CAN ENGINEERED COLLABORATION OVERCOME THE DELEGATION PARADOX? A PRODUCTION-FUNCTION FRAMEWORK FOR HUMAN–AI DIAGNOSTICS

Akira Murata
Juntendo University
Faculty of International Liberal Arts
a.murata.hf@juntendo.ac.jp
February 2026

Prepared for the NBER Conference on Applications of Artificial Intelligence in Healthcare
Cambridge, MA • May 7–8, 2026

PRELIMINARY AND INCOMPLETE — PLEASE DO NOT CITE WITHOUT PERMISSION

Abstract

Regulatory frameworks for medical AI assume that human–AI teams outperform either agent alone. Recent evidence contradicts this assumption. In diagnostic radiology, physicians using AI assistance perform no better than—and sometimes worse than—standalone algorithms. Agarwal et al. (2024) formalize this as the "delegation paradox," tracing it to systematic failures in Bayesian updating: automation neglect and signal dependence neglect.

This paper proposes a structural intervention. We model collaboration as an endogenous factor of production, introducing a collaborative production function with a synergy term governed by a collaboration-efficiency parameter (η) that responds to training interventions. We define causal estimands for super-additive diagnostic synergy and propose Quality-Adjusted Diagnostic Throughput (QAT) as the primary endpoint, translating clinical accuracy into productivity terms. Our minimum detectable effect—a 2.6 percentage-point improvement in

QAT—corresponds to freeing 8–12 percent of radiologist capacity at constant quality, equivalent to 1.0–1.4 additional FTE radiologists per 100,000 studies.

To identify these parameters, we design a four-arm within-reader experiment using retrospective, de-identified diagnostic cases. The key innovation is a structured cognitive friction protocol (Arm T4) that mandates pre-AI impression recording, calibration training, audit checklists, and escalation rules for discordant cases. Power calculations indicate that 500 reads per arm suffice to detect the minimum detectable effect. The paper is pre-empirical: we provide a fully specified, IRB-ready protocol that hospital systems can implement, but do not present trial data. If engineered synergy proves achievable, it would alter optimal human–AI task allocation in healthcare and reshape physician labor demand.

Keywords: human–AI collaboration, delegation paradox, diagnostic AI, collaborative production function, healthcare economics, healthcare productivity, operations management, experimental design, cognitive friction

JEL Codes: I11, J24, O33, D83, C93

1. Introduction

The integration of artificial intelligence into clinical diagnostics promises to alleviate escalating case volumes and improve healthcare productivity. Regulatory bodies—including the FDA, MHRA, and the European AI Act—presume that human oversight adds value to algorithmic predictions, placing regulatory emphasis on human–AI team performance rather than autonomous algorithm performance.[1]

Empirical evidence increasingly contradicts this presumption. Across diagnostic radiology, dermatology, pathology, and emergency triage, AI systems operating independently frequently outperform clinicians equipped with the same algorithmic assistance. In a recent randomized trial, standalone large language models achieved specialist-level diagnostic accuracy on standardized vignettes, yet physicians using the same models as decision support showed no significant improvement over conventional resources.[2] The pattern appears systematic rather than domain-specific.

Agarwal, Moehring, Rajpurkar, and Salz (2024) formalize this phenomenon as the "delegation paradox."[3] In a pre-registered experiment with professional radiologists, they show that clinicians systematically fail to realize gains from AI assistance due to two Bayesian updating errors. Automation neglect leads clinicians to underweight AI predictions relative to the optimum. Signal dependence neglect leads them to treat their own visual assessment and the AI prediction as independent, when both derive from the same underlying image. Under the estimated behavioral parameters, optimal delegation rules frequently assign cases entirely to the AI or entirely to the human, with few cases benefiting from joint decision-making.

This paper treats the delegation paradox not as an immutable constraint but as an engineering problem. We design a structured collaboration protocol for real hospital workflows and specify an experimental test of whether engineered collaboration capital can overcome the behavioral failures documented by Agarwal et al. The paper is pre-empirical: our goal is to solve

the design and identification problem, providing a blueprint for subsequent retrospective and prospective trials.

