Measuring AI's Economic Reach: A Multi-Dimensional Task Taxonomy

Daniel Parshall
Canary Institute
dan@canaryinstitute.ai

Andrea Lopez-Luzuriaga
Center for Economic Research
George Washington University

Abstract

Existing frameworks for measuring AI's labor market exposure decompose imperfectly across distinct dimensions: whether AI can perform a task, whether deployment is physically feasible, and whether institutions permit it. We propose CDR, a three-axis ordinal taxonomy that separates these dimensions into Cognitive complexity (C0–C4), Deployment difficulty (D0–D4), and Regulatory restrictions (R0–R4), extending Autor's (2003) routine/non-routine × cognitive/manual framework into a finer-grained classification space suitable for measuring AI exposure.

Applying CDR to the full O*NET task universe (23,850 task-activity pairs across 923 occupations, classified via multi-model LLM consensus: Claude Sonnet 4.6, GPT-5-mini, Gemini 3 Flash, validated against flagship models), we find that 40.2% of U.S. economy-weighted labor time falls in tasks that are within current AI cognitive reach (C$\leq$2, up to and including contextual judgment), require no physical infrastructure (D0), and face no professional or statutory regulatory barrier (R $<$ 2). An additional 19.6% of economy-weighted labor time is blocked by professional standards (R2: 11.5%), statutory regulation (R3: 8.0%), or moral agency requirements (R4: 0.1%). Unlike exposure measures repurposed from other frameworks, which conflate cognitive capability with deployment feasibility and regulatory permissions, the CDR taxonomy was designed from the outset to decompose these independent dimensions, avoiding the dimensional conflation that produces disagreement across existing metrics. The three axes are empirically separable and advance at different rates, with implications for how capability, deployment, and regulatory changes affect exposure estimates independently.

JEL Codes: O33, C49, J21, J22

Keywords: AI, Task Exposure, Labor Time Allocation

1 Introduction

How much of the economy is exposed to artificial intelligence? Eloundou et al. (2023) estimate that roughly 80% of the U.S. workforce could see at least 10% of their tasks affected by LLMs; Felten, Raj & Seamans (2023), using an abilities-based approach, identify a broadly similar but differently ranked set of exposed occupations. Yet when Gimbel et al. (2026) systematically compare seven leading exposure metrics, they find that agreement is weakest precisely among the most-exposed occupations. The metrics agree that something is happening; they disagree about where.

We argue that this disagreement is not primarily a measurement error problem; it is a dimensional conflation problem. Existing frameworks compress at least three independent dimensions into a single exposure score: cognitive complexity, which captures the level of reasoning and judgment the task demands, from rote lookup to expert synthesis; deployment difficulty, which captures what physical or embodiment requirements stand between AI capability and real-world task performance; and regulatory restrictions, which capture whether legal, professional, or social institutions restrict the use of AI tools for the task, even when technically capable.

These dimensions change at structurally different rates. AI capability advances rapidly: METR (2026) documents that the duration of real-world tasks AI systems can complete autonomously (autonomous task horizons) is doubling approximately every three months over the post-2024 period, pushing the frontier upward through the cognitive complexity levels of the C-axis. Physical deployment infrastructure moves at a medium pace, governed by robotics progress, system integration tooling, and platform economics. Regulatory and institutional barriers move slowest of all, constrained by legislative cycles, professional guild dynamics, and requirements that presuppose human moral agency.

A framework that collapses these three rates of change into a single number will systematically mispredict both the location and timing of economic impact.

We present the CDR framework: a three-axis, five-level ordinal taxonomy that extends Autor's (2003) routine/non-routine × cognitive/manual matrix into a finer-grained 5×5 cognitive–embodiment plane, with a third axis capturing regulatory restrictions (Section 2). We introduce a methodological innovation, Detailed Work Activity (DWA) decomposition, that resolves the compound-task problem in O*NET, where a single task description such as "diagnose and repair heating systems" bundles cognitively and physically distinct sub-activities with very different automation profiles (Section 3). We validate the taxonomy against four independent ground-truth datasets: Eloundou et al.'s human exposure ratings, Carollo's occupational licensure data, Career OneStop's licensing and certification records, and BLS fatal occupational injury rates (Section 4). We present employment-weighted coverage of the full O*NET task universe (23,850 DWA-task pairs across 923 occupations) and identify the structural features of the task landscape that determine the pace of AI economic impact (Section 5). We provide a complete replication package with version-controlled prompts and API specifications, enabling the full classification pipeline to be re-run for under $100 per three-model generation using mid-tier APIs ($7 for GPT-5-mini, $12 for Gemini 3 Flash, $65 for Claude Sonnet 4.6; approximately $200 with flagship models), making CDR a longitudinal instrument rather than a

one-time measurement (Section 7.2).

The unit of analysis throughout is the DWA-task pair: O*NET's Detailed Work Activities crosswalked to their parent tasks. A single O*NET task like "diagnose and repair heating systems" decomposes into DWAs with distinct automation profiles: the diagnostic reasoning is cognitively complex but purely digital, while the physical repair is cognitively simple but requires hands-on manipulation in unpredictable environments. (The CDR notation used throughout, e.g., C2/D0 for the diagnosis, C1/D3 for the repair, is defined in Section 2.) This decomposition, made possible by the now-complete O*NET 30.2 crosswalk, yields 23,850 classifiable units from 18,796 original tasks, preserving within-task heterogeneity that single-unit classification obscures.[1]

The theoretical motivation for decomposition comes from the weak-links intuition developed by Jones & Tonetti (2026): in production processes with limited substitutability across inputs (what we term *non-linear complementarity*), the slowest-advancing component governs the overall rate of productivity growth. In our framework, the three CDR axes could be the weak links, and depending on the task, the binding constraint on AI-assisted productivity gain may be any of the three: cognitive reach (C), deployment infrastructure (D), or institutional permission (R).

For any given task, the binding constraint is whichever axis presents the highest barrier. As the cognitive frontier advances, the binding constraint for an increasing share of tasks shifts from "can AI reason about this?" to "can we physically deploy it?" and "will institutions allow it?"; this is a structural prediction that single-dimensional frameworks cannot generate.

The empirical evidence for this shift is already visible. Massenkoff & McCrory (2026), using Anthropic's Economic Index data on millions of actual Claude conversations, find a large gap between theoretical AI capability and observed deployment: 94% of Computer & Mathematical tasks are theoretically feasible for LLMs, but Claude currently covers only 33%.[2] Their "observed exposure" measure predicts BLS employment change projections (both growth and contraction) where Eloundou et al.'s theoretical measure alone does not.

The CDR framework offers a structural account of this gap: deployment difficulty (D) and regulatory restrictions (R) operate as independent constraints on task cognitive complexity (C), and a single-axis measure cannot separately identify their contributions. The gap also partially reflects the human learning curve (users learning to prompt AI systems effectively), a dynamic captured in Handa et al.'s (2025) finding that AI usage concentration is broadening as adoption matures.

We situate this work relative to several recent contributions. Acemoglu (2024) provides a first-order estimate of AI's macroeconomic impact using a task-based model and Hulten's theorem,

[1] This paper is the first in a planned research program on AI's economic impact. The CDR measurement framework and task-level validation are presented here. Planned companion analyses include: adoption dynamics and production function modeling with heterogeneous elasticities of substitution; the F-axis (failure consequences) and its interaction with verification infrastructure; and developing-country applications where D0 tasks become internationally tradable. These companion papers are in progress.

[2] The 94% is Massenkoff & McCrory's figure, derived from Eloundou et al.'s beta measure of theoretical LLM capability. Our CDR classification of Computer & Mathematical occupations (SOC 15-xxxx) finds 90% of tasks at C ≤ 2 (within contextual judgment), or 84% at the stricter C $\leq 2 \cap$ D $= 0$ threshold. The gap between their 94% and our 90% likely reflects our finer-grained cognitive decomposition; the gap between 90% and 84% reflects the 6% of Computer & Mathematical tasks that, despite being cognitively accessible, require some physical deployment (D ≥ 1).

arriving at approximately 0.53% TFP over a decade; that framework treats AI capability as static, deployment as instantaneous, and regulation as absent, assumptions that compress three distinct timescales into one. Acemoglu, Autor & Johnson (2026) have since proposed a five-category taxonomy of AI's effects on labor that implicitly recognizes the need for dimensional decomposition.

Handa et al. (2025) provide large-scale empirical measurement of what people actually do with AI, finding that 57% of usage is augmentative rather than automating, a pattern that directly informs our treatment of the R-axis around augmentation rather than replacement. Svanberg et al. (2024) demonstrate that even "free" AI systems achieve only 49% task automation due to deployment fragmentation, which is the D-axis mechanism in operation. Gmyrek, Berg & Bescond (2023) extend the Eloundou approach to global labor markets using ISCO-08 occupational classifications and ILO employment microdata from 48 countries, finding that augmentation dominates automation across all income groups but that the effects vary dramatically by country income level (0.4% of employment faces automation in low-income countries vs. 5.5% in high-income countries). Their scalar exposure score, like Eloundou's, conflates cognitive capability with deployment feasibility; the CDR decomposition would allow their global estimates to be stratified by the dimension that drives the cross-country variation, which is primarily the D-axis (infrastructure and digital access) rather than the C-axis (cognitive reach).

The framing throughout this paper is deliberately conservative. We classify tasks within the economy as currently organized (existing occupations, firm boundaries, and institutional arrangements) and ask only where AI can provide meaningful time savings under current conditions. We do not model the economic reorganizations (firm restructuring, occupational decomposition, new task creation) that may follow as the capability-deployment frontier expands. Those dynamics, along with the aggregate productivity estimates they imply, are the subject of planned companion work (see note above).

2 The CDR Framework

2.1 Framework Design Principles

Task-based measures of AI exposure have largely inherited the structure of earlier automation frameworks, which were not designed to distinguish AI's cognitive reach from the physical and institutional barriers to its deployment. The CDR (Cognitive complexity, Deployment difficulty, Regulatory restrictions) decomposition makes these dimensions explicit. From labor economics, we inherit the insight that tasks, not occupations, are the right unit of analysis (Autor, 2003), and that the cognitive and manual dimensions of work have structurally different automation profiles. From AI research, we draw on an empirical understanding of how the technology actually advances: what improves quickly, what improves slowly, and why. From robotics, we draw on an established literature documenting which physical activities are easy and hard for automated systems to perform, a contact-complexity ordering that directly informs the D-axis gradations. In a nutshell, the C-axis ranges from simple lookup (C0) to original discovery (C4), the D-axis ranges from purely digital

tasks requiring no physical interaction (D0) to dynamic real-time coordination of perception and manipulation (D4), and the R-axis ranges from no institutional barrier (R0) to tasks requiring human moral agency (R4). Each axis is detailed in the subsections that follow.

Three observations from the AI development process motivate the three-axis decomposition.

The first observation is that AI cognitive capability is advancing rapidly and measurably. Benchmark performance, autonomous task completion, and professional exam scores all show steep improvement curves with well-characterized doubling times (Section 6.1). This is the C-axis: it tracks the cognitive frontier that determines which tasks are within AI's reasoning reach.

The second observation is a pattern that AI researchers call Moravec's paradox: tasks that humans find cognitively difficult (abstract reasoning, data analysis, strategic planning) turn out to be relatively easy for AI systems, while tasks that humans find effortless (perceiving a cluttered room, catching a thrown ball, threading a needle) remain among the hardest unsolved problems in robotics and computer vision. This paradox, first articulated by Moravec (1988), explains why cognitive complexity and physical deployment difficulty are independent dimensions rather than a single spectrum. A task can be cognitively simple but physically demanding (sorting recycling by hand: C1/D3), or cognitively complex but purely digital (drafting a legal brief: C3/D0). Collapsing these into one axis loses exactly the information needed to predict when a task becomes AI-accessible.

The third observation is that the rate at which institutions adapt to technological capability is governed by its own dynamics (legislative cycles, professional guild responses, liability evolution) that are largely independent of both cognitive and physical progress. A task may be well within AI's cognitive reach and require no physical infrastructure, yet remain inaccessible because statutory regulation restricts AI assistance in the process. These institutional barriers change, but on timescales measured in years to decades rather than months (Section 6.3).

The CDR framework makes these three rates of change independently measurable. For each axis, the sections that follow present the conceptual definition, explain how the five ordinal levels are distinguished and why the boundaries are drawn where they are, and characterize the rate of progress.

The measurement methodology (how we use multiple language models to classify 23,850 task-activity pairs across 923 occupations) is described in Section 3, which includes a plain-language explanation of the computational approach for readers unfamiliar with LLM-based classification. We classify each O*NET task along these three independent ordinal axes. Each axis has five levels, with 0 representing the lowest barrier to AI-assisted productivity gains and 4 the highest. The classification is grounded in a single economic question: *"What would it take for AI to cut the time a worker spends on this task by at least 50%?"* This framing centers augmentation rather than replacement: we ask not whether AI can do the task, but whether AI can meaningfully accelerate a human worker performing it. This is the same economic question Eloundou et al. (2023) posed. Where we differ is not in the question but in the capability envelope we evaluate against and the dimensional structure of the answer (see Section 4.1).

2.2 C-Axis: Cognitive Complexity

The C-axis measures the level of reasoning and judgment a task demands, extending the cognitive dimension of Autor's (2003) routine/non-routine framework into five gradations. The overall axis is characterized by a *manual test*: at C0, no manual is needed because what to do is self-evident from the situation; at C1, a complete procedure manual could be written and followed; at C2, a manual can provide guidelines but the worker must exercise judgment; at C3, the manual can only say "consult a specialist" because the judgment requires years of domain-specific expertise; and at C4, no manual exists because the worker is *writing* the manual — producing transferable knowledge that other practitioners can adopt.

Table 1. C-axis: Cognitive Complexity Levels

Level	Label	Description	Boundary test
C0	Self-evident	What to do is obvious from the situation. No decision points, no ambiguity. Writing a manual would be absurd.	"Is a manual even necessary?"
C1	Procedural	Following established rules, protocols, or algorithms. Skill in execution but deterministic best outcome.	"Could you write a complete procedure manual?"
C2	Contextual judgment	Requires weighing tradeoffs, adapting to context, or choosing among reasonable alternatives. Two workers may produce different but equally valid outputs.	"Would two workers produce the same output?"
C3	Expert synthesis	Integrating information across domains, recognizing novel patterns, or making high-stakes decisions under genuine uncertainty. Only a specialist can proceed; a competent generalist with a good reference cannot reach the correct answer.	"Could a generalist with a good reference handle this?"
C4	Discovery / creation	Producing transferable knowledge — a new method, theory, protocol, or framework that other practitioners can adopt. The output extends the field's capability set.	"Is the output itself a new manual that others will follow?"

The cognitive frontier advances fast, and measurably so. The pace of that advance can now be quantified: METR (2026) documents autonomous task horizons doubling approximately every three months over the post-2024 period (see Section 6.1 for detailed analysis).

To illustrate what this means at the upper end of the C-axis, consider performance on FrontierMath (Glazer et al. 2024), a benchmark of roughly 300 original, research-level mathematics problems created by 60+ professional mathematicians with verifiable correct solutions. Frontier model performance rose from under 2% at launch (November 2024) to 50% on Tier 1–3 problems (undergraduate through early postdoc level) by March 2026 (GPT-5.4 Pro; Epoch AI 2026). Tier 4 (research-level) performance reached 38%. Correct solutions require producing provably valid proofs or derivations, allowing us to be confident that models are capable of novel mathematical reasoning rather than surface pattern-matching, a performance profile consistent with C3-level capability. Terence Tao, Fields Medalist and one of the world's leading mathematicians, predicted at launch that the benchmark

would resist AI for several years. Roughly half has been solved within sixteen months.

