swiftly but not instantaneously.

Our analysis uses the third variant $\gamma = \text{E1} + \text{E2}$, which awards full credit to indirectly exposed tasks. This scenario corresponds to rapid diffusion of purpose-built software and therefore produces the largest exposure levels|estimates of γ average 0.46. Conceptually, γ should be interpreted as an upper bound on short-run productivity gains: it assumes not only frictionless access to frontier

LLMs but also rapid roll-out of complementary software. Our analysis uses γ as an ordinal number, indicating occupations' (and therefore respondents') relative exposure to AI, and is not sensitive to absolute scale.

Exposure scores are assigned to survey respondents by matching respondents' occupations to SOC classifications and assigning the correct exposure. Figure **??** shows the distribution of exposure scores within our sample.