Contributions

This paper makes three contributions.

First, we formalize human–AI collaboration as a collaborative production function with an endogenous collaboration-efficiency parameter (η), distinct from human effort and AI capability. Unlike standard productivity-shifter or task-based automation models, our framework treats collaboration capital as a separate production input that must be deliberately constructed through training and workflow design.

Second, we define causal estimands for super-additive diagnostic synergy and introduce Quality-Adjusted Diagnostic Throughput (QAT), a productivity metric that maps diagnostic outcomes into hospital-level economics. QAT enables administrators to compare collaboration protocols in terms of operational capacity and economic value, not clinical accuracy alone.

Third, we design a four-arm within-reader experiment with a structured cognitive friction protocol (T4) capable of identifying whether engineered collaboration can overcome the delegation paradox. The protocol mandates pre-AI impression recording, calibration training, audit checklists, and escalation rules—features that hospital systems and regulators can operationalize in procurement standards and post-market surveillance. We derive power calculations and identification conditions that make the protocol ready for implementation.

Positioning Relative to Agarwal et al. (2024)

Agarwal et al. document the delegation paradox under existing workflows. Our contribution is to propose and analyze a structural intervention designed to shift η upward.[3] Where their finding implies that current deployment strategies often fail, our framework provides a roadmap for engineering the collaboration capital necessary to make human–AI teams super-additive. Success

would change the policy prescription from “automate or don’t” to “automate with structured human collaboration.”

The paper connects to several strands of the economics literature. It extends task-based models (Acemoglu and Restrepo 2018; Autor 2024) by introducing collaboration as a production input distinct from standard substitution and complementarity.[4] It provides structural foundations for the skill compression hypothesis documented by Brynjolfsson, Li, and Raymond (2023).[5] And it responds to the identification challenge posed by Mullainathan and Obermeyer (2022), whose machine learning approach to diagnosing physician error revealed patterns that structured collaboration might correct.[6]

The paper proceeds as follows. Section 2 develops the theoretical framework. Section 3 presents mathematical formulations. Section 4 describes the experimental methodology. Section 5 addresses data governance and econometric integrity. Section 6 discusses general equilibrium implications. Section 7 concludes.

2. Theoretical Framework

2.1 Limitations of Existing Models

Standard production function frameworks (Cobb-Douglas, CES) treat technology as a productivity shifter, implicitly assuming that gains from adoption are realized automatically upon deployment. Task-based automation models, notably the Acemoglu-Restrepo framework, represent AI as a substitute for human labor in specific task domains, with the key margin being task allocation between humans and machines.[7]

Neither approach captures the delegation paradox. The productivity-shifter model cannot explain why giving clinicians a superior diagnostic tool fails to improve their performance. Task-based models predict reallocation to the agent with comparative advantage but do not model the possibility that joint production could exceed either agent's independent capacity. The delegation paradox arises in this joint-production space, where Bayesian updating failures prevent the realization of complementary gains.

2.2 Collaboration as an Endogenous Factor of Production

We propose that the value of AI in diagnostics depends on a collaboration technology constituting a distinct factor of production, separate from clinician human capital and algorithmic capability. The mere co-presence of human and AI in a workflow is insufficient; their productive interaction must be deliberately constructed through institutional design, training protocols, and workflow architecture.

This parallels the literature on team production and organizational capital.[8] Just as firm-specific human capital requires investment, human–AI collaboration capital requires deliberate investment in calibration, trust formation, and structured interaction protocols. The delegation paradox, in this view, reflects underinvestment in collaboration capital—not fundamental incompatibility between human and algorithmic cognition.