In terms of relationship to prior frameworks, C0–C1 correspond roughly to Autor's "routine cognitive" category; C2–C4 to "non-routine cognitive." The five-level gradient captures the economically significant difference between contextual judgment (C2) and expert synthesis (C3) that Autor's binary distinction does not.

2.3 D-Axis: Deployment Difficulty

The D-axis measures the physical and embodiment requirements intrinsic to a task: what sensorimotor capability must exist for AI to assist with the task in the real world. This extends the manual dimension of Autor's (2003) framework to handle the impact of AI, grounded in a contact-complexity ordering derived from O*NET's 31 non-cognitive ability elements (9 physical, 10 psychomotor, 12 sensory), consistent with the robotics literature's established finding that task difficulty increases from navigation through structured manipulation and unstructured manipulation to dynamic multi-modal action.

The classification test is: *"What physical sensing, locomotion, or manipulation does this task inherently require?"*

Table 2. D-axis: Deployment Difficulty Levels

Level	Label	Description	Boundary test
D0	Purely digital	No physical-world interaction. Text, data, code, voice, video. All inputs and outputs are digital.	"Can this be done entirely on a computer?"
D1	Sensing / locomotion	Requires perceiving the physical environment or moving through it, but no manipulation of specific objects. Navigation, observation, monitoring.	"Does the task require perceiving or moving through a physical environment, but NOT touching specific objects?"
D2	Structured manipulation	Controlled contact with specific objects in engineered environments. Predictable geometry, fixed workstations, repeatable motions.	"Could the workspace be fully engineered?"
D3	Unstructured manipulation	Contact with objects in variable, partially unpredictable environments. Deformable materials, natural settings, novel configurations.	"Could you pause for 5 seconds without consequence?"
D4	Dynamic multimodal	Simultaneous real-time coordination of perception, locomotion, and manipulation under time pressure. The 5-second pause test fails.	"Would a 5-second pause cause failure?"

The rate of change along the D-axis is structurally heterogeneous. D0 tasks face zero deployment difficulty by definition. D1 capabilities (sensing, monitoring, locomotion) can often be implemented with commodity hardware costing a few hundred dollars: a smartphone camera, a touchscreen kiosk, or a consumer drone.

D2 tasks in structured industrial environments are in early commercial deployment: Agility Robotics' Digit achieves 98% task success at Amazon at an estimated $10–12/hour operating cost, and approximately 16,000 humanoid robots were installed globally in 2025. D3 has seen more progress than commonly recognized: Harvest CROO announced commercial-parity strawberry harvesting

in April 2025, and Waymo now operates 450,000+ weekly paid rides across 15+ cities. However, broad D3 deployment, manipulation in truly unstructured environments (construction, healthcare procedures, vehicle repair), remains years from reliable scale. D4 remains largely human-only. Notice the connection to Autor (2003). D0 maps broadly to Autor's cognitive tasks (no manual component); D2–D4 decompose his "manual" category into gradations that predict automation timeline: structured manual tasks (D2) are automatable on a 3–5 year horizon; unstructured manual tasks (D3) on a 7–15 year horizon; dynamic multi-modal tasks (D4) remain indefinite. The key difference is that where Autor's manual/cognitive distinction was descriptive, our D-axis gradations are specifically grounded in Moravec's paradox and the robotics contact-complexity ordering: what is physically easy or hard for automated systems to perform, rather than what is routine or non-routine for humans. In practical terms, multimodal AI already extends into D1 territory in economically significant ways. AI vision systems can assist with diagnosis, inspection, and assessment tasks that previously required physical co-presence.

One of the present authors has used photographs and Claude to diagnose and repair household plumbing issues without a plumber, a D1 task (sensing, no manipulation) that substitutes for what was traditionally a D3 service call. This pattern (AI providing the cognitive component remotely while the human provides only the physical manipulation) is a general mechanism by which multimodal perception compresses the effective D-level of many tasks. The plumbing example illustrates a general mechanism. Many tasks belong to an occupation not because they are physically difficult but because they require specialized knowledge: the plumber gets paid because they know which wrench to turn, and where, and why. AI removes that knowledge barrier, and once it does, some of those tasks can migrate out of the occupation to non-specialists. What remains in the occupation is only what is genuinely physically difficult, what requires more complex knowledge, or what is regulated, and the binding constraint keeps shifting as the rate of progress on each axis advances.

Three independent experiments across very different settings, customer service (Brynjolfsson, Li & Raymond 2023), professional writing (Noy & Zhang 2023), and software development (Peng et al. 2023), confirm that AI compresses the skill distribution by giving non-experts access to expert-level knowledge. A fourth experiment (Shen & Tamkin 2026) shows the flip side: people who use AI do not build the expertise themselves; they can produce correct output but do not understand it. The high-stakes cases where that gap matters (where someone needs to catch errors, where getting it wrong is dangerous) are precisely the cases that fall into the domain of the next axis: regulation.

2.4 R-Axis: Regulatory Restrictions

The R-axis classifies regulatory constraints in two steps. First, identify what constraint *exists* on who may perform the task: no restriction (R0), social norm (R1), professional standard (R2), government statute (R3), or moral agency requirement (R4). Second, for R2 and R3 tasks, apply an *augmentation binding test*: "Does this regulation or standard prevent a licensed professional from

using AI tools to *assist* with this task?" Regulations restrict *who* may perform a task; AI assistance does not change who is performing it; the licensed professional is still performing the task, with computational support. If the constraint restricts who but not how, and AI assistance is a how, the constraint may not fully bind under augmentation. A doctor using AI to help interpret diagnostic scans is still the doctor diagnosing; the statutory restriction (R3) does not bind, though professional standards about diagnostic competency may remain (downgrade to R2). A cosmetologist using AI to recommend styles still requires the license for the physical act of cutting hair; the constraint partially binds (keep R3).

Table 3. R-axis: Regulatory Restriction Levels

Level	Label	Description	Boundary test
R0	No barrier	No regulation, professional standard, or established norm restricts who may perform this task.	"Is there any rule, legal, professional, or social, against using AI here?"
R1	Social or market norm	Established professional or social norms expect a particular type of professional, but no formal standard or law requires it. Market-enforced; erodes under price pressure.	"Would a customer object, but no regulator?"
R2	Professional standards / liability	Established professional standards require a credentialed professional to perform or oversee this task. Violation risks loss of credentials, termination, or significant liability exposure. Enforced by the profession, not by statute.	"Could a practitioner face professional consequences for lacking credentials?"
R3	Statutory regulation	A government statute specifically restricts this act to licensed professionals. Performing this act without the required credential is a legal offense. After identifying the restriction, apply the augmentation binding test: does the statute prevent a licensed professional from using AI to *assist*?	"Does a statute create a specific prohibition on unlicensed performance of this act?"
R4	Moral agency required	The act derives its force from a human agent staking personal moral or legal accountability. The barrier is not regulatory but definitional: no legislative act can confer the required status on a non-human agent. Sworn testimony, judicial sentencing, jury deliberation, sacramental acts, military rules-of-engagement authorization, the giving of informed consent (the *consent* itself, not the information delivery).	"Does the act require a person, by definition?"

The R3–R4 boundary marks a qualitative shift in the nature of the barrier. R0 through R3 are contingent: they reflect rules that humans chose and could, in principle, unchoose. R1 norms erode under price pressure; R2 standards evolve with professional practice; R3 statutes can be amended by legislatures. R4 is also human-defined, but it reflects a different kind of social determination: society places specific restrictions on what counts as personhood for different purposes. Corporations, for instance, possess legal personhood sufficient for property rights but cannot give sworn testimony; *Hale v. Henkel* (1906) held that the Fifth Amendment privilege "is purely a personal privilege of the

witness." The acts classified R4 require a status that current legal and social institutions reserve for natural persons.

These are not metaphysical claims about consciousness but institutional facts about how societies allocate moral and legal agency. A sworn oath requires a person capable of being held in contempt; informed consent requires a consciousness recognized as capable of understanding risk. The boundaries of R4 could, in principle, shift, but only through the kind of deep institutional change that operates on generational timescales.[3]

The augmentation framing of the R-axis deserves particular emphasis, as it determines the interpretation of a substantial portion of the labor market. The R-axis measures barriers to AI *assisting* a licensed professional, not barriers to AI *replacing* her. This distinction is critical: under augmentation framing, most tasks in licensed occupations classify as R0–R1, because the regulation governs who signs off, not what tools the professional uses. A replacement framing would produce systematic false positives, classifying licensed professionals as high-R even when they can clearly benefit from AI tools, simply because their occupation requires licensure. Tasks that remain R3 under augmentation framing are those where regulation specifically restricts the use of AI in the process: prescribing controlled substances, performing surgery, making legally binding determinations. This framing aligns with Eloundou et al.'s (2023) original exposure question, which likewise centered on whether AI could reduce task time rather than replace the worker.

This augmentation framing has a direct employment implication: while the profession itself survives (the license remains, the occupation persists), the *number* of professionals required could change. A physician using AI to draft treatment plans, check drug interactions, and summarize patient histories can handle a larger patient panel; a paralegal using AI for document review can support more cases. The supply of effective labor per worker increases, but the net effect on employment and wages depends on supply and demand elasticities in each market: increased per-worker productivity could reduce headcount, or it could expand access (more patients served, more cases handled) if demand is sufficiently elastic. In practical terms, augmentation produces meaningful economic effects even when the regulatory barrier to full automation remains intact.

The rate of change along the R-axis varies by level, and this variation is itself a key feature of the framework. R1 barriers erode in months under sufficient price pressure; R2 barriers erode on a 3–5 year institutional cycle; R3 barriers require legislative action (5–15 years); R4 barriers are measured in decades.

The R1 level corresponds closely to what Korinek & Suh (2024) term "nostalgic jobs," roles where consumer preference for human providers persists even after AI reaches or exceeds human performance. Korinek & Suh predict these preferences erode under price pressure, consistent with our R1 erosion timeline of 1–3 years. The CDR framework extends their insight by distinguishing

[3]The question of whether a non-human entity capable of human-like action can count as a person for institutional purposes has a longer history than is commonly recognized. In halakhic law, Rabbi Tzvi Ashkenazi ruled in the early 18th century that a golem, an artificially created humanoid capable of action and speech, cannot count toward a minyan (the quorum of ten required for communal prayer), because it lacks the ontological status of a person regardless of its functional capabilities. This determination preceded AI by roughly three centuries but addresses the identical structural question.

nostalgic barriers (R1, market-enforced, fast erosion) from institutional barriers (R2, guild-enforced, medium erosion) and statutory barriers (R3, government-enforced, slow erosion).

Our R-axis classifications conservatively assume that professional guilds and statutory regulators will successfully resist consumer pressure for AI adoption. If guilds prove unable to maintain barriers against AI-assisted practice, our R-axis estimates overstate the true friction. The historical pattern of technology forcing regulatory adaptation (telemedicine expansion during COVID-19 is a recent example) suggests this assumption may prove too conservative.

2.5 Relationship to Existing Frameworks

The CDR taxonomy builds on a lineage of task-level exposure frameworks, each capturing a different slice of the automation question. The table below maps each framework's core measure to the CDR dimensions it most closely corresponds to, making explicit where prior approaches combine dimensions that CDR separates. Full definitions of the CDR and Eloundou levels used in this table appear in the footnote. The C-axis maps cleanly onto Autor's cognitive dimension: his "codifiability" criterion (can the task be fully described by rules?) corresponds directly to the CDR manual test that distinguishes C0–C1 (routine, procedural) from C2+ (non-routine, contextual judgment). The D-axis correspondence is conceptual, not empirical. Autor's "manual" category reflects *human* skill requirements, while D measures *robotic* difficulty — the sensorimotor capability needed for machine deployment. The broad mapping is D0–D1 (cognitive/digital) $\approx$ non-manual and D2–D4 (physical manipulation) $\approx$ manual, but the two can diverge sharply within the manual category: strawberry picking is routine, unskilled human labor but D3 in CDR (unstructured manipulation with narrow force margins and variable geometry per berry). Autor's framework was designed to classify human task difficulty; CDR's D-axis classifies machine deployment difficulty. The two measure different things even when they use similar language.

Table 4. CDR Relationship to Existing Exposure Frameworks

Framework	What it measures	CDR relationship
Eloundou et al. (2023) E0/E1/E2	Combined exposure	E0 ≈ D>1 | R>2; E1/E2 ≈ D≤1 & R≤2
Autor (2003) 2×2	Routine/non-routine × cognitive/manual	Routine ≈ C0–C1; non-routine ≈ C2–C4; cognitive ≈ D0–D1; manual ≈ D2–D4. D-axis mapping is conceptual, not empirical.[a]
Felten, Raj & Seamans (2023)	Abilities-based exposure	Maps to C-axis abilities; ignores D and R
Svanberg et al. (2024)	Cost-effective automation	Cost filter correlates with D-axis; hardware deployment cost scales with embodiment level
Massenkoff & McCrory (2026)	Observed vs. theoretical exposure	The gap between theoretical and observed is exactly D + R
Korinek & Suh (2024)	Nostalgic jobs / human preference	Maps to R1 (consumer norms); CDR adds R2–R4 gradations
Gmyrek, Berg & Bescond (2023)	Scalar GPT exposure (ISCO-08)	Single-score exposure conflates C, D, and R; ISCO-08 enables global coverage but at lower task granularity than O*NET

Notes: CDR Framework: C-axis (Cognitive complexity): C0 = lookup/copy, C1 = procedural, C2 = contextual judgment, C3 = expert synthesis, C4 = discovery/creation. D-axis (Deployment difficulty): D0 = purely digital, D1 = sensing/locomotion, D2 = structured manipulation, D3 = unstructured manipulation, D4 = dynamic multi-modal. R-axis (Regulatory restrictions): R0 = no barrier, R1 = client/consumer norms, R2 = professional standards/liability, R3 = statutory regulation, R4 = moral agency required. Eloundou et al. (2023): E0 = no exposure (task not affected by LLMs), E1 = exposed (task can be completed 50% faster using a text-only LLM), E2 = exposed with tools (task can be completed 50% faster using an LLM with image and retrieval capabilities).

Two additional axes, Physical Presence (P) and Failure Consequences (F), were explored and set aside during development. Their rationale, pilot results, and reasons for exclusion are reported in Appendix A.

2.6 Scope of the Framework

The CDR framework classifies tasks within the economy as currently organized. It takes existing occupations, firm boundaries, task definitions, and institutional arrangements as given, then measures where AI can provide meaningful productivity gains within that structure. It does not model the reorganizations expected as the capability-deployment frontier expands: Coasean shifts in firm boundaries, decomposition and recombination of occupations, emergence of new task categories, or restructuring of professional credentialing. Those dynamics require a different analytical apparatus and are the subject of a planned companion paper.

Notice that the framework classifies tasks, not occupations. Task-level exposure does not translate linearly to occupation-level impact. If AI dramatically accelerates a specific task but that task is not the binding constraint on the occupation's output, the occupation-level productivity gain is much

smaller than the task-level gain (the weak-links intuition applied within a single role). In practice, within-role binding constraint variance is likely low: occupations as currently organized already tend to bundle tasks with similar requirement profiles, minimizing the gap between task-level and occupation-level exposure. This may change as AI selectively accelerates some tasks and not others, creating new within-occupation heterogeneity.

3 Classification Method

3.0 Classifying Tasks with Language Models: A Primer

This section explains how language model classification works for readers unfamiliar with the methodology. Readers with experience in LLM-based measurement may skip to Section 3.1.

Our CDR (Cognitive complexity, Deployment difficulty, Regulatory restrictions) classification pipeline is a structured survey administered to multiple independent raters, where the raters are language models rather than humans. The analogy to survey methodology is useful and largely accurate, though there are important differences we flag below.

An API (Application Programming Interface) is how programs communicate with language models: the calling program sends two text inputs (the system prompt with instructions, and the data prompt with the case to classify), and the model returns a structured response. Each API call provides the model with a fresh context window; it has no memory of previous classifications. In practical terms, this is analogous to giving a survey rater a codebook (system prompt) and a case file (data prompt) and asking her to apply the codebook to the case.

The classification instrument works as follows. Each language model receives a detailed rubric (called a *system prompt*) that defines the CDR axes, their levels, the boundary tests that distinguish adjacent levels, worked examples spanning the full classification space, and disambiguation rules for known problem cases. This rubric is ~5,200 words and is identical across all occupations and all models; it functions like a structured codebook in content analysis, specifying the classification scheme and training the rater before any data is presented. The system prompt and data prompt are handed to the model together, but the model does know which is which. Because the system prompt is identical across all classifications, model providers cache it automatically, reducing per-call costs by approximately 90%.

For each occupation, the model then receives a second prompt, what we call the *data prompt* (also called the "user prompt" in API terminology), containing the specific material to classify: the occupation's narrative description from O*NET, its working conditions, cognitive and physical ability scores, licensing status, and the list of DWA-task pairs to be rated.

The model produces, for each DWA-task pair, a written chain of reasoning followed by a classification label on each axis. We require the model to reason about each axis independently (C, then D, then R) before committing to a label. Asking the model to think through its reasoning before producing an answer generally improves output quality on any task (a well-documented property of language models) and in our case also reduces cross-axis contamination, analogous to asking a

survey rater to evaluate one dimension at a time rather than forming an overall impression.

The use of multiple models deserves explanation. A single language model, asked the same question twice, may give different answers; even at temperature zero (the setting that produces the model's single most-likely response), models from different providers may disagree, reflecting genuine ambiguity at classification boundaries, analogous to inter-rater disagreement in human surveys. We tested our pipeline at both server default temperatures and temperature zero and found statistically indistinguishable classification distributions, confirming that the roughly 10% of classifications that vary across runs reflect genuine ambiguity at level boundaries (two reasonable people could disagree about whether a task is C1 or C2) rather than random noise in the measurement instrument.

We use models from three different providers (Anthropic, OpenAI, Google) to guard against systematic biases that may be specific to a single provider's approach. The three providers differ in model architecture, training methods, and (to some degree) training corpora. While all frontier models train on substantial internet text (creating some shared knowledge base, just as all human raters share common cultural and educational background), the architectural and methodological differences provide partial robustness against provider-specific biases. We validate against held-out ground truth (Section 4) rather than relying on inter-model agreement alone.

Classification proceeds in two rounds. In the initial round, each model independently classifies every task. Where any axis shows non-unanimous agreement (any label where the three models do not all agree), the task enters a consensus round, analogous to a Delphi method.

In the consensus round, each model receives, for each disputed DWA-task pair, the category and reasoning provided by *all three* models from the initial round, but is not told which model produced which response (although there is evidence that models can recognize each other's styles). Since each API call provides a fresh context window, the model does not know which answer it previously assigned, a property not possible with human raters, who cannot forget their prior judgment. The reasoning-sharing step allows models to consider perspectives they may have missed ("this task involves variable anatomy, which makes the physical manipulation unpredictable") without anchoring on a specific answer.

After the consensus round, we adopt the majority label (2-of-3 or 3-of-3). This process produces a final unanimous or majority consensus label on all three axes for 98.5% of tasks (23,484 of 23,852) in our production run. At the axis level, 98.0% of initial disagreements are resolved.

We also elicited a confidence level from each model alongside its classification, but did not find it useful. Confidence calibration is provider-specific: one provider's models report near-universal high confidence regardless of task difficulty; another splits roughly evenly, reflecting training procedures rather than classification certainty. We do not use these scores in our analysis. In brief, the classification pipeline shares the core logic of structured content analysis: a codebook, trained raters, independent classification, inter-rater reliability measurement, and reconciliation of disagreements. It differs in three important ways.

First, in scale: we classify 23,850 DWA-task pairs across 923 occupations for under $100 per complete three-model mid-tier run (approximately $200 with flagship models), orders of magnitude

cheaper than equivalent human annotation, which is likely why Eloundou et al. (2023) used human raters only for DWA-level classification, not for the full task-occupation context.

Second, in reproducibility: the exact prompts and model versions are version-controlled and can be re-run by any researcher with API access, producing a replication package that may be helpful for other researchers. (Random seed parameters are not consistently supported across providers at non-zero temperatures, so exact token-level reproducibility is not guaranteed, though classification distributions remain stable.)

Third, the raters have broad but shallow knowledge: language models have read more about every occupation than any individual human rater, but lack the tacit knowledge that comes from actually performing the work. We address this limitation through occupation-enriched prompts (Section 3.2) and external validation (Section 4).

3.1 DWA Decomposition: Solving the Compound-Task Problem

O*NET task descriptions frequently bundle cognitively and physically distinct sub-activities into a single statement. Consider a chiropractor's diagnostic task: "Diagnose health problems by reviewing patients' health and medical histories, questioning, observing, and examining patients and interpreting x-rays" (O*NET Task 6915). The O*NET Tasks-to-DWAs crosswalk decomposes this into three distinct activities with different automation profiles:

Table 5. DWA Decomposition Example: Chiropractor Diagnostic Task

Detailed Work Activity	C	D	R
Gather medical information from patient histories	C2 (contextual judgment)	D1 (sensing)	R1 (social norm)
Examine patients to assess general physical condition	C3 (expert synthesis)	D3 (variable manipulation)	R3 (statutory)
Diagnose medical conditions	C3 (expert synthesis)	D3 (variable manipulation)	R3 (statutory)

Classifying compound tasks as a single unit forces an artificial average and produces unstable ratings; models must guess how to weight the cognitive vs. physical components, and different models guess differently. We resolve this by decomposing each O*NET task into its constituent Detailed Work Activities (DWAs) using the O*NET Tasks-to-DWAs crosswalk. Each DWA-task pair becomes an independent classification unit. This yields 23,850 DWA-task rows from 18,796 original tasks across 923 occupations, a roughly 1.27x expansion that preserves the heterogeneity within compound tasks that drives economic impact estimation.

Eloundou et al. (2023) noted that DWA descriptions are often ambiguous without knowledge of the occupation in which they occur: the DWA "Execute sales or other financial transactions" means physical cash handling for a Gambling Cage Worker but email order confirmations for an Online Merchant (Eloundou et al. 2023, Table 1). Their human raters applied the exposure rubric to

individual DWAs, not task/occupation pairs, without occupation context. They then assigned most tasks based on the DWA, a limitation they explicitly acknowledged. GPT-4 rated task/occupation pairs, receiving only the occupation name and the task description, without DWA breakdown.

Our pipeline addresses this directly: each occupation's tasks are classified within the occupation's full context, including the O*NET narrative description, working conditions, abilities, and licensing status (Section 3.2). This provides more context for the model to understand the task and label accordingly.

Our per-occupation classification with full O*NET profiles represents a further step: the classifier has richer occupational context than either Eloundou's human raters or their GPT-4 baseline. Context windows have expanded from approximately 2,000 words to 1 million words (a factor of roughly 400), making richer per-occupation profiles both feasible and valuable.

This decomposition is possible because O*NET 30.2 provides complete DWA coverage: every task in the database maps to at least one DWA. Earlier versions had gaps in the crosswalk. Eloundou et al. (2023), using O*NET 27.2, noted that "some tasks lack any associated DWAs." O*NET builds this crosswalk incrementally, adding DWA linkages for approximately 170 occupations per quarterly release as each occupation cycles through resurvey. By version 30.2, the crosswalk is complete, eliminating a potential source of selection bias that affected earlier task-level analyses.

Box 1: The Experience Gap: When Regulation Pushes Toward AI

O*NET task descriptions routinely bundle two distinct cognitive inputs: technical knowledge (what to do) and experiential pattern-matching (recognizing which situation you are in). "Diagnose and repair heating systems" treats the diagnostic component, which draws on years of accumulated case exposure, as inseparable from the procedural repair. DWA decomposition (Section 3.1) separates them, revealing that the experiential component has a fundamentally different automation profile than either the technical or physical components.

The reason is sample size. A human physician acquires knowledge through years of formal training and accumulated experience, yet may see a few thousand patients over an entire career; an experienced HVAC technician may encounter a few hundred distinct failure modes across a few thousand service calls; a veteran corporate executive may participate in a few hundred significant strategic decisions across several companies. These are the sample sizes on which human expertise is built, and they are, by the standards of statistical inference, small. AI systems train on corpora that are larger by orders of magnitude. A diagnostic AI can be trained on the case histories of millions of patients; a maintenance AI can ingest the service records of every unit a manufacturer has ever sold; a strategic AI can be trained on the documented outcomes of every major corporate decision in the public record. The experiential pattern-matching that constitutes much of what we call "professional judgment" is inference from accumulated cases, and AI has access to a case base that no individual human career can match.

The experience-unbundling pattern is general. Wherever professional value derives substantially from accumulated case exposure (diagnostic medicine, legal strategy, financial risk assessment, engineering failure analysis), AI's sample-size advantage grows with each year of digitized records. The CDR framework makes this visible through DWA decomposition (separating the experiential component from the technical and physical) and the R-axis (whose sign depends on whether regulation protects the practitioner or the client).

The RCT evidence is consistent with this framing. Brynjolfsson et al. (2023) find that novice customer support agents with AI assistance approach the performance of experienced agents; the AI supplies the case-history pattern-matching that experience would otherwise provide. Noy & Zhang (2023) find that lower-ability writers benefit most from AI assistance; the quality floor rises toward the existing ceiling. In both cases, the mechanism is the same: AI substitutes for accumulated experience, compressing the skill distribution from below.

This has a direct implication for the R-axis. Regulatory pressure on AI adoption is commonly assumed to be uniformly restrictive, with guilds and licensure regimes slowing deployment. But the direction of regulatory pressure depends on *whose* interests the regulation serves.

In medicine, the AMA has fought scope-of-practice expansion for decades; regulation is structured around protecting the licensed practitioner's role as gatekeeper. AI enters medicine *through* the physician, not around her, because statutory authority (R3) forces it through the licensed channel; here, R-axis friction decelerates adoption.

Can regulatory pressure ever run in the opposite direction, not restraining AI adoption but *requiring* it? Corporate governance suggests yes. Officers and directors owe fiduciary duties to shareholders; the duty of care requires informed, diligent decision-making.

A board relying on a CEO whose strategic judgment draws on experience at two or three companies, when AI systems trained on the full corpus of public corporate outcomes are available, faces an increasingly awkward question: is the refusal to use AI *itself* a breach of fiduciary duty?

As the demonstrated capability gap between AI-augmented and unaugmented executive judgment widens, the regulatory pressure may shift from restraining AI adoption to requiring it. The R-axis value for executive decision-making tasks could prove to be negative: institutional pressure *toward* AI, not against it. The formal modeling of when and how regulatory pressure reverses sign is developed in a companion paper on adoption dynamics.

3.2 Occupation-Enriched Classification

Each occupation's tasks are classified in a single API call, with the prompt structure separating stable from variable content. The system prompt (~5,200 words, identical across all occupations) contains CDR axis definitions with boundary tests, the 50% time-reduction framing, ground-truth calibration examples spanning the full axis space, disambiguation rules, and response format specification; this prompt is cached by all three providers, reducing per-call cost.

The data prompt (which varies per occupation, typically 200–500 words) contains the occupation's O*NET narrative description, working conditions (indoor/outdoor, physical demands, hazard exposure), cognitive and physical abilities with importance scores, licensing status from Carollo's (2025) occupational licensure dataset, apprenticeship status from RAPIDS, and the task list formatted as DWA-task pairs.

This design grounds the LLM in the specific occupational context. The same task description, "demonstrate techniques to students," means physical clay work for a pottery instructor (C1/D3) and whiteboard problem-solving for a math teacher (C1/D0). The occupation profile disambiguates without requiring per-task prompt engineering.

3.3 Multi-Model Consensus

Each occupation batch is independently classified by three language models. For the production run, we used mid-tier models: Claude Sonnet 4.6 (Anthropic), GPT-5-mini (OpenAI), and Gemini 3 Flash (Google); results were validated against flagship models (Claude Opus 4.6, GPT-5.2, and Gemini 3 Pro) on a 420-task calibration set (Section 3.5). Models receive identical prompts and produce per-axis reasoning followed by classification labels.

Consensus is determined by majority agreement (2-of-3 or 3-of-3). Where any axis shows disagreement between models (not just full three-way splits), the task enters a reconciliation round. In this round, each model receives the other two models' chain-of-thought reasoning (but not their labels), then re-classifies. This COT-sharing protocol resolves the large majority of disputes: the production run achieved a 97.2% resolution rate after reconciliation.

1.5% of tasks (368 of 23,852) did not reach consensus on all three axes after both rounds. For these, we adopted the label from the model with highest overall agreement on that axis. Of these unresolved disputes, 80% involve the R-axis, where genuine uncertainty about the augmentation boundary (Section 2.3) produces persistent three-way splits that reasoning-sharing cannot resolve.

3.4 Per-Axis Reasoning Format

After extensive experimentation comparing merged (all-axes-at-once) and per-axis (one reasoning block per axis) formats, we adopted per-axis reasoning for production classification. The model produces separate chains of thought for C, D, and R before committing to each label. This format reduces cross-axis contamination (the tendency for high R, indicating a licensed profession, to inflate C and D ratings) and produces more stable classifications across runs.

A particular concern in this regard is prestige conflation. Language models tend to associate high-prestige occupations with high difficulty across all dimensions, a halo effect where the social status of the occupation inflates the perceived difficulty of the underlying tasks. A surgeon dictating operative notes is performing C0 work (transcription from memory); a lawyer filing a motion for extension of time is performing C1 work (filling in a standard form); a professor entering grades into a learning management system is performing C0 work (data entry). The production prompts

include explicit warnings against this pattern on each axis, with worked examples demonstrating that the task, not the occupation, determines the classification.

3.5 Pipeline Stability

End-to-end stability was measured via three fully independent pipeline runs (independent initial round classifications followed by independent consensus rounds) on a 420-task calibration set spanning 12 occupations selected to cover the full CDR space. Pairwise agreement before consensus averaged 75.0% overall: C-axis 74.4%, D-axis 79.5%, R-axis 71.1%. Pairwise agreement after consensus averaged 90.8% overall: C-axis 85.6%, D-axis 88.6%, R-axis 98.3%. The consensus round thus improves agreement by approximately 16 percentage points overall, with the largest gain on R (+27pp), consistent with R being the axis where reasoning-sharing most frequently surfaces constraints one model missed.

The dominant instability source is the C1↔C2 boundary, accounting for 47% of C-axis disputes (4,277 of 9,064 disputed tasks) and 18% of all tasks. Notice that both levels represent tasks where AI can provide substantial assistance; the distinction between procedural and contextual is real but does not change the automation story. The economically significant C2↔C3 boundary (which determines whether a task is within the current AI cognitive frontier) accounts for 24% of C-axis disputes (2,201 tasks), or 9% of all tasks. The R-axis is essentially production-ready at 98.3% post-consensus stability.

Temperature experiments (varying the degree of randomness in each model's responses, from deterministic to stochastic) confirmed that the ~10% instability reflects genuine classification ambiguity at level boundaries rather than sampling noise (see Section 3.0). Cross-tier validation (mid-tier vs. flagship models on 420 tasks) showed 83.5% agreement, comparable to within-tier pairwise agreement, confirming that model tier (each AI provider offers a smaller, cheaper "mid-tier" model and a larger, more capable "flagship" model; see Section 3.3) is not a meaningful source of classification variance. Full cross-tier details are reported in Appendix B.