2.3 Microfoundations: Bayesian Updating and Cognitive Friction

The behavioral mechanism underlying the delegation paradox is well characterized by Agarwal et al. (2024). Optimal Bayesian updating requires clinicians to correctly assess three quantities: the prior probability of disease, the diagnostic accuracy of the AI conditional on disease status, and the correlation between the AI signal and the clinician's own visual assessment. Automation neglect occurs when clinicians underweight AI predictions. Signal dependence neglect occurs when clinicians treat the AI output and their own assessment as independent, despite both being generated from the same image.

Our central proposition is that these updating failures are not inherent cognitive limitations but defaults that emerge absent structured interaction protocols. Structural cognitive friction—mandatory steps requiring clinicians to record independent assessments before viewing AI output, compare reasoning explicitly, and follow audit procedures for discordant cases—can shift η upward, potentially converting antagonistic interactions into super-additive ones.

2.4 Skill Compression and Labor Market Implications

This framework provides microfoundations for the skill compression hypothesis. Brynjolfsson, Li, and Raymond (2023) find that generative AI disproportionately benefits less-experienced workers, compressing the productivity distribution.[9] In our model, this operates through η: junior clinicians with less-entrenched heuristics and greater openness to algorithmic guidance may exhibit higher initial η values and steeper learning curves than senior clinicians.

We pre-specify an interaction test of T4 × clinician experience level to evaluate whether the structured intervention differentially benefits less-experienced readers. If the T4 protocol equalizes diagnostic performance across experience levels, the implications for medical wage premiums, residency training design, and returns to clinical experience are substantial.

3. Mathematical Formulations

3.1 The Collaborative Production Function

We define per-period diagnostic output for task i as:

$$Y_i = \alpha L_i + \beta A_i + \gamma S_i \quad (1)$$

where L_i denotes human clinical effort (independent diagnostic assessment), A_i denotes AI capability (independent algorithmic prediction), and S_i is the synergy term capturing additional output from collaboration. The parameters α, β, and γ are output elasticities. Super-additivity holds when $\gamma S_i > 0$.

The synergy term is a function of three factors:

$$S_i = f(\eta_i, A_i, L_i, C_i) \quad (2)$$

Here η_i is the collaboration-efficiency parameter (trust calibration, verification discipline, Bayesian updating quality) and C_i captures task context and complexity. Super-additivity is established when $\gamma S_i > 0$ at conventional significance levels.

The collaboration-efficiency parameter evolves dynamically:

$$\eta_{i,t+1} = (1-\delta)\eta_{i,t} + \kappa S_{i,t} + \tau \cdot Training_{i,t} \quad (3)$$

where δ captures depreciation (habituation decay, alert fatigue), κ captures learning-by-doing, and τ captures the direct effect of structured training. The T4 protocol is designed to shift η upward through the τ channel.

3.2 Quality-Adjusted Diagnostic Throughput (QAT)

To translate diagnostic accuracy into a hospital economics framework, we define QAT as the primary endpoint. The metric divides an accuracy-weighted utility function by the financial opportunity cost of clinician effort:

$$QAT_{it} = U_{it} / (Time_{it} \times c_k) \quad (4)$$

where c_k is the wage rate per unit of time.[10] The utility function is pre-specified as:

$$U_{it} = \mathbb{1}[correct_{it}] - \lambda^{fp}\mathbb{1}[FP_{it}] - \lambda^{fn}\mathbb{1}[FN_{it}] \quad (5)$$

The asymmetric penalty weights λ(fp) and λ(fn) are locked prior to the study via a two-round Delphi consensus survey, ensuring they reflect localized clinical risk profiles and are not subject to post-hoc manipulation.[11] This eliminates a significant degree of researcher freedom in outcome definition—a concern central to the credibility revolution.

3.3 Synergy Estimands

The protocol computes two pre-specified synergy estimands at the read-instance level.[12] The first is a weighted (partial-substitution) measure:

$$S_weighted(\alpha) = Y(T3) - [\alpha Y(T1) + (1-\alpha)Y(T2)] \quad (6)$$

The second is a conservative benchmark:

$$S_conservative = Y(T3) - max\{Y(T1), Y(T2)\} \quad (7)$$

A positive S_conservative establishes super-additivity in the strong sense: the team exceeds the best individual agent. A negative value signifies antagonism, consistent with the delegation paradox. The effect of the T4 intervention on η is evaluated via:

$$\Delta\eta = Y(T4) - Y(T3) \quad (8)$$

Numerical Example

Suppose in 100 diagnostic cases:

- Y(T1) = 90% (AI-only accuracy)
- Y(T2) = 85% (human-only accuracy)
- Y(T3) = 87% (standard collaboration)
- Y(T4) = 94% (enhanced collaboration)

With α = 0.5, weighted synergy for T3 is 0.87 – 0.875 = –0.005 (antagonism). Conservative synergy for T3 is 0.87 – 0.90 = –0.03 (team underperforms best agent). The intervention effect is 0.94 – 0.87 = 0.07 (7 percentage points). Under T4, conservative synergy becomes 0.94 – 0.90 = +0.04: super-additivity achieved. The T4 protocol converts antagonism into genuine complementarity.

3.4 Notation Summary

Table 1 summarizes the key notation.

Symbol	Definition
L_i	Human clinical effort (independent diagnostic assessment)
A_i	AI capability (independent algorithmic prediction)
S_i	Synergy term (extra output from human–AI collaboration)
η_i	Collaboration-efficiency parameter
C_i	Task context and complexity
QAT	Quality-Adjusted Diagnostic Throughput
U_{it}	Utility function (accuracy-weighted, asymmetric error penalties)
$\lambda(fp)$, $\lambda(fn)$	Penalty weights for false positives/negatives (Delphi-calibrated)
c_k	Clinician wage rate per unit of time
Y(T1)–Y(T4)	Output from AI-only, human-only, standard, and enhanced collaboration arms
$\Delta\eta$	Treatment effect of T4 on collaboration efficiency
S_weighted	Weighted synergy measure (partial-substitution benchmark)
S_conservative	Conservative synergy measure (exceeds best individual agent)

4. Experimental Methodology and Identification Strategy

4.1 Design Overview

We propose a four-arm experimental reading study using historical, fully de-identified diagnostic cases. The study employs within-reader randomization: each clinician reads cases under all four conditions, with cases randomly assigned to arms. This design eliminates between-reader heterogeneity and enables clean estimation of within-reader treatment effects.[13]

Power calculations based on variability estimates from Agarwal et al. (2024) indicate that detecting a 2.6 percentage-point QAT improvement is achievable with approximately 500 tasks per arm, assuming 80 percent power at the 5 percent significance level with 30 readers.

4.2 Arm Definitions

T1 (AI-Only Benchmark). Offline AI inference evaluated against ground truth, establishing the algorithmic baseline Y(T1) corresponding to βA_i. No human involvement.

T2 (Human-Only Baseline). Standard-of-care clinical reading without AI assistance, establishing Y(T2) corresponding to αL_i.

T3 (Standard Collaboration). Clinicians record a pre-AI impression, view AI output, and revise their assessment. This arm captures human–AI interaction as typically deployed—the condition under which the delegation paradox has been documented.

T4 (Enhanced Collaboration). The primary intervention. T3 is augmented with: (i) mandatory calibration training using practice cases with known ground truth; (ii) standardized audit checklists requiring explicit documentation of reasoning for discordant cases; and (iii) escalation rules mandating peer review when pre-specified discordance criteria are met.[14]

T4 targets signal dependence neglect by forcing clinicians to articulate independent reasoning before viewing AI output, interrupting the heuristic of treating the two information sources as independent. The audit checklist requires effortful engagement with the AI's reasoning rather than passive acceptance or reflexive dismissal.