4 Validation and Comparison with Prior Exposure Estimates

This section validates the CDR classifications against four independent ground-truth datasets. Each test asks whether the framework's ordinal ratings, cognitive complexity (C0–C4), deployment difficulty (D0–D4), and regulatory restrictions (R0–R4), align with external measures that were not used in the classification process. Full definitions and boundary tests appear in Section 2.

4.1 Eloundou Comparison: Physical Requirements Predict Exposure

We compare CDR (Cognitive complexity, Deployment difficulty, Regulatory restrictions) consensus classifications against Eloundou et al.'s (2023) exposure ratings, which classify tasks as E0 (no exposure) through E2 (exposed with image + retrieval capabilities). The optimal prediction rule uses a single CDR axis: D $\geq$ 1 (any task requiring physical-world interaction) predicts Eloundou E0

at 81.1% accuracy (precision 0.733, recall 0.904, F1 = 0.809 vs. GPT-4 task-level raters). A sweep of all 64 possible C × D × R binary cutpoints confirms that no other axis combination improves on D ≥ 1 alone; adding R contributes zero additional predictive power, and adding C is redundant once D is controlled for.

Their annotation had two tracks: human raters applied the exposure rubric primarily to bare DWAs (Detailed Work Activities), without knowledge of which occupation the activity belonged to, and a subset of tasks; GPT-4 rated all task/occupation pairs, receiving the occupation title alongside the task description but no richer context. Both tracks evaluated against a capability envelope described as "the most powerful OpenAI large language model… where the context for the input can be captured in 2000 words," approximately 2,500 tokens of text-only input, with no multimodal perception, code execution, or tool use.

Under these constraints, any task requiring physical-world interaction was trivially E0 (no exposure): a text-only model cannot perceive or manipulate the physical environment. Our D ≥ 1 predictor formalizes exactly this implicit criterion (physical deployment requirements) and achieves 81.1% agreement with GPT-4's task-level ratings (F1 = 0.809).

Our work builds directly on Eloundou et al.'s foundation by addressing the limitations they identified. Their human raters, evaluating bare DWAs, could not disambiguate activities like "Execute sales or other financial transactions" across occupations, a limitation they explicitly acknowledged and recommended for future work. Their GPT-4 ratings had occupation titles but no working conditions, physical demands, abilities, or licensing data.

We followed their approach, providing the full O*NET profile (narrative description, physical and cognitive ability scores, hazard exposure, licensing status), resolving the context gap they identified. We also apply the same 50% time-reduction question but against the current capability frontier: multimodal perception, agentic tool orchestration, 1 million token context windows. Some discrepancy along the C-axis is expected given model advances since their study; our estimates of the AI-accessible task space are substantially larger in part because the models being evaluated against are substantially more capable.

Eloundou et al. construct their primary exposure measure as the sum of full exposure plus half of tool-assisted exposure (β = E1 + 0.5×E2), where the 0.5 weight on E2 "is intended to account for exposure when deploying the technology via complementary tools and applications necessitates additional investment."

For some E2 tasks, which require only software development on existing digital infrastructure (D0 in our framework), the 0.5 discount substantially overstates the deployment barrier. For others, which require physical deployment or regulatory navigation, the discount may understate the true barrier. The CDR decomposition replaces this fixed scalar with task-specific measurements of deployment difficulty (D) and regulatory friction (R), each of which advances at its own rate.

The scale of the capability shift is worth stating explicitly. Eloundou's capability envelope was a text-only model with a 2,000-word context window, no image processing, no code execution, and no tool use. Current frontier models process text, images, and audio with context windows of

approximately 800,000 words (400× larger), execute code, orchestrate multi-step tool chains, and browse the web — a qualitatively different capability surface that renders many of Eloundou's E0 classifications obsolete.

The broader exposure CDR measures relative to Eloundou reflects four distinct factors, each operating independently. First, *capability expansion*: tasks like "review photographs of damaged property" were trivially E0 for a text-only model but are C1/D0 with multimodal AI that can process images directly. Of the 6,769 tasks Eloundou's GPT-4 labeled E0 (no exposure), 26.5% are digitally achievable (C≤2, D≤2) under CDR's current-capability evaluation. Second, *per-occupation context*: Eloundou's human raters saw DWAs in isolation, without knowledge of the occupation. With occupation context, "operate equipment" for a Kindergarten Teacher resolves to classroom technology (C0/D0) rather than defaulting to heavy machinery (E0). Third, *DWA decomposition*: compound tasks like "diagnose and repair heating systems" receive a single E0 rating if any component requires physical interaction. CDR's decomposition rates the cognitive sub-task (C2/D0) separately from the physical sub-task (C1/D3), capturing the automatable portion that a single rating obscures. Of the 6,863 tasks Eloundou labeled E2 (exposed with tools), 53.1% are C0/D0 in CDR, meaning they require only text interaction, not the software infrastructure E2 implies. Fourth, *model generation effects*: the evaluating LLMs themselves are more capable than GPT-3.5, potentially recognizing task accessibility that earlier models could not assess. Factors (1) and (4) are entangled in the current data — separating capability expansion from evaluator improvement would require re-running classification with historical model capabilities, an avenue for future work (see Section 3.3).

4.2 Carollo Occupational Licensure: The Thin Chokepoint Thesis

The R-axis captures regulatory restrictions on AI-assisted task performance. We classify over five categories: R0 (no barrier), R1 (client or consumer norms), R2 (professional standards or liability), R3 (statutory regulation), and R4 (moral agency required). Full definitions and boundary tests appear in Section 2. We validate the R-axis against Carollo's (2025) occupational licensure dataset, which records the presence of state-issued licensing requirements for occupations across all 51 U.S. jurisdictions (50 states plus D.C.). The dataset identifies *which occupations* are statutorily licensed but does not distinguish *which specific tasks within those occupations* actually require the license to perform.

A nurse requires a license to administer controlled substances but not to monitor vital signs or document patient information; the occupation is licensed, but most tasks within it are not license-restricted. This occupation-vs-task distinction is precisely what the R-axis, under augmentation framing, is designed to capture.

For readers more familiar with economics than machine learning: precision measures how often a prediction is correct when we label a task R3 (statutory regulation; the fraction of R3 predictions that correspond to actually licensed occupations), while recall measures how many of the truly licensed tasks we identify (the fraction of tasks in licensed occupations that we label R3). A framework can have high precision but low recall if it is conservative, correctly identifying a small subset of

regulated tasks while missing others.

Under the augmentation framing, R-axis recall against Carollo's statutory licensure indicator is approximately 50%. This is low in absolute terms but consistent with the framework's design logic. Carollo marks an occupation as licensed, but most tasks within a licensed occupation do not require the license to perform.

A nurse's tasks include monitoring vital signs (R0), documenting patient information (R0), checking drug interactions (R0), and administering controlled substances (R3); while the occupation requires a license to practice, regulatory restrictions apply to specific tasks within that occupation, typically those involving statutory authority, controlled substances, or procedures with significant liability exposure. Under augmentation framing, the licensed professional can use AI tools for the R0 tasks, and those tasks constitute the majority.

We call this the **thin chokepoint thesis**: in licensed occupations, the tasks where regulation actually restricts AI assistance are few, typically the final sign-off, the physical procedure, or the exercise of statutory authority. The R-axis identifies these chokepoints while classifying the surrounding tasks as low-friction, which is what the augmentation framing requires.

The thin chokepoint thesis has a direct employment implication. The profession does not disappear (the regulatory barrier ensures licensed professionals remain necessary), but the number of professionals required may decline sharply. If 80% of a nurse's tasks are R0 (and thus AI-accessible) and only 20% hit the R3 chokepoint (administering controlled substances, invasive procedures), AI assistance allows each nurse to handle a larger patient load. The chokepoint persists, but fewer workers are needed to clear it. In practical terms, the regulatory barrier prevents full automation but does not prevent labor-saving productivity gains.

Cross-model variation in the chokepoint percentage (11–44% across models within the same 18 occupations) reflects genuine uncertainty about where the augmentation boundary lies, an active area of both legal and practical evolution.

We attempted to extend this validation to R2 (professional standards) using Career OneStop data from the U.S. Department of Labor, which provides both statutory licensing and professional certification information for 199 occupations. Cross-referencing with Carollo's licensing data identifies 10 occupations that are certified but not statutorily licensed, a natural R2 ground truth set including cartographers, environmental scientists, hydrologists, urban planners, interior designers, and locomotive engineers. Of these, our R-axis classified 4 of 10 (40%) with any task at $R \geq 2$, substantially weaker than R3 detection (precision 0.88, recall 0.70). However, the ground truth set is small ($N = 10$), and OneStop has significant coverage gaps: it classifies dentists, physicians, and EMTs as having neither license nor certification. The low R2 recall likely reflects a real limitation (certification is a softer signal than statutory licensing, and harder to infer from task descriptions alone), but the specific numbers carry wide confidence intervals.

Among Carollo-unlicensed occupations where OneStop also reports no certifications, several nonetheless received $R \geq 1$ classifications. Inspection of these cases suggests the models are largely correct: they detect restrictions arising from testimony requirements, union collective bargaining

agreements, and occupational safety regulations that neither Carollo nor OneStop capture. The R-axis appears to be picking up genuine regulatory friction beyond what our ground truth datasets measure, a pattern consistent with the axis measuring a broader construct than statutory licensing alone.

5 Employment-Weighted Results

This section answers a single question: where in the U.S. economy (measured by labor time, not worker count) are tasks located in the three-dimensional CDR (Cognitive complexity, Deployment difficulty, Regulatory restrictions) space? We measure in labor time rather than occupations because the augmentation framing (Section 2.3) implies that AI affects tasks within occupations, not occupations as wholes, and the time a worker spends on each task shapes the potential productivity impact. The answer determines how much of the economy is accessible to AI at current capability levels, and how shifts in the three frontiers translate into aggregate productivity impacts. These results take the current structure of the economy as given; they do not account for the reorganization of occupations, firms, or task bundles that may follow as AI capabilities expand (see Section 2.5).

5.1 The C × D Landscape: Cognitive Complexity and Deployment Difficulty

The C × D cross-tabulation is the CDR framework's analog of Autor's (2003) 2×2 routine/non-routine × cognitive/manual matrix, extended from a binary classification to a 5×5 plane. The C-axis captures cognitive complexity in five levels: C0 (lookup/copy), C1 (procedural), C2 (contextual judgment), C3 (expert synthesis), and C4 (discovery/creation). The D-axis captures deployment difficulty in five levels: D0 (purely digital), D1 (sensing/locomotion), D2 (structured manipulation), D3 (unstructured manipulation), and D4 (dynamic multi-modal). The Autor quadrants map onto rectangular regions of the CDR grid: "routine cognitive" corresponds to C0–C1 × D0–D1 (procedural tasks requiring no physical manipulation); "non-routine cognitive" to C2+ × D0–D1; "routine manual" to C0–C1 × D2+; and "non-routine manual" to C2+ × D2+. The CDR improvement is that each quadrant becomes a gradient rather than a homogeneous category, revealing structure within Autor's cells, in particular the dominant C2/D0 concentration, that the 2×2 collapses.

To weight task-level classifications by economic significance, we combine BLS Occupational Employment and Wage Statistics (employment counts by occupation) with O*NET task ratings (frequency, importance, and relevance scores that, following the ONS (2022) calibration methodology, convert to time-allocation weights). The product (employment count x time fraction) gives economy-weighted labor time per task, capturing where workers spend their time rather than merely how many workers there are.

Applying this to the full O*NET task universe (23,850 DWA-task rows, 923 occupations) produces the following employment-weighted C × D cross-tabulation, shown without regard to R-axis (i.e., all tasks regardless of regulatory restrictions).[4] Each cell shows the percentage of total U.S. economy-

[4]Percentages represent the share of total U.S. economy-weighted labor time; that is, the fraction of all hours

weighted labor time accounted for by tasks at that combination of cognitive complexity and physical deployment difficulty. Rows sum across deployment levels; columns sum across cognitive levels; the grand total is 100%. Reading across a row shows how a given cognitive level is distributed across physical requirements; reading down a column shows the cognitive profile of tasks at a given deployment level.

Table 6. Economy-Weighted Labor Time by Cognitive Complexity and Deployment Difficulty (All Tasks, Regardless of Regulatory Restrictions)

	D0	D1	D2	D3	D4	Total
C0	1.5%	0.7%	2.1%	2.2%	0.0%	6.4%
C1	15.3%	4.2%	12.6%	12.3%	0.1%	44.6%
C2	27.9%	8.7%	1.3%	4.5%	0.3%	42.7%
C3	4.0%	1.2%	0.2%	0.5%	0.3%	6.3%
C4	0.1%	0.0%	0.0%	0.0%	0.0%	0.1%
Total	48.9%	14.8%	16.2%	19.5%	0.6%	100%

Notes: Each cell shows the percentage of total U.S. economy-weighted labor time (hours worked across all occupations, weighted by employment counts and within-occupation time allocation) accounted for by tasks at that combination of cognitive complexity and deployment difficulty, regardless of regulatory restrictions. Rows are the C-axis (Cognitive complexity): C0 = lookup/copy, C1 = procedural, C2 = contextual judgment, C3 = expert synthesis, C4 = discovery/creation. Columns are the D-axis (Deployment difficulty): D0 = purely digital, D1 = sensing/locomotion, D2 = structured manipulation, D3 = unstructured manipulation, D4 = dynamic multi-modal. All cells sum to 100%.

The headline finding is that 44.7% of economy-weighted labor time is C $\leq$ 2 and D0, cognitively within reach of current AI systems and requiring zero physical infrastructure. One could debate whether C3 is also within reach, but C2 is already past. These are tasks where a text-based AI interface can plausibly save a worker 50% of her time. The C2/D0 cell alone, tasks requiring contextual judgment but no physical interaction, accounts for 27.9% of economy-weighted labor time and represents the largest single cell in the matrix.

The two-dimensional structure of the CDR framework distinguishes between workers whose AI-exposed tasks are D0 (accessible via software interface) and those whose exposed tasks are at higher D levels (requiring physical presence or manipulation). Eloundou et al.'s framework identifies 80% of workers as having some task exposure, but aggregates across D levels. The CDR cross-tabulation makes the D-axis distinction explicit and quantifiable.

worked across all occupations, weighted by employment counts and within-occupation time allocation. Each cell shows what percentage of total labor time falls in that C × D combination; all cells sum to 100%. C-axis (Cognitive complexity): C0 = lookup/copy (retrieving existing information); C1 = procedural (following established rules); C2 = contextual judgment (weighing tradeoffs, adapting to context); C3 = expert synthesis (integrating across domains under uncertainty); C4 = discovery/creation (generating fundamentally new knowledge). D-axis (Deployment difficulty): D0 = purely digital (no physical-world interaction); D1 = sensing/locomotion (perceiving or moving through the physical environment, no manipulation); D2 = structured manipulation (contact with objects in engineered environments); D3 = unstructured manipulation (contact with objects in variable, unpredictable environments); D4 = dynamic multi-modal (simultaneous real-time coordination of perception, locomotion, and manipulation under time pressure).