4.3 Econometric Specification

The analysis follows an intention-to-treat framework. The primary regression specification is: [15]

$$QAT_{ir} = \beta_0 + \beta_1 T3_i + \beta_2 T4_i + X_i'\delta + \alpha_r + \theta_i + \varepsilon_{ir} \quad (9)$$

where α_r and θ_i are reader and case fixed effects, X_i is a vector of case-level controls (imaging modality, pathology type, complexity score), and standard errors are clustered at the reader level. The coefficients β_1 and β_2 identify average treatment effects of standard and enhanced collaboration relative to the human-only baseline. The contrast $\beta_2 - \beta_1$ estimates $\Delta\eta$.

We mandate Benjamini-Hochberg adjustments for secondary outcomes, including subgroup analyses by clinician experience, case complexity, and AI confidence. Pre-registration locks the outcome hierarchy before data collection.

4.4 Identification Assumptions and Threats

Causal identification relies on two assumptions: balanced case difficulty across arms (ensured by random assignment) and absence of carryover effects across arms. We address the latter through randomized arm ordering, mandatory washout intervals, and pre-specified tests for order effects.

The principal threat to external validity is the Hawthorne effect: clinicians may perform differently in experimental settings than in routine practice. We discuss this in Section 6.

4.5 Implementation Roadmap

The protocol follows a two-phase strategy.

Phase I (Retrospective/Offline). The four-arm protocol operates in a controlled environment using de-identified cases, isolating the collaboration mechanism in partial equilibrium. This phase enables clean identification of η and synergy estimands, power calibration, and institutional feasibility testing. This paper presents the complete Phase I specification. The protocol is IRB-ready and can be implemented by any hospital system with retrospective imaging data and a frozen AI model.

Phase II (Prospective Clinical Trial). Conditional on Phase I demonstrating positive $\Delta\eta$ and super-additive synergy under T4, the next step embeds the protocol into live clinical workflows via a staggered rollout with radiologist-shift pair randomization, 12-month duration, and quarterly interim analyses. Phase II design is beyond this paper's scope but represents the necessary step from partial to general equilibrium evaluation.

5. Data Governance and Econometric Integrity

Healthcare AI research requires alignment with regulatory standards (HIPAA, GDPR) and emerging evaluation frameworks (DECIDE-AI, STARD-AI).[16] For this economic evaluation, data governance is not merely a compliance requirement but a component of the identification strategy.

The study operates in a retrospective, offline, non-clinical environment. An institutional Honest Broker achieves full de-identification before cases enter the experimental pipeline, and no algorithmic outputs are written to the live EMR. This isolation matters econometrically for two reasons. First, it guarantees exogenous assignment of the AI treatment—predictions are generated from the same fixed image data available to each reader. Second, it prevents AI outputs from contaminating future ground-truth data, avoiding a feedback loop that would undermine outcome validity.

The protocol mandates continuous interaction telemetry—capturing pre-AI impressions, dwell times, override behaviors, and checklist completion—without collecting patient identifiers. This telemetry enables decomposition of the treatment effect into automation neglect reduction versus signal dependence neglect reduction, and supports monitoring for habituation decay as the cognitive friction in T4 potentially devolves into perfunctory compliance.

AI model weights are frozen for the duration of the study. Any externally triggered model updates require a complete restart of the affected experimental block, ensuring that η and synergy estimands are not confounded by algorithmic improvements.

6. Discussion

6.1 External Validity and Phase II

The offline design isolates the collaboration mechanism in partial equilibrium but strips away clinical liability, time pressure, and workflow interruptions. The estimated parameters therefore serve as bounds to inform a subsequent prospective trial. Phase II would embed T3 and T4 into live PACS workflows using a staggered rollout with radiologist-shift pair randomization over 12 months, testing whether offline η gains translate to sustained improvements under real clinical conditions.

6.2 General Equilibrium and Welfare Implications

If the T4 protocol produces super-additive synergy, the implications extend beyond partial equilibrium workflow adjustments. The current efficiency logic prescribes delegating cases entirely to AI or entirely to humans.[17] Demonstrated complementarity would shift optimal system design away from this binary.

The welfare effects operate through several channels. First, aggregate diagnostic capacity: super-additive teams increase the effective supply of diagnostic services without proportional increases in physician headcount, addressing projected shortages in radiology and pathology. A back-of-envelope calculation illustrates the magnitude. Consider a hospital system performing 100,000 radiology studies annually with 12 FTE radiologists at $207/hour.[18] A 2.6 percentage-point QAT improvement—our minimum detectable effect—frees 8–12 percent of radiologist capacity at constant quality, equivalent to 1.0–1.4 virtual FTEs and $430K–$600K in avoided recruitment costs.

Second, physician labor demand: demonstrating that humans and AI are complements under structured protocols alters the labor demand curve for diagnostic physicians, creating sustained demand for human judgment conditional on collaboration capital investment.[19]

Third, skill compression: if T4 equalizes accuracy across experience levels, it compresses the skill premium in radiology, with implications for staffing models and geographic equity.[20] Rural facilities struggling to recruit experienced radiologists could achieve comparable quality through AI-augmented junior clinicians under high-η protocols. Fourth, healthcare costs: the QAT framework enables direct cost-effectiveness comparison across protocols. Sahni et al. (2023) estimate AI adoption could save $200–$360 billion annually in U.S. healthcare spending; our framework provides a method for measuring optimized collaboration's contribution.[21] Fifth, algorithmic equity: our design enables subgroup analysis of synergy estimands by patient demographics, testing whether the protocol mitigates or exacerbates disparities in algorithmic performance.[22]

6.3 Regulatory Implications

Current regulations presume the value of human oversight but provide no guidance on how that oversight should be structured. If T4 demonstrates super-additivity while T3 shows antagonism, this provides an empirical basis for mandating structured collaboration protocols as a condition of regulatory approval.

The collaboration-efficiency parameter η could serve as a regulatory benchmark, analogous to how sensitivity and specificity benchmark standalone devices. Possible pathways include: pre-market approval requiring minimum η thresholds in teaming studies; post-market surveillance of QAT metrics; reimbursement rates tied to documented QAT improvements; and procurement standards requiring T4-compatible workflows.

Hospital administrators can operationalize findings through residency curriculum integration, credentialing requirements, real-time QAT dashboards, and incentive alignment linking compensation to documented η improvements.

7. Conclusion

The delegation paradox—the finding that clinicians equipped with AI assistance often perform no better than either agent independently—challenges the assumptions underlying medical AI regulation and deployment. This paper proposes treating human–AI collaboration as a distinct, endogenous factor of production that must be deliberately engineered.

We develop a collaborative production function with a measurable synergy term governed by η, define causal estimands for super-additive synergy, and propose QAT as a primary endpoint that translates clinical accuracy into economic productivity. Our four-arm experiment, centered on a structured cognitive friction protocol, provides a clean identification strategy for the causal effect of collaboration engineering on diagnostic synergy.

The paper's objective is to specify what a credible test requires: how to define η and QAT operationally, how to design a trial that identifies super-additive synergy, and what institutional features must be in place. We see the paper as an economic and experimental design standard for evaluating human–AI teams in healthcare, complementing clinical reporting standards such as DECIDE-AI and STARD-AI.

If the structured protocol succeeds in generating super-additivity, the implications extend beyond radiology to any domain where AI augments expert judgment—pathology, emergency medicine, mental health, and beyond healthcare to legal analysis and financial auditing. The delegation paradox may prove to be a solvable engineering problem, contingent on appropriate collaboration institutions. Taking this blueprint to data is the next step.

References

Acemoglu, D., & Restrepo, P. (2018). The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. American Economic Review, 108(6): 1488–1542.

Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2024). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. NBER Working Paper No. 31422.

Alchian, A. A., & Demsetz, H. (1972). Production, Information Costs, and Economic Organization. American Economic Review, 62(5): 777–795.

Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Angrist, J. D., & Pischke, J.-S. (2009). Mostly Harmless Econometrics. Princeton University Press.