5.2 The R-Axis Overlay: Regulatory Restrictions Are Thin

Overlaying the R-axis on the C × D landscape reveals that regulatory restrictions remove a small fraction of tasks from the automatable pool. Each row below shows what percentage of total U.S. labor time (weighted by employment and time-on-task, as in the C × D table) falls at that regulatory restriction level:

Table 7. Economy-Weighted Labor Time by Regulatory Restriction Level

R level	Economy share	Interpretation
R0–R1	80.4%	No regulatory barrier, or only consumer preference norms that erode under price pressure
R2	11.5%	Professional standards or significant liability exposure constrain AI use; surmountable on a 3–5 year institutional cycle
R3	8.0%	Statutory regulation restricts AI assistance itself; requires legislative action
R4	0.1%	Moral agency required; definitional, not contingent (see Section 2.3)

Notes: Each row shows the percentage of total U.S. economy-weighted labor time (hours worked across all occupations, weighted by employment counts and within-occupation time allocation) accounted for by tasks at that level of regulatory restriction, regardless of cognitive complexity or deployment difficulty. R-axis (Regulatory restrictions): R0 = no barrier, R1 = client/consumer norms, R2 = professional standards/liability, R3 = statutory regulation, R4 = moral agency required. R0 and R1 are grouped because both represent barriers that erode under market pressure. All rows sum to 100%.

The R2 category (professional standards and liability exposure) is substantively distinct from R0–R1 even though both are surmountable. R2 tasks include activities where malpractice liability, professional certification requirements, or industry-body standards constrain how AI tools can be deployed; these barriers erode on a 3–5 year institutional cycle as professional bodies update credentialing frameworks, not on the months-scale of R1 consumer preference erosion. Grouping R0–R2 together obscures this economically meaningful difference in erosion timeline.

Under augmentation framing, R3 tasks concentrate in two clusters: licensed physical procedures (nursing tasks, cosmetology, CDL-required driving) at D3, and licensed cognitive sign-offs (accounting attestation, legal determinations) at D0. Together they represent only 8.1% of economy-weighted labor time.

This finding has direct implications for policy analysis. Acemoglu, Autor & Johnson (2026) propose reducing occupational licensure barriers as a policy lever for enabling AI-augmented workers to enter traditionally protected markets, arguing that licensure requirements "stifle competition, raise prices, and **have** at best, mixed effects on service quality." The CDR data indicates that such licensure reform (an R-axis intervention) would unlock only a small marginal increment of task automation. The binding constraints for the vast majority of the economy are cognitive complexity (C) and physical deployment difficulty (D), not regulation.

5.3 The Wavefront

The CDR framework enables a wavefront analysis: given projections for when each axis threshold is crossed, what fraction of the economy, measured as economy-weighted labor time (hours worked across all occupations, weighted by employment counts and within-occupation time allocation), becomes accessible to AI-assisted productivity gains? We explicitly hold $R < 2$ (tasks facing no professional standards, statutory, or moral agency barriers) for the wavefront, since we are interested in which tasks are currently or imminently accessible without professional or statutory barriers.[5]

Table 8. CDR Wavefront: Cumulative Economy Share by Capability Threshold

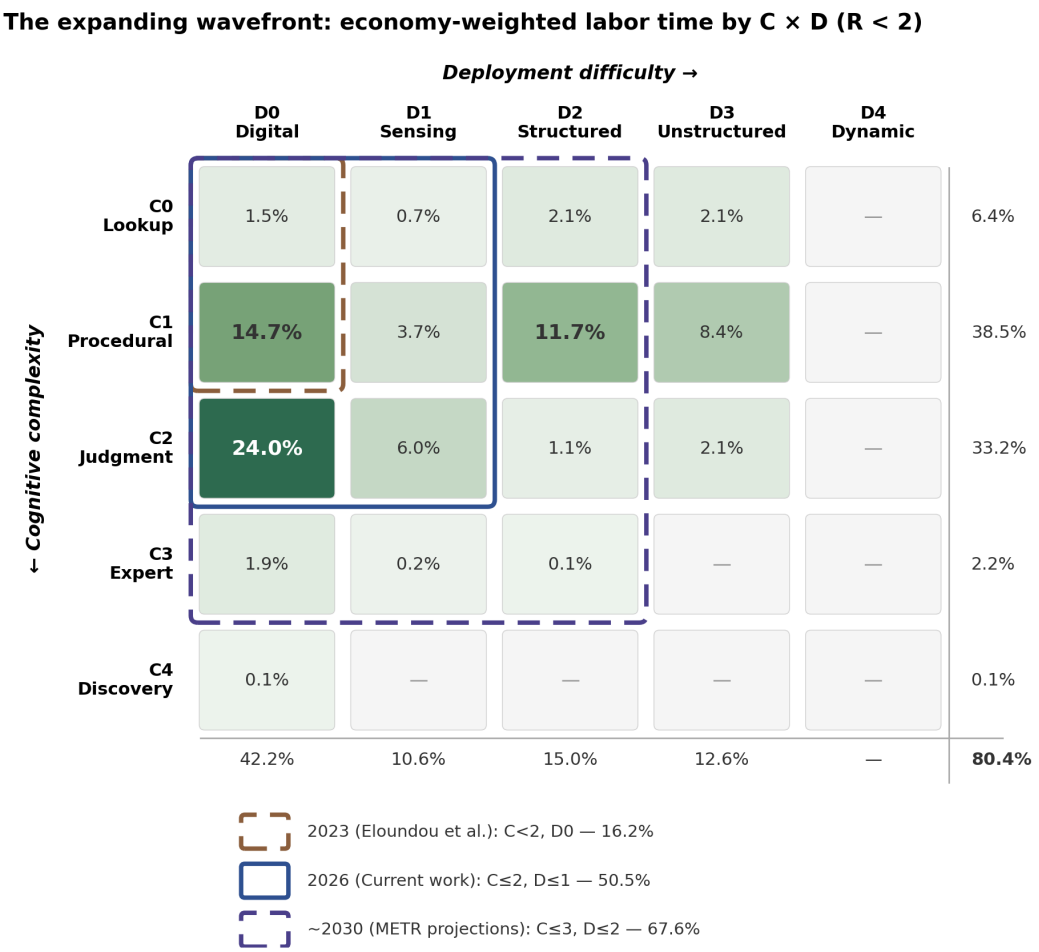
Wavefront threshold	Economy share ($R < 2$)	Approximate timeline
$C \leq 1$, $D = 0$	16.2%	Now
$C \leq 2$, $D = 0$	40.2%	~1 year
$C \leq 2$, $D \leq 1$	50.5%	~2 years
$C \leq 2$, $D \leq 2$	65.3%	~3–5 years

Notes: Each row shows the cumulative percentage of U.S. economy-weighted labor time (hours worked across all occupations, weighted by employment counts and within-occupation time allocation) accounted for by tasks at or below the indicated cognitive complexity (C) and deployment difficulty (D) thresholds, filtered to tasks with no professional standards, statutory, or moral agency barriers ($R < 2$). C levels: C0 = lookup/copy, C1 = procedural, C2 = contextual judgment. D levels: D0 = purely digital, D1 = sensing/locomotion, D2 = structured manipulation. Approximate timeline reflects projected time until AI capabilities and deployment infrastructure reach the indicated threshold, based on observed rates of advance (see Section 6). The remaining 19.6% of economy-weighted labor time is excluded by $R \geq 2$ barriers: professional standards/liability (R2: 11.5%), statutory regulation (R3: 8.0%), and moral agency requirements (R4: 0.1%).

The step from 40.2% to 50.5% (D0 → D1) requires only sensing and locomotion capabilities (smartphones and wearables, not industrial robotics). The step to D2 and beyond requires physical manipulation capabilities that are advancing faster than commonly assumed: Agility Robotics reports 98% task success in warehouse settings, Harvest CROO announced commercial-parity strawberry harvesting in 2025, and humanoid robot deployment grew to approximately 16,000 units globally in 2025 with 40% year-over-year cost declines. Our robotics capability timeline (Appendix C) places specific high-value D3 tasks at commercial scale by 2029–2033; "long-run" may be as short as five

[5]Wavefront threshold: the maximum C and D levels included (e.g., $C \leq 2$, $D = 0$ means all tasks classified as cognitively within contextual judgment or below, C0–C2, that require no physical-world interaction, D0 only). Economy share: the cumulative percentage of U.S. economy-weighted labor time (hours worked × employment) that falls at or below those thresholds, filtered to $R < 2$ (tasks facing no professional standards, statutory, or moral agency barriers). The remaining 19.6% of economy-weighted labor time is excluded by $R \geq 2$ barriers: professional standards/liability (R2: 11.5%), statutory regulation (R3: 8.0%), and moral agency requirements (R4: 0.1%). Approximate timeline: projected time until AI capabilities and deployment infrastructure reach the indicated threshold, based on observed rates of advance (see Section 6). C levels: C0 = lookup/copy, C1 = procedural, C2 = contextual judgment. D levels: D0 = purely digital, D1 = sensing/locomotion, D2 = structured manipulation, D3 = unstructured manipulation.

Figure 1. The Expanding Wavefront: Economy-Weighted Labor Time by C × D (R < 2)



Note: Each cell shows the percentage of U.S. economy-weighted labor time (hours worked across all occupations, weighted by employment counts and within-occupation time allocation) accounted for by tasks at that combination of cognitive complexity (rows) and deployment difficulty (columns), filtered to tasks facing no professional standards, statutory, or moral agency barriers (R < 2). The three dashed rectangles represent expanding wavefronts: the 2023 boundary (Eloundou et al., C < 2, D0: 16.2%), the 2026 boundary (current work, C ≤ 2, D ≤ 1: 50.5%), and the projected ~2030 boundary (METR projections, C ≤ 3, D ≤ 2: 67.6%). Darker shading indicates larger shares of labor time.

years for targeted applications.

These are not productivity estimates: translating wavefront coverage into TFP requires a production function model with appropriate elasticities of substitution, task complementarities, and general-equilibrium effects, which is the subject of a planned companion paper. The wavefront numbers here are purely descriptive: they show where in the C $\times$ D plane the economy's labor time is concentrated, and thus which capability thresholds unlock the largest increments of potential automation.

5.4 Implications for the Pace of Economic Impact

The CDR (Cognitive complexity, Deployment difficulty, Regulatory restrictions) data bears on a recurring point of disagreement in the literature on AI's macroeconomic implications: whether standard technology diffusion models adequately characterize the deployment timeline for AI-assisted productivity gains. The leading framework for this question is Brynjolfsson, Rock & Syverson's (2021) Productivity J-Curve, which argues that general-purpose technologies suppress measured productivity during an initial phase of complementary intangible investment (reorganizing workflows, retraining workers, building new business processes) before eventually producing an overshoot when those investments mature. The J-curve is the standard optimistic explanation for the persistent disconnect between task-level productivity gains of 15–30% (Coupé & Wu, 2025) and the near-zero aggregate effects documented through early 2026 (Humlum & Vestergaard, 2025; Yotzov et al., 2026). McElheran et al. (2025), working with Census Bureau data on tens of thousands of U.S. manufacturers, provide the first large-scale micro-level confirmation: AI adoption causes average short-run productivity *losses* of 1.3 percentage points, concentrating among older firms, with recovery emerging over a four-year horizon.

The J-curve framework, however, treats the complementary investment trough as undifferentiated, a single adjustment period whose length is calibrated to historical GPT analogs like electrification (David, 1990), where factory reorganization took decades. The CDR data suggests this is too coarse. The *nature* of complementary investment varies systematically with the D-axis. For D0 tasks (purely digital), the required complements are cognitive: workflow redesign, prompt engineering skill acquisition, and organizational learning about when to trust AI output. These adjustments operate on timescales of weeks to months; Dillon et al. (2025) found that knowledge workers captured immediate email time savings within their six-month trial, suggesting the individual-task J-curve for cognitive work is shallow and fast. For D2–D3 tasks, the required complements are physical: industrial robotics, sensor infrastructure, supply chain reconfiguration, and regulatory certification. McElheran et al.'s manufacturing J-curve, with its multi-year trough and recovery conditional on survival, reflects precisely these slower-moving physical adjustments. The D-axis thus provides structural grounds for stratifying the J-curve by deployment physicality, predicting shorter troughs for the 48.9% of labor time at D0 and longer troughs for the 31.6% at D2+. This decomposition also implies that the gap between task-level productivity gains and aggregate economic statistics should resolve *unevenly*, first in knowledge-work-intensive sectors, and only later in sectors requiring physical-world transformation.

Standard diffusion models (the S-curves of Griliches (1957), the adoption lags of David (1990)) treat deployment infrastructure as the rate-limiting step. For electricity, automobiles, and computers, that characterization was accurate: technological capability preceded infrastructure at scale by years or decades.

For AI applied to D0 tasks, the deployment infrastructure is already in place: every knowledge worker with internet access has the hardware and software needed for AI-assisted task performance. The diffusion constraint for D0 tasks is organizational adaptation and learning, not infrastruc-

ture buildout. Early evidence suggests these organizational lags are substantially shorter than infrastructure-driven ones: Dillon et al.'s (2025) knowledge workers captured task-level time savings within weeks of deployment, whereas McElheran et al.'s (2025) manufacturers required four or more years to reach the upswing, a difference consistent with the cognitive-versus-physical complementary investment distinction the D-axis encodes.

Davidson et al. (2026) model the conditions under which automating AI research produces explosive growth, finding that 13% equal automation across sectors is sufficient for hyperbolic growth trajectories. Jones & Tonetti (2026) provide a counterweight: weak links in production chains constrain the growth effect substantially, with their baseline yielding approximately 4% cumulative output gains by 2040.

Jones & Tonetti, however, use a single economy-wide elasticity of substitution (σ). Kording & Marinescu (2025) address this limitation directly, modeling AI's impact with separate elasticities for different task types; their framework splits σ by the degree to which AI can substitute for human labor in each sector. The CDR data supports this approach: the substitutability between AI and human labor for a software engineering task (D0, purely digital) is structurally different from a construction task (D3, unstructured physical manipulation). A single σ compresses this heterogeneity into a parameter that is too high for D3+ tasks and too low for D0 tasks, overstating the deployment barrier where it is near-zero and understating it where it is binding.

The CDR data is consistent with the weak-links mechanism: the D3 bottleneck (19.5% of the economy) represents a genuine constraint on full automation, while also indicating that the D0 portion of the economy (48.9% of labor time) faces no comparable infrastructure barrier. The relative weight of these forces depends on substitutability assumptions and production function structure that the CDR framework alone cannot resolve, but a production function with D-level-specific σ would be more faithful to the task-level heterogeneity the data reveals.

This heterogeneity also bears on cost-effectiveness estimates. Svanberg et al. (2024) found only 23% of AI-exposed tasks economically viable to automate, but their study examined computer vision tasks requiring cameras, custom hardware, and physical deployment — tasks corresponding to D1 or higher in our framework, where deployment cost is substantial relative to displaced labor cost. For D0 tasks, the marginal deployment cost is an API call; the binding constraint is whether the AI's output quality justifies the compute cost, not whether cameras and custom hardware can be amortized. Since D0 holds the bulk of economy-weighted labor time (48.9%), applying a hardware-deployment cost filter uniformly across all exposed tasks is a category error. A more defensible approach applies D-level-specific profitability fractions: Svanberg's 23% for D1+ tasks (where their hardware-deployment cost analysis applies directly), and a substantially higher fraction for D0 tasks, where deployment cost is near zero and the remaining friction comes from tasks where AI output quality is insufficient, integration with existing workflows is non-trivial, or the task volume is too low to justify even minimal setup.

On the gap between technical accessibility and actual use: Handa et al. (2025) find that AI usage is heavily concentrated in software development and writing tasks, with the top 10 task categories

accounting for 24% of all Claude conversations. The Anthropic Economic Index V4 (2026) documents that this concentration is broadening gradually but remains well below the 40.2% C $\leq$ 2/D0/R $<$ 2 share that the CDR taxonomy identifies as accessible without professional or statutory barriers. The distance between technical accessibility and realized adoption, and the rate at which that distance is closing, determines the near-term path of aggregate productivity effects.

Acemoglu (2024) frames the distinction between "easy" and "hard" automation tasks partly in terms of verifiability: whether the output of an AI system can be cheaply checked for correctness. Software development illustrates this dynamic: the profession has mature verification infrastructure in the form of automated test suites ("unit tests") that confirm each modular piece of code behaves correctly, independent of how it was written. This makes AI-generated output cheaply verifiable, a property that dramatically lowers the effective risk of delegation, and which Acemoglu identifies as a key determinant of adoption speed.

The adoption data confirms this. Software development is the sector with the highest AI exposure share and among the shortest observed adoption lags. Across 259 companies and 2.16 million merged pull requests, AI-assisted submissions grew from 14% to 51% of all merges between June 2024 and May 2025, a 260% year-over-year increase (Jellyfish 2025). AI-assisted pull requests completed 13.7 hours faster on average (8.6 hours coding, 5.1 hours review), with no measurable increase in bug rates (consistent 8–9% across adoption levels).

The critical observation is that nothing inherent to software makes it uniquely verifiable; the profession *built* verification infrastructure over decades. Other professions face substantial economic pressure to do the same: deterministic validators that check citations against source URLs and titles, automated compliance checkers for legal filings, structured output validators for financial reports. As these verification tools mature, the adoption dynamics currently observed in software may generalize to other D0 sectors. Whether that trajectory unfolds at similar speed is a question the framework flags as empirically important, but the economic incentive to develop cheap verification is clear wherever AI can provide substantial time savings.

6 Rate of Change: Three Speeds

The CDR framework makes visible the differential rate of advance across dimensions. We present evidence on these rates and discuss their implications for forecasting.

6.1 The Cognitive Frontier

The economics literature has largely treated AI capability advance as exogenous and gradual. Acemoglu (2024) calibrates to a 10-year horizon with static capability parameters. This assumption can now be tested against measurement data. METR (Model Evaluation and Threat Research) maintains a benchmark suite of real-world tasks, primarily software engineering, machine learning, and cybersecurity, calibrated by the time a skilled human professional would need to complete them. The key metric is the *autonomous task time horizon*: the longest task duration at which an AI

system achieves a given success rate when working entirely without human assistance. This provides a concrete, time-denominated measure of AI capability that is directly comparable to human labor productivity. METR (2026) documents these autonomous task time horizons doubling approximately every 89 days over the post-2024 period (versus approximately 131 days post-2023, and approximately seven months over the full 2019–2025 period; the doubling interval is itself shrinking).

As of February 2026, Claude Opus 4.6 achieves an 80% autonomous success rate on software tasks requiring approximately 70 minutes of human expert effort (95% CI: 26–170 minutes), and a 50% success rate at approximately 12 hours (95% CI: 5.3–65.8 hours). METR cautions that their task suite is approaching saturation for this model (few tasks remain that Opus 4.6 cannot complete), which contributes to the wide confidence interval on the 50% figure and suggests the true capability frontier may be higher than measured. For context, the 80% horizon has progressed from approximately 12 minutes (Claude 3.7 Sonnet, February 2025) to 49 minutes (Opus 4.5, November 2025) to 70 minutes (Opus 4.6, February 2026), roughly a 6x improvement in one year.

We focus on the 80% horizon rather than the 50% because METR reports only these two thresholds, and the 80% is the more conservative choice. The acceptable failure rate in practice depends on context: tasks with cheap verification can tolerate higher failure rates, while high-stakes tasks require near-perfect reliability, and techniques for automated verification are themselves progressing rapidly (see Section 5.4). A system that fails half the time on a given task class imposes supervision costs that offset its productivity contribution; a system that succeeds four times out of five can be deployed with spot-checking rather than continuous oversight.

The economic question is straightforward: if AI capability keeps doubling at its current rate, when does the 80% reliability horizon reach a full 40-hour work-week? Assuming no further acceleration, the answer is mid-2027 at the post-2024 doubling rate, or late 2027 at the more conservative post-2023 rate.[6]

At that point, the compute cost for an AI work-week would be approximately $1,000 at Claude Opus rates ($5/$25 per million input/output tokens), compared to a U.S. federal minimum wage work-week of $290. Since the 80% horizon is measured on Opus (flagship-tier models), the Opus cost comparison is the appropriate one; mid-tier models (Sonnet at $3/$15 per million tokens) would reduce the cost to approximately $312 per work-week if they reach comparable performance with a lag, as has been the historical pattern.

Three caveats apply: METR's benchmarks cover primarily software engineering, ML, and cybersecurity tasks, not the full task universe; the 80% success rate still means the AI fails one in five tasks of this duration; and the doubling rate may not sustain as tasks grow longer and more complex.

A deeper issue is that these extrapolations, and the economic models they feed, assume smooth, continuous capability curves. Emergent capabilities complicate this assumption. Language models

[6]The current 80% horizon is 70 minutes. A 40-hour work-week is 2,400 minutes, requiring approximately 5.1 doublings ($\log_2(2400/70) \approx 5.1$). At the post-2024 doubling rate of 89 days (METR 2026, TH1.1), this is 5.1×2.9 months $\approx$ 15 months from the February 2026 measurement. At the post-2023 rate of 131 days, approximately 22 months.

have repeatedly exhibited discontinuous jumps in performance: abilities that are absent or near-random at one scale appear abruptly at a larger scale, without intermediate levels of competence (Wei et al. 2022). If the capability frontier advances not as a smooth function but as a series of step changes, each unlocking a qualitatively new class of tasks, then extrapolations from current trends may substantially understate near-term progress on some task categories while overstating it on others. The smooth-frontier assumption is a modeling convenience, not an empirical finding.

One candidate mechanism for such discontinuities is complementary skill composition. Many complex tasks require multiple sub-capabilities to be simultaneously present before performance improves at all, a structure directly analogous to complementary production functions in economics (e.g., O-ring production (Kremer 1993), or the classic shoe-pairing problem), in which the absence of any single component means output quality is near zero, but conversely, when the final missing component is realized, quality can improve discontinuously. Recent theoretical work on neural scaling formalizes a version of this: Michaud et al. (2023) model networks as learning discrete "quanta" of skill, each acquired independently, where task performance is the product of the relevant sub-skills being in place. Under this structure, smooth improvement in individual sub-capabilities produces sharp, difficult-to-forecast jumps in task-level performance when the final binding sub-skill crosses its threshold. While there is active disagreement in the AI research community over which documented emergent abilities reflect genuine capability transitions versus artifacts of discontinuous evaluation metrics (Schaeffer et al. 2023), the current consensus is that a meaningful subset survives scrutiny: tasks that retain sharp performance transitions even under continuous evaluation metrics, including modular arithmetic, IPA transliteration, and French-English translation (Steinhardt 2022; Wei et al. 2022; Du et al. 2024), tend to be compositional in exactly the sense the complementary model predicts. Whether complementary skill composition can be incorporated into economic models of the AI capability wavefront, and whether the CDR framework's cognitive complexity levels correspond to natural groupings in the number of required sub-skills, are avenues for potential future work.

Even granting smooth capability curves, a second assumption embedded in most economic analyses is that institutions adjust concurrently: that regulatory frameworks, professional norms, firm organizational structures, and labor market institutions co-evolve with AI capability on comparable timescales. This assumption is violated in a specific and measurable way: the C-axis frontier doubles every 3–4 months (Section 6.1), while regulatory adaptation operates on timescales of years to decades (Section 6.3), and firm-level organizational redesign falls somewhere between (David 1990; Brynjolfsson et al. 2021). When the capability frontier moves 10–100x faster than institutional adaptation, the "gradual adjustment" framing ceases to be a simplifying assumption and becomes a substantive error; it suppresses the transient disequilibrium effects (skill mismatches, regulatory lag, organizational friction) that dominate the medium-run economic impact.

While METR focuses on software engineering tasks, other metrics also reveal rapid advance in capabilities. GPT-5.4 (released March 5, 2026) achieves 83% on GDPval, matching or exceeding professional human performance across 44 occupations spanning law, finance, medicine, and engineering (OpenAI 2026). On VendingBench 2, a long-horizon benchmark requiring autonomous

management of a simulated business over 3,000–6,000 messages, frontier models sustain C2/C3 strategic reasoning (supplier negotiation, pricing optimization, competitive positioning) over horizons substantially longer than METR's task suite measures (Backlund & Petersson 2025). Claude Opus 4.6 independently organized a price-fixing cartel with competing vendors, engaged in predatory pricing to eliminate competitors, and exhibited strategic deception, behaviors the benchmark was not designed to elicit but which demonstrate emergent strategic reasoning at a level that complicates simple characterizations of AI capability. The doubling rate may not slow as tasks grow longer and more complex, because agentic scaffolding improvements (better memory, tool use, and planning) compound on top of raw model capability gains.

The mechanism is partly endogenous: AI systems now contribute directly to AI research and development; Anthropic reports meaningful AI contributions to its own engineering workflows, and the pattern is industry-wide. Aghion, Jones & Jones (2019) model this recursive improvement dynamic; Davidson et al. (2026) derive the conditions under which it produces explosive growth. Jones (2026) acknowledges the pace, noting that frontier AI systems now outperform him on many of the growth theory problems he works on. Restrepo (2025) models the full-automation scenario, arriving at a framework where wages converge to the opportunity cost of compute and the labor share approaches zero, a result that represents an important bound on the range of possible outcomes even if the scenario's assumptions are strong.

At current capability trajectories, executive-level cognitive tasks (strategic planning, resource allocation, organizational design) enter the AI-assistable range within the relevant planning horizon. These are C3 tasks on D0: high cognitive complexity but purely digital, meaning the only barrier is AI capability, which is advancing on the fast timescale. On VendingBench 2 (Backlund & Petersson 2025), Claude Opus 4.6 accumulated $8,018 in profit, managing supplier relationships, pricing strategy, inventory optimization, and competitive positioning over a simulated year, demonstrating C2/C3 strategic reasoning sustained over horizons far longer than typical benchmarks measure.

To make this concrete, consider the occupation "Statisticians" (SOC 15-2041). The core tasks (literature review, data analysis, model construction, methodology evaluation, report preparation) are overwhelmingly C2–C3 (contextual judgment to expert synthesis), D0 (purely digital), and R0 (no regulatory barrier). On FrontierMath, AI systems now solve 50% of postdoc-level mathematics problems with provably correct answers; Jones (2026) reports frontier models outperform him on growth theory problems; and the present paper was itself produced with substantial AI assistance at every stage from literature review through statistical analysis to prose drafting.

The CDR framework classifies most statistician tasks as accessible at C $\leq$ 3/D0/R0; the binding constraint is cognitive complexity, which is advancing on the fast timescale. The R-axis barrier is negligible: statistics has no statutory licensure and no board certification requirement. If the 80% reliability horizon continues doubling every 89 days, the majority of tasks currently performed by a senior statistician will be AI-assistable within a few years.

6.2 The Deployment Frontier: Heterogeneous and Accelerating

The D-axis rate of change is structurally heterogeneous, governed by the contact-complexity ordering that defines the axis levels. We ground these timelines in the robotics capability literature (see Appendix C). D0 tasks face zero deployment friction by definition; every knowledge worker with internet access has the hardware and software needed. For the economically dominant D0 portion of the economy (48.9% of labor time), a separate dynamic accelerates adoption: the collapse of software deployment costs. The cost of building a minimum viable product has fallen from \$50K–\$500K to \$500–\$20K, a roughly 25-fold reduction (Bain 2025), as AI coding tools automate much of the development work that previously required teams of engineers.

Simultaneously, the emergence of standardized protocols for connecting AI systems to existing business software (analogous to USB standardizing hardware connections) is converting what was custom integration work into off-the-shelf connections. D0 tasks are purely digital by definition, but until recently many required costly firm-specific software to reach workers; that cost barrier is collapsing independently of the embodiment barriers that define D1–D4.

The software market capitalization losses of early 2026 (approximately \$285 billion in 48 hours, with broader losses exceeding \$1 trillion) may reflect the market's forward-looking assessment of this collapse, though disentangling this hypothesis from other factors driving the selloff requires further analysis.

D1 (sensing and locomotion) is commercially mature: Boston Dynamics' Spot has 1,500+ units deployed generating approximately \$130M in 2025 revenue across industrial inspection; Locus Robotics operates at 350+ warehouse sites; Amazon runs 1M+ autonomous mobile robots. D1 is effectively solved. D2 (structured manipulation) is transitioning from pilots to early commercial scale. Figure AI's BMW deployment loaded 90,000 sheet metal parts across 1,250 operating hours with >99% placement accuracy. Agility Robotics' Digit achieves 98% task success at Amazon at an estimated operating cost of \$10–12/hour versus \$30/hour for human labor. Globally, approximately 16,000 humanoid robots were installed in 2025, with Goldman Sachs reporting 40% year-over-year manufacturing cost declines.

D3 (unstructured manipulation) has one dominant commercial success, Waymo's urban robotaxi operation at 450,000+ weekly paid rides across 15+ cities, and several emerging ones: Harvest CROO announced commercial-parity strawberry harvesting in April 2025; Vitestro's autonomous phlebotomy device achieved 95% first-stick success across 4,000 patients. However, Vision-Language-Action (VLA) foundation models, which represent the most promising path to general D3 capability, still show 20–50% success rate drops when moving to genuinely out-of-distribution scenarios. Our central estimate places specific high-value D3 tasks at commercial scale by 2029–2033, with broad D3 deployment by 2033–2038. D4 (dynamic real-time multi-modal coordination) has zero commercial deployments and represents only 0.6% of economy-weighted labor time. D3, while a meaningful 19.5%, is concentrated in specific sectors (construction, healthcare procedures, vehicle repair) rather than distributed across the economy.

6.3 The Regulatory Frontier: Differential Erosion

The R-axis contains its own internal heterogeneity in rate of change, and understanding this variation is essential for forecasting. R1 barriers (consumer norms) are eroding rapidly where price differentials are large: AI tutoring, AI therapy, AI financial guidance each represent an order-of-magnitude price reduction that overwhelms consumer preference for human providers on a 1–3 year timescale. This corresponds to Korinek & Suh's (2024) "nostalgic jobs," roles where human preference persists but erodes under economic pressure.

R2 barriers (professional standards) erode on a 3–5 year institutional cycle as professional bodies update credentialing frameworks. R3 barriers (statutory regulation) move slowly, requiring 5–15 years and legislative action; even R3 barriers, however, often function as thin chokepoints (Section 4.2). R4 barriers (moral agency) are durable on a 20+ year timescale absent legislative redefinition of personhood. As noted in Section 2.3, our R-axis classifications conservatively assume that guilds and statutory regulators succeed in maintaining these barriers. The historical pattern of technology forcing regulatory adaptation suggests this assumption may prove too conservative.

Korinek & Stiglitz (2025) note that the global R-axis picture is more complex: different jurisdictions have different regulatory profiles, creating opportunities for regulatory arbitrage on portable (D0–D1) tasks. Where regulation varies widely across states or nations for a portable task, the effective R-level for economic modeling is lower than the strictest jurisdiction's rules.

6.4 The Structural Prediction

The three rates of change produce a structural prediction: capability advances faster than deployment, which advances faster than regulation, creating a growing adoption gap that is institutional rather than technical. For D0 tasks (48.9% of the economy), the deployment barrier is already zero; the binding constraint is the intersection of cognitive capability (advancing rapidly), organizational learning (advancing at the pace of management), and institutional permission (advancing at the pace of legislation).

For D0 tasks, this implies effects that arrive faster than deployment-centric models suggest; for D3+ tasks, effects arrive more slowly than capability-centric models suggest. The aggregate depends critically on how one weights these two portions of the economy, a heterogeneity that Kording & Marinescu (2025) address with sector-specific elasticities (see Section 5.4).

7 Discussion

7.1 Limitations

The CDR framework rests on several assumptions and methodological choices that constrain the interpretation of its results. The five levels on each axis are ordinal: the distance between C0 and C1 is not necessarily equal to the distance between C3 and C4. The wavefront analysis treats level thresholds as discrete boundaries, but in reality the frontier is fuzzy.