Athey, S., & Imbens, G. W. (2017). The Econometrics of Randomized Experiments. In Handbook of Economic Field Experiments, Vol. 1, pp. 73–140. Elsevier.

Autor, D. (2024). Applying AI to Rebuild Middle Class Jobs. NBER Working Paper No. 32140.

Bansal, G., Wu, T., Zhou, J., et al. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. CHI 2021.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at Work. NBER Working Paper No. 31161.

Dranove, D., & Garthwaite, C. (2022). Artificial Intelligence, the Evolution of the Healthcare Value Chain, and the Future of the Physician. NBER Working Paper No. 30607.

Food and Drug Administration (FDA) & Medicines and Healthcare products Regulatory Agency (MHRA). (2021). Good Machine Learning Practice for Medical Device Development: Guiding Principles.

Goh, E., Gallo, R., Hom, J., et al. (2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Network Open, 7(10): e2440969.

Haider, S. A., et al. (2025). Novel Interaction Effects Metrics for Additive Augmentation. SSRN Electronic Journal. Pre-print.

Janke, S. M., et al. (2024). Data Scientists' and Clinicians' Assessments of Optimal Human-AI Teaming in the Intensive Care Unit: Delphi Study. JMIR, 26: e50130.

Mullainathan, S., & Obermeyer, Z. (2022). Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. Quarterly Journal of Economics, 137(2): 679–727.

Noy, S., & Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Science, 381(6654): 187–192.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science, 366(6464): 447–453.

Prescott, E. C., & Visscher, M. (1980). Organization Capital. Journal of Political Economy, 88(3): 446–461.

Sahni, N., Stein, G., Zemmel, R., & Cutler, D. (2023). The Potential Impact of Artificial Intelligence on Health Care Spending. NBER Working Paper No. 30857.

Vasey, B., et al. (2022). DECIDE-AI: new reporting guidelines to bridge the translational gap. Nature Medicine, 28(5): 924–933.

Vickers, A. J., & Elkin, E. B. (2006). Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Medical Decision Making, 26(6): 565–574.

Notes

[1] FDA & MHRA (2021). Good Machine Learning Practice for Medical Device Development. See also European AI Act (2024), Art. 14 on human oversight for high-risk AI systems.

[2] Goh, E., et al. (2024). Large Language Model Influence on Diagnostic Reasoning. JAMA Network Open, 7(10): e2440969.

[3] Agarwal, N., et al. (2024). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. NBER Working Paper No. 31422.

[4] Acemoglu & Restrepo (2018); Autor (2024).

[5] Brynjolfsson, Li, & Raymond (2023). See also Noy & Zhang (2023).

[6] Mullainathan & Obermeyer (2022). Quarterly Journal of Economics, 137(2): 679–727.

[7] Acemoglu & Restrepo (2018); Autor (2024).

[8] Alchian & Demsetz (1972); Prescott & Visscher (1980).

[9] Brynjolfsson, Li, & Raymond (2023). See also Noy & Zhang (2023).

[10] Wage rate calibrated from AAMC (2023) physician compensation data for diagnostic radiologists ($207/hr). Sensitivity analysis varies ±30%.

[11] Janke et al. (2024). Penalty weights calibrated via Delphi survey following Vickers & Elkin (2006).

[12] Haider et al. (2025).

[13] Athey & Imbens (2017).

[14] Amershi et al. (2019); Bansal et al. (2021).

[15] Angrist & Pischke (2009).

[16] Vasey et al. (2022); STARD-AI Steering Committee (2023).

[17] Agarwal et al. (2024).

[18] AAMC (2023) physician compensation data. See also BLS (2024) Occupational Employment and Wage Statistics.

[19] Dranove & Garthwaite (2022).

[20] Brynjolfsson, Li, & Raymond (2023). See also Noy & Zhang (2023).

[21] Sahni et al. (2023). NBER Working Paper No. 30857.

[22] Obermeyer et al. (2019). Science, 366(6464): 447–453.