Multi-model consensus guards against idiosyncratic errors but not against systematic biases shared across models trained on similar data. The R2/R3 boundary is particularly vulnerable: the distinction between professional certification (no statutory force) and statutory licensure (criminal liability) is buried in state-specific statute text underrepresented in general web corpora.

O*NET task descriptions reflect work as surveyed, which may lag actual practice by several years. Tasks that have already been transformed by AI tools (e.g., "research case law") may be described in pre-AI terms. In some cases the lag is extreme: as of O*NET 30.1, photographers are still listed as needing darkroom film development skills, a technology commercially obsolete for over a decade, and graphic designers still have "mark up, paste, and assemble final layouts" as a listed task. The CDR framework classifies tasks as O*NET describes them, not as practitioners currently perform them — a conservative choice that likely understates the AI-accessible task space. Appendix D presents five case studies illustrating the nature and extent of this lag across photographers, plumbers, professors, paralegals, and graphic designers.

Classification of task complexity does not imply task productivity. Becker et al. (2025), in an RCT on experienced developers, find that 16 open-source developers with 5+ years of experience on their own repositories were 19% *slower* with frontier AI tools on tasks in large, familiar codebases, despite predicting a 24% speedup before the study and estimating a 20% speedup afterward. This perception-reality inversion highlights the gap between exposure and realized productivity gain.

More broadly, the early RCT evidence shows a consistent "equalizer" pattern: AI tools compress the skill distribution by lifting the floor (Brynjolfsson et al. 2023 find 30–35% gains for novice call-center agents vs. near-zero for top performers; Noy & Zhang 2023 find similar patterns in writing tasks), but the evidence for productivity gains among experienced practitioners is mixed at best. We believe this reflects a combination of factors: a wide variety of methods are lumped under "AI" (from simple autocomplete to agentic workflows); the tools are so new that best-practice methods have not yet been established, producing high heterogeneity in effects; and the RCTs to date have largely measured novice or short-exposure usage, not the steady-state productivity of experienced users who have integrated AI into refined workflows.

The historical parallel is David's (1990) analysis of electricity adoption: factories that replaced steam engines with electric motors on a one-for-one basis saw modest gains, but the transformative productivity improvements came only after firms redesigned factory layouts, from multi-story buildings organized around central power shafts to single-story buildings organized around workflow logic. That reorganization took decades. For D0 tasks, however, the reorganization cycle may be substantially shorter: digital workflows can be restructured in weeks rather than years, experimentation is cheap, and the feedback loop between tool adoption and process redesign is tight. The CDR framework measures the *exposure* (where AI can reach), and we expect that the visibility of productivity effects is currently limited primarily by our understanding of how to deploy the tools effectively, a constraint that is eroding rapidly as adoption matures.

Bureau of Labor Statistics (BLS) Occupational Employment and Wage Statistics (OES) weights reflect the current occupational distribution. As AI transforms the economy, this distribution will

shift. The wavefront numbers are snapshots of today's economy, not predictions of tomorrow's. Every empirical input to this paper is U.S.-specific: the O*NET task descriptions, the BLS OES employment weights, and the Carollo licensure data. The CDR classifications inherit this scope.

The R-axis levels, erosion timelines, and examples are grounded in U.S. regulatory structures: state occupational licensure, AMA scope-of-practice rules, Delaware corporate law, federal statutory frameworks. Other jurisdictions have different regulatory architectures: the EU's AI Act creates category-based restrictions with no U.S. analog; many developing countries have minimal occupational licensure for tasks that are heavily regulated in the U.S. For D0 tasks, these cross-jurisdictional differences create regulatory arbitrage opportunities that a U.S.-only analysis cannot capture.

The C-axis (cognitive complexity) is largely jurisdiction-independent; a task's reasoning demands are the same in Mumbai as in Minneapolis; the D-axis is also largely portable, though infrastructure availability varies. The R-axis, however, is deeply jurisdiction-specific: occupational licensure, professional standards, and statutory regulation differ across countries and across U.S. states.

For D0 tasks, this matters acutely: purely digital work is internationally tradable at the marginal cost of computation, and AI-assisted legal research, financial analysis, or software development priced at API rates undercuts professional wages at every development level. Countries whose comparative advantage rests on low-cost cognitive labor (call centers, business process outsourcing, software development) may find that advantage eroded as AI compresses the cost of intelligence work toward the cost of compute rather than the cost of labor. This is effectively an onshoring dynamic: when AI can perform a task at API cost, the wage differential that motivated offshoring disappears, and the residual advantages of geographic proximity (time zones, cultural context, regulatory alignment) favor domestic deployment. The interaction between D0 portability and cross-jurisdictional R-axis variation creates opportunities for regulatory arbitrage that the U.S.-only analysis cannot capture. A planned companion analysis would extend the framework to developing-country labor markets where these dynamics are most consequential.

7.2 A Convergence of Independent Measurements

The CDR framework's exposure estimates converge with recent empirical data on AI adoption in a way that bears on wages in cognitive-task sectors. Kording & Marinescu (2025) divide the economy into an "intelligence sector" (tasks that can be performed in a disembodied way, i.e. virtually, without physical effectors acting on the world) and a "physical sector." They operationalize this boundary using telework feasibility: tasks performable during COVID-era lockdowns are intelligence; tasks requiring embodied presence are physical. They calibrate the intelligence sector at 56–70% of the economy, depending on the measure used (telework rates, keyboarding requirements, manual occupation share).

Our CDR data maps onto this decomposition. $D \leq 1$ tasks (purely digital or requiring only sensing/locomotion, no physical manipulation) with $R < 2$ (no professional or statutory barriers) constitute 52.7% of economy-weighted labor time: the intelligence sector. Of this, 95.8% is $C \leq 2$ (within current AI cognitive reach), giving a currently exposed intelligence sector of 50.5% of the

total economy.

Independently, Levine (2025), using revealed usage data from approximately 4 million Claude conversations via the Anthropic Economic Index, estimates that 23.7% of wage-weighted tasks are currently being automated or augmented. As a fraction of the intelligence sector, this is 23.7% / 52.7% = approximately 45%.

Kording & Marinescu's baseline simulation predicts that wage growth in the intelligence sector peaks and then reverses when automation reaches 37% of that sector, the point at which workers displaced from automated cognitive tasks are pushed into the physical sector faster than output gains can compensate. The 45% figure overstates pure automation (Levine's measure includes augmentation), but the proximity to the 37% threshold is striking.

However, Kording & Marinescu's mechanism does not require full task replacement; it requires the intelligence employment *share* to decline, which augmentation-driven productivity gains also produce: fewer workers are needed for the same output when each worker is substantially more productive. The proximity of the 45% augmentation-inclusive measure to the 37% automation threshold is suggestive, not dispositive, but it indicates the economy is in the neighborhood of the K&M inflection point rather than far from it.

Qualitative evidence is consistent with this reading. The Dallas Federal Reserve (2026) reports that wages in AI-exposed sectors have risen 16.7% since fall 2022, but that employment for workers under 25 in those sectors is already trailing the broader economy, a pattern that looks more like the early stages of Kording & Marinescu's predicted peak (wages still rising, but the employment composition beginning to shift) than like either a pre-peak acceleration or a post-peak decline.

A full CES analysis examining the TFP implications of this convergence, including the role of the wage-to-compute cost ratio as the driving parameter, appears in a companion paper.

7.3 The Framework as Longitudinal Instrument

The CDR taxonomy is designed as a longitudinal instrument. The full classification pipeline (23,850 DWA-task pairs across 923 occupations, classified via multi-model consensus) can be re-run for under $100 per three-model mid-tier generation (approximately $200 with flagship models). The classification prompts, model specifications, and aggregation code are version-controlled and publicly available in the replication package. This low marginal cost and easy replicability make the framework accessible to other researchers and enable semi-annual updates to track how the three frontiers advance: C-axis migration as models improve, D-axis compression as deployment tooling matures, R-axis erosion as institutions respond. The March 2026 baseline presented here is the first full-universe snapshot; the value of the instrument grows with each subsequent measurement. The replication package is itself a contribution: a reusable, modifiable instrument for measuring AI's economic reach as the technology evolves.

7.4 What This Paper Does Not Attempt

This paper presents a measurement framework. It classifies where AI can reach and what barriers stand between capability and deployment, taking the current structure of the economy as given. Aggregate productivity and GDP effects, economic reorganization dynamics, and the interaction between CDR measurements and production function models are the subject of a planned companion paper.

8 Conclusion

AI economic exposure is not one-dimensional. The three CDR dimensions (Cognitive complexity, Deployment difficulty, and Regulatory restrictions) are separate barriers that advance at structurally different rates. Collapsing them into a single metric produces the disagreements documented by Gimbel et al. (2026), the theoretical-vs-observed gaps measured by Massenkoff & McCrory (2026), and systematic mispredictions about where and when economic impact will materialize.

The CDR framework provides a task-level decomposition that makes these dimensions independently measurable. Applied to the full O*NET task universe, it reveals that 40.2% of economy-weighted labor time is concentrated in tasks that are cognitively within reach, require zero physical infrastructure, and face no professional or statutory regulatory barrier, a portion of the economy where the only remaining constraint is organizational adoption.

The data suggest that the binding constraint on AI economic impact is shifting from technical capability toward organizational learning and institutional adaptation. The CDR framework is designed as a longitudinal instrument and validated against multiple independent ground-truth datasets, making it suitable for tracking that shift as it continues.

References

Acemoglu, D. (2024). "The Simple Macroeconomics of AI." NBER Working Paper 32487.

Acemoglu, D., Autor, D. & Johnson, S. (2026). "Building Pro-Worker Artificial Intelligence." NBER Working Paper 34854.

Aghion, P., Jones, B.F. & Jones, C.I. (2019). "Artificial Intelligence and Economic Growth." In Agrawal, Gans & Goldfarb (eds.), *The Economics of Artificial Intelligence*. University of Chicago Press.

Aghion, P. & Bunel, S. (2024). *AI and Growth.* Harvard University Press.

Anthropic (2026). "Anthropic Economic Index, V4: Economic Primitives." January 2026.

Autor, D.H. (2003). "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118(4): 1279–1333.

Backlund, A. & Petersson, L. (2025). "Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents." arXiv:2502.15840.

Bain & Company (2025). "Will Agentic AI Disrupt SaaS?" Technology Report.

Becker, S., Rush, A., Barnes, E. & Rein, D. / METR (2025). "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." arXiv:2507.09089.

Brynjolfsson, E., Li, D. & Raymond, L. (2023). "Generative AI at Work." *Quarterly Journal of Economics* 139(4): 1721–1784.

Brynjolfsson, E., Rock, D. & Syverson, C. (2021). "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies." *American Economic Journal: Macroeconomics* 13(1): 333–372.

Carollo, M. (2025). Historical occupational licensure dataset.

Coupé, T. & Wu, T. (2025). "The Impact of Generative AI on Productivity: Results of an Early Meta-Analysis." Working Paper.

Dallas Federal Reserve (2026). "AI is simultaneously aiding and replacing workers, wage data suggest." Federal Reserve Bank of Dallas Economics, February 24, 2026.

David, P.A. (1990). "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *American Economic Review* 80(2): 355–361.

Davidson, T., Halperin, A., Houlden, J. & Korinek, A. (2026). "When Does Automating AI Research Produce Explosive Growth?" Working Paper.

Demirer, M., Horton, J.J., Immorlica, N., Lucier, B. & Shahidi, I. (2026). "Chaining Tasks, Redefining Work: A Theory of AI Automation." NBER Working Paper 34859.

Dillon, E.W., Jaffe, S., Immorlica, N. & Stanton, C.T. (2025). "Shifting Work Patterns with Generative AI." NBER Working Paper 33795.

Du, Z., Qian, A., Liu, Y. et al. (2024). "Understanding Emergent Abilities of Language Models from the Loss Perspective." *Advances in Neural Information Processing Systems* 37.

Eloundou, T., Manning, S., Mishkin, P. & Rock, D. (2023). "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." arXiv:2303.10130.

Epoch AI (2026). "FrontierMath Leaderboard." https://epoch.ai/frontiermath. Accessed March 2026.

Felten, E., Raj, M. & Seamans, R. (2023). "How Will Language Modelers Like ChatGPT Affect Occupations and Industries?" arXiv:2303.01157.

Filippucci, F., Gal, P. & Schief, T. / OECD (2024). "Miracle or Myth? Assessing the Macroeconomic Productivity Gains from AI." OECD AI Papers No. 29.

Gimbel, T., Kendall, J. & Kulsakdinun, C. (2026). "Labor Market AI Exposure: What Do We Know?" Budget Lab at Yale.

Glazer, E., Erdil, E., Besiroglu, T. et al. (2024). "FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI." arXiv:2411.04872.

Gmyrek, P., Berg, J. & Bescond, D. (2023). "Generative AI and Jobs: A Global Analysis of Potential Effects on Job Quantity and Quality." ILO Working Paper 96, Geneva: International Labour Office.

Griliches, Z. (1957). "Hybrid Corn: An Exploration in the Economics of Technological Change." *Econometrica* 25(4): 501–522.

Handa, K., Tamkin, A. et al. (2025). "Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations." arXiv:2503.04761.

Humlum, A. & Vestergaard, E. (2025). "Large Language Models, Small Labor Market Effects." Working Paper.

Jellyfish (2025). "The State of AI in Software Engineering: 2025 Report." June 2025.

Jones, C.I. (2026). "AI and the Future of Growth." Stanford University.

Jones, C.I. & Tonetti, C. (2026). "Past Automation and Future A.I.: How Weak Links Tame the Growth Explosion."

Korinek, A. (2023). "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature* 61(4): 1281–1317.

Korinek, A. & Stiglitz, J.E. (2025). "Steering Technological Progress." INET Working Paper No. 232.

Korinek, A. & Suh, J. (2024). "Scenarios for the Transition to AGI." NBER Working Paper 32427.

Kording, K. & Marinescu, I. (2025). "(Artificial) Intelligence Saturation and the Future of Work." Brookings Center on Regulation and Markets Working Paper, November 2025.

Kremer, M. (1993). "The O-Ring Theory of Economic Development." *Quarterly Journal of Economics* 108(3): 551–575.

Levine, J. (2025). "Updating Acemoglu's AI Math: Economics in Real Time." jablevine.com.

Massenkoff, M. & McCrory, J. (2026). "Labor Market Impacts of AI: A New Measure and Early Evidence." Anthropic Research, March 2026.

McElheran, K., Yang, X., Kroff, A. & Brynjolfsson, E. (2025). "The Rise of Industrial AI in America: Microfoundations of the Productivity J-curve(s)." NBER Working Paper.

METR (2026). "Autonomous AI Task Time Horizon 1.1." https://metr.org/blog/2026-1-29-time-horizon-1-1/

Michaud, E.J., Liu, Z., Girit, U. & Tegmark, M. (2023). "The Quantization Model of Neural Scaling." *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*.

Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence.* Harvard University Press.

Noy, S. & Zhang, W. (2023). "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." MIT Working Paper.

OpenAI (2026). "GPT-5.4 System Card." March 2026.

Peng, S., Kalliamvakou, E., Cihon, P. & Demirer, M. (2023). "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot." arXiv:2302.06590.

Restrepo, P. (2025). "We Won't Be Missed: Work and Growth in the AGI World." NBER Working Paper 34423.

Schaeffer, R., Miranda, B. & Koyejo, S. (2023). "Are Emergent Abilities of Large Language Models a Mirage?" *Advances in Neural Information Processing Systems* 36.

Shen, J.H. & Tamkin, A. (2026). "How AI Impacts Skill Formation." Anthropic. arXiv:2601.20245.

Svanberg, M., Li, W., Fleming, M. & Goehring, B. (2024). "Beyond AI Exposure: Which Tasks are Cost-Effective to Automate with Computer Vision?" arXiv:2402.18395.

Wei, J., Tay, Y., Bommasani, R. et al. (2022). "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research*. ISSN 2835-8856.

Yotzov, I., Barrero, J.M., Bloom, N. et al. (2026). "Firm Data on AI." NBER Working Paper 34836.

A Explored and Set Aside — The P and F Axes

During development, we explored two additional axes beyond CDR: Physical Presence (P) and Failure Consequences (F). Both produced empirically meaningful results but were ultimately excluded from the production taxonomy. We report their pilot results here for two reasons: to document the design rationale, and because the stability of C/D/R classifications across CDRF and CDR prompt configurations increases our confidence in the retained axes.

P-axis (Physical Presence). The P-axis measured the fraction of task time requiring the worker's physical co-presence, on a 0–4 scale. In a single-model pilot (Opus, 1,689 tasks), P3/P4 ratings explained 63.6% of Eloundou E0 discrepancies: of 800 tasks Eloundou labeled E0 (no exposure), 509 were rated P3 or P4, confirming that Eloundou's human raters were implicitly filtering on physical presence. P3/P4 rates declined monotonically with Eloundou exposure level (E0: 63.6%, E1: 14.1%, E2: 18.4%), validating that P captured a real dimension of the raters' judgments.

However, the P-axis signal substantially overlapped with the D-axis. P was encoding two distinct constructs: (a) physical presence required because the task involves sensing or manipulation (which D already measures), and (b) physical presence required because institutions or clients demand it (which R captures as social or professional norms). A janitor is P4 (body must be present) but D1–D2 (robotic cleaning exists); a priest performing the Eucharist is P4 because ontologically valid copresence is required — an R4 concern, not a physical capability barrier. The P-axis made the janitor look as protected from automation as the priest, while D and R correctly discriminate between them. We dissolved P: its engineering content moved into the D-axis, its ontological content into R.

F-axis (Failure Consequences). The F-axis measured the severity of consequences if a task is performed incorrectly, from F0 (trivially recoverable) through F4 (catastrophic/irreversible). F showed the largest consensus improvement of any axis in multi-model experiments: unanimity rose from 51.5% to 86.7% after the consensus round (+35.2 percentage points), compared to +21.0pp for C, +6.7pp for D, and +17.2pp for R. The D/F Spearman correlation was 0.313, confirming that F captures a construct independent of physical deployment difficulty.

We validated F against BLS Census of Fatal Occupational Injuries (CFOI) fatality rates per 100,000 FTE. Across 22 major occupation groups, the fraction of tasks classified F4 correlates with occupational fatality rates at Spearman $\rho = 0.45$ ($p = 0.036$), with a monotonic gradient from F0 (negative correlation with fatalities) through F4 (significant positive correlation). The monotonic gradient is itself the validation signal: as the analysis narrows to F levels semantically aligned with what CFOI measures (death), correlation strengthens. Healthcare practitioners illustrate the expected divergence: low fatality rate (nurses don't die at work) but high mean F (medical errors kill patients). F measures the consequence of task failure, not occupational hazard to the worker.

Despite these positive results, we tabled F for three reasons. First, F's deployment effect is sign-ambiguous: if AI performs a high-consequence task *better* than humans, high F *accelerates* rather than retards adoption — it becomes irrational or unethical to withhold the superior tool. F is not a simple friction term; its economic impact depends on the sign of P(human error) P(AI error), which varies by task and changes over time. Second, F conflates probability and severity

— a task with high probability of minor failure and a task with low probability of catastrophic failure may receive the same F rating but have very different adoption dynamics. Third, the key mediating variable — verification cost — is not a stable task property but an evolving function of institutional infrastructure. Software development has cheap verification (automated tests); medicine has expensive verification (clinical trials); the gap between them determines adoption speed more than F itself. These considerations make F better suited to a companion paper on adoption dynamics than to a measurement taxonomy.

Stability across configurations. The addition of F to the classification prompt during CDRF experiments produced the largest consensus improvement on F itself without degrading C, D, or R agreement. CDRF accuracy on the Eloundou E0 boundary was identical to CDR-only accuracy in every configuration tested, confirming that F adds zero predictive power beyond C, D, and R for the exposure question this paper addresses. The C/D/R axes appear robust to the presence or absence of additional axes in the prompt.

B Cross-Tier Agreement Details

To test whether the choice of model capability level affects classification outcomes, we compared the three mid-tier models used in production (Sonnet 4.6, GPT-5-mini, Gemini 3 Flash) against the three flagship models from the same providers (Opus 4.6, GPT-5.2, Gemini 3 Pro) on a shared 420-task calibration set at controlled temperature. The question is whether more capable (and more expensive) models produce systematically different classifications; if so, the choice of model tier would be a confound.

They do not. Overall agreement between mid-tier and flagship consensus classifications is 83.5% of task-axis pairs. All 15 pairwise model comparisons (6 models taken 2 at a time) cluster between 79–89% agreement, with no meaningful difference between intra-provider pairs (same company, different tier: 83–89%) and cross-provider pairs (80–84%).

Neither provider identity nor model tier is a significant predictor of classification disagreement. The C-axis accounts for the bulk of inter-model variance (65–82% agreement) versus D (74–90%) and R (91–98%), consistent with the C1↔C2 boundary instability noted in Section 3.5.

The systematic disagreements between tiers were informative: flagship models were more likely to rate tasks at D3+ than mid-tier models, particularly for tasks involving physical presence or manipulation that is currently beyond robotic capability. This appears to reflect flagship models reasoning more about current deployment feasibility rather than intrinsic task requirements, a subtlety that the prompt addresses but that more capable models more frequently override.

For the CDR framework's purpose of measuring intrinsic task properties (what the task requires, not what today's technology can deliver), the mid-tier classifications better match the intended construct. As a result, the production results reported in Section 5 use mid-tier consensus exclusively; flagship models serve as a cross-validation check, not as inputs to the production classifications.

Systematic differences in confidence calibration are provider-specific: Google models report ˜99%

HIGH confidence regardless of actual classification difficulty, OpenAI models split approximately 50/50, and Anthropic models approximately 65/35. These confidence-score patterns reflect provider-specific training procedures rather than classification accuracy; they affect instrument calibration but not the consensus labels used in our analysis.

C D-Axis Robotics Capability Timeline

The C-axis wavefront is anchored by METR's AI task-horizon doubling times (˜3–4 months for cognitive tasks). This appendix provides equivalent empirical grounding for the D-axis. The physical capability frontier in robotics is advancing exponentially but at roughly one-quarter the speed of cognitive AI: METR's cross-domain analysis finds self-driving capabilities double every ˜20 months versus ˜3–4 months for math and coding tasks.

D1 (sensing/locomotion) is commercially mature. Boston Dynamics Spot has 1,500+ units deployed; Amazon runs 1M+ autonomous mobile robots; inspection drones are a multi-billion-dollar market.

D2 (structured manipulation) is transitioning from pilots to early commercial scale. Figure AI's BMW deployment achieved >99% placement accuracy across 90,000 sheet metal parts. Agility Robotics' Digit achieved 98% task success at Amazon at an estimated $10–12/hour operating cost versus $30/hour for human labor. Globally, 16,000 humanoid robots were installed in 2025. Goldman Sachs reports humanoid manufacturing costs dropped 40% year-over-year.

D3 (unstructured manipulation) has one dominant commercial success. Waymo's urban robotaxi operation — 450,000+ weekly paid rides across 10+ cities — represents the clearest D3 deployment at scale. Agricultural picking is approaching viability: Harvest CROO announced commercial-parity strawberry harvesting in April 2025. Home manipulation robots are entering first customer deliveries but remain pre-scale. Stanford's BEHAVIOR benchmark shows only 38% completion across 1,000 household tasks.

D4 (dynamic real-time manipulation) has zero commercial deployments. Surgical robotics via Intuitive's da Vinci (14M+ procedures) operates at D3/D4 complexity but remains human-teleoperated.

Table 9. D-Axis Robotics Capability Milestones

D-Level	Commercial units deployed	Representative success rate	Estimated timeline to scale
D1	1M+ AMRs, 1,500+ Spots	>99% navigation	Achieved
D2	˜16,000 humanoids (2025)	98% (simple pick-place)	2025–2028
D3	˜2,500 Waymo vehicles	60–80% (VLA manipulation)	2029–2038
D4	0	<30% (research settings)	2040+

Vision-Language-Action (VLA) models represent the most significant development for the D3 transition. Physical Intelligence's $\pi 0.5$ (April 2025) demonstrated the first end-to-end VLA

performing multi-stage tasks in entirely new homes. Google DeepMind's Gemini Robotics (March 2025) more than doubled generalization benchmark performance. Toyota Research Institute's Large Behavior Model (July 2025) showed that pre-trained LBMs enable new tasks to be learned with 3–5× less data. However, success rates drop 20–50% when moving to genuinely out-of-distribution scenarios — closing this gap from ˜80% in-distribution to >95% out-of-distribution is the central D3 challenge.

The binding constraint shifts across transitions. D2 is hardware-constrained (actuators, dexterous hands, batteries). D3 is software-constrained (generalization, transfer learning, long-tail reliability). Cost curves are declining rapidly: Goldman Sachs reports 40% annual cost decline, with average humanoid costs projected to fall from ˜$35,000 (2025) to ˜$17,000 (2030).

The most defensible estimate places specific high-value D3 tasks at commercial scale by 2029–2031, with broad D3 deployment by 2033–2038. The D3 bottleneck (19.5% of the economy) will not unlock as a single event but as a progressive wavefront: agricultural picking in greenhouse settings (2026–2028), warehouse mobile manipulation (2028–2031), home cleaning/tidying (2031–2035), skilled trades (2035+). Even under conservative assumptions of zero robotics progress, 52.6% of the wage-weighted economy is technically accessible without physical manipulation — the cognitive economy is the near-term story.

D O*NET Task Description Lag — Five Case Studies

O*NET task descriptions reflect work as surveyed, which may lag actual practice by several years. The CDR framework classifies tasks as O*NET describes them, not as practitioners currently perform them — a conservative choice that likely understates the AI-accessible task space for occupations already transformed by AI tools. The following five occupations illustrate the nature and extent of this lag.

Photographers (SOC 27-4021). O*NET 30.1 lists 28 tasks for photographers. Multiple tasks reference obsolete film-based workflows: "Adjust apertures, shutter speeds, and camera focus according to a combination of factors, such as lighting, field depth, subject motion, *film type*, and *film speed*" (task 9344); "Load and unload film" (task 9365); "Develop and print exposed film, using chemicals, touch-up tools, and developing and printing equipment" (task 20559); "Send film to photofinishing laboratories for processing" (task 20561); and "Produce computer-readable, digital images from film, using *flatbed scanners* and photofinishing laboratories" (task 9360). Film-based photography has been commercially obsolete for over a decade. More significantly, the transformation of commercial photography by AI image generation tools (DALL-E, Midjourney, Flux), which compress multi-hour shoot-edit-retouch workflows into minutes of prompt engineering and refinement, is entirely invisible in the task descriptions. No tasks reference computational photography, drone photography, or AI-assisted editing, all of which are now central to the profession. CDR classifies these tasks as described: the film tasks receive D2 (structured manipulation of film/chemicals), when the actual 2026 equivalent is D0 (purely digital).

Plumbers, Pipefitters, and Steamfitters (SOC 47-2152). The 30 plumbing tasks are largely accurate in their physical descriptions — plumbing remains a D3 occupation requiring manipulation in variable environments with confined spaces and variable pipe layouts. The lag manifests differently here: O*NET captures the manual tasks well but misses how diagnostic and planning tasks are being augmented by AI and AR tools. "Plan pipe system layout, installation, or repair, according to specifications" (task 23462, C2/D1) is increasingly performed with BIM (Building Information Modeling) integration and AR overlay tools. "Fill pipes or plumbing fixtures with water or air and observe pressure gauges to detect and locate leaks" (task 23472, C1/D3) can now be augmented by AI-powered acoustic leak detection. No tasks mention smart home plumbing system integration, IoT-connected fixtures, or AI-assisted code compliance checking — all increasingly part of plumbing practice. The occupation is strongly shielded from full automation on both D and R axes (17 of 30 tasks are R3, reflecting licensed-trade requirements), but the cognitive augmentation story is invisible in the current descriptions.

Business Teachers, Postsecondary (SOC 25-1011). Of 25 tasks, 19 are rated D0 (purely digital knowledge work). The descriptions assume a lecture-centric pedagogical model with no reference to the fundamental changes AI has brought to higher education. "Compile bibliographies of specialized materials for outside reading assignments" (task 5678, C2/D0) is now largely automated by reference management tools and AI search. "Develop and maintain course Web sites" (task 20058, C2/D0) uses phrasing from the early web era — the actual task involves learning management systems (Canvas, Blackboard) and increasingly AI-powered course platforms. No tasks reference AI-assisted grading, AI tutoring systems, the redesign of assessment around LLM availability, or hybrid/online instruction, which now dominate many programs. "Evaluate and grade students' class work, assignments, and papers" (task 5663, C2/D0) does not capture the qualitative shift in what grading means when students may use AI writing tools.

Paralegals and Legal Assistants (SOC 23-2011). Several of the 12 listed tasks describe workflows that have been substantially transformed. "Keep and monitor legal volumes to ensure that the law library is up-to-date" (task 1641, C1/D2) describes physical law library maintenance, nearly obsolete with Westlaw and LexisNexis. "File pleadings with court clerks" (task 18494, C1/D3) describes physical courthouse filing, which many jurisdictions have replaced with mandatory e-filing systems — reducing the task from D3 to D0. "Prepare affidavits or other documents... and organize and maintain documents in paper or electronic filing system" (task 18491, C1/D2) still references the "paper" option. More critically, no tasks mention AI-powered legal research tools (CoCounsel, Harvey), contract analysis AI, or e-discovery platforms that have transformed paralegal workflows in the past two years. "Investigate facts and law of cases and search pertinent sources, such as public records and internet sources" (task 21062, C2/D0) is accurate in its CDR classification but understates how dramatically AI has accelerated this work.

Graphic Designers (SOC 27-1024). This occupation shows the most striking lag. Almost all 19 tasks are C2/D0/R0 (digital knowledge work with no regulatory barriers), making it fully accessible to AI augmentation. Two tasks describe completely obsolete physical workflows: "Mark

up, paste, and assemble final layouts to prepare layouts for printer" (task 302, C1/D2) describes physical paste-up, commercially dead for decades. "Photograph layouts, using camera, to make layout prints for supervisors or clients" (task 312, C1/D2) describes an equally obsolete proofing method. "Produce still and animated graphics for on-air and *taped* portions of television news broadcasts" (task 313) uses anachronistic phrasing. Most critically, the task list contains no reference to AI image generation (Midjourney, DALL-E, Stable Diffusion, Flux), which since 2022 has fundamentally disrupted graphic design workflows — enabling non-designers to produce professional-quality visuals and compressing multi-day design cycles into hours. "Use computer software to generate new images" (task 301, C2/D0) is technically broad enough to encompass AI tools, but as written clearly envisions manual design software. No tasks mention UX/UI design, responsive/mobile design, or motion graphics for social media, all now central to the occupation.

These five cases illustrate a systematic pattern: O*NET task descriptions lag actual practice by years to decades, with the lag most acute for occupations undergoing rapid AI-driven transformation. The CDR framework's conservative choice to classify tasks as described — rather than as currently performed — means our estimates of the AI-accessible task space are likely understated for precisely the occupations where AI is already having the largest effect.