# Disutility-Weighted Automation Prioritization: Directing Agentic R&D Toward High-Hazard, Low-Intrinsic-Value Human Toil

**Dima Baker**
Collective Intelligence Network

November 2026

**Abstract**

Market-driven artificial intelligence R&D allocates capital toward tasks with the highest nominal wage bill or lowest digital friction. Empirical labor audits—including the Stanford WORKBank study of 844 occupational tasks and task-based models of technological change—reveal a structural tension: unconstrained wage-bill optimization targets high-wage, high-intrinsic-value cognitive and creative craft while leaving workers in hazardous physical toil, routine drudgery, and supervisory cleanup loops. We propose Disutility-Weighted Automation Prioritization (`DWAP-Bounty`), a mechanism-design framework that steers R&D budgets using occupational hazard $h_j$ and noisy consensus agency preferences $\hat{v}_j$ under inter-worker preference disagreement ($\sigma_{\text{pref}}$). Using an illustrative 24-descriptor task catalog grounded in O*NET and WORKBank profiles ($J = 60$ paired tasks across budgets $B \in \{15\%, 30\%, 45\%, 60\%\}$) with equalized R&D conversion and handover risk functions, we compare five policies and trace the Pareto frontier between wage surplus and toil elimination. At $B = 45\%$ ($M_{\text{phys}} = 1.5\times, \sigma_{\text{pref}} = 0.18$), `DWAP-Bounty` eliminates 60.49% [60.45, 60.53] of hazardous toil and achieves 74.29% [74.23, 74.35] worker-agency alignment, whereas Constrained Surplus Optimization achieves higher market surplus (48.29% vs. 37.26%) at lower toil elimination (51.23%), and Market Wage Greedy achieves 53.41% surplus but 44.80% toil elimination and 39.11% alignment. Sweeps across toil prevalence $\pi_{\text{toil}} \in [0.15, 0.50]$, disagreement $\sigma_{\text{pref}} \in [0.05, 0.32]$, and robotics cost $M_{\text{phys}} \in [1.0\times, 6.0\times]$ show that the 4.0× embodied-cost crossover is conditional on this budget model.

## 1 Introduction

A foundational promise of technological automation is the liberation of human beings from dangerous, exhausting, and repetitive drudgery so that human attention can flourish in creative inquiry, interpersonal care, and high-agency craftsmanship [15, 18]. In classic economic models of skill-biased and task-based technological change, Autor et al. [3] and Acemoglu and Restrepo [1] showed that traditional industrial and software automation primarily substituted for routine manual and routine cognitive tasks while complementing non-routine analytical and interpersonal roles. However, the deployment of generative and agentic AI between 2024 and 2026 has altered this classical equilibrium: because foundation models operate at low marginal cost over digital text, code, and media, venture and enterprise R&D capital increasingly targets high-wage, non-routine creative and analytical occupations while under-investing in high-hazard physical manipulation and tedious operational maintenance [12, 19].

### 1.1 The "Reverse Centaur" pathology and the WORKBank agency audit

Empirical evidence across human-AI interaction and labor economics highlights the social cost of undirected automation. When Shao et al. [19] audited 844 occupational tasks across the U.S. workforce using the WORKBank database and the 5-point Human Agency Scale (HAS), they uncovered a sharp mismatch between developer R&D targets and worker preferences: workers strongly desire full automation (`HAS-1`/`HAS-2`) for repetitive scheduling, hazardous inspection, and boilerplate data entry, yet demand high human agency (`HAS-4`/`HAS-5`) over tasks involving creative synthesis, domain craftsmanship, and interpersonal judgment. Simultaneously, longitudinal evaluations of autonomous agent task completion horizons by Kwa et al. [12] show that while agents excel on short-horizon digital subroutines, reliability drops sharply on long-horizon, open-ended projects.

When organizations deploy partially reliable agents driven purely by wage-bill minimization (`Market Wage Greedy`) or digital ease (`Feasibility First`), premature replacement of high-agency work produces what Bainbridge [4] famously termed the *ironies of automation*: machines execute the expressive or routine synthesis steps while human operators are relegated to high-stress, low-autonomy exception monitoring and post-hoc error cleanup [6, 9]. Empirical studies confirm that uncalibrated AI delegation degrades junior skill formation [20], accumulates cognitive debt [11, 17], erodes independent critical thinking under time pressure [13, 24], and induces strategic confidence miscalibration [2, 8].

### 1.2 Redirecting R&D via disutility weighting and shared-control governance

Reversing this misallocation requires treating automation R&D prioritization as an explicit multi-objective mechanism-design problem under worker preference heterogeneity [1, 23]. When a cooperative collective or public R&D consortium allocates a finite automation budget $B$, which tasks should receive full-replacement R&D bounties, and which should be reserved for human-steered augmentation [5, 21]? Grounded in the task-based framework of Acemoglu and Restrepo [1], the WORKBank agency audit of Shao et al. [19], and inalienable human-in-the-loop governance [5, 6], we formalize Disutility-Weighted Automation Prioritization (`DWAP-Bounty`), where full-automation priority is monotonic in occupational hazard $h_j$ and inverse in noisy observed worker agency desire $\hat{v}_j$, and compare it directly against constrained market-surplus optima and the full Pareto frontier.

### 1.3 Contributions

In this position and formal modeling paper, we make five contributions:

1. **Taxonomy of worker-centric automation steering (Section 2).** We synthesize classical task-displacement theory [1, 3], systematic meta-analyses of human-AI synergy [21], and 2025–2026 agency and shared-control research [5, 6, 19] into a five-part taxonomy (Table 1).

2. **Formal disutility-bounty, preference disagreement, and Pareto-frontier model (Section 3).** We formulate R&D portfolio allocation over an illustrative 24-descriptor occupational task catalog $\mathcal{J}$ with hazard $h_j$, routine index $r_j$, heterogeneous individual worker preferences $v_{j,w}$ ($\sigma_{\text{pref}} \in [0.05, 0.32]$), noisy planner agency estimates $\hat{v}_j$, market wage $w_j$, and physical-to-digital cost multiplier $M_{\text{phys}}$. We evaluate all policies on four independent outcomes under strictly equalized R&D efficiency and handover risk functions and trace the multi-objective Pareto frontier (Equations 1–4).

3. **Three-stage `DWAP-Bounty` architecture and parameter provenance (Section 4).** We specify an operational architecture integrating WORKBank preference elicitation [7, 19],

shared-control handover gates [5, 6], and counterfactual Shapley transition dividends [14, 22] (Figure 1, Table 2, Table 3).

4. **Paired Monte Carlo benchmark and three sensitivity sweeps (Section 5).** We execute a paired Monte Carlo evaluation (`experiment.py`) comparing `DWAP-Bounty`, Constrained Surplus Optimization (`constrained_surplus_opt`), Routine-Biased Replacement [3], Feasibility-First, and Market Wage Greedy across budgets $B \in \{15\%, 30\%, 45\%, 60\%\}$, alongside sensitivity sweeps over physical-toil prevalence $\pi_{\text{toil}} \in \{0.15, 0.33, 0.50\}$, inter-worker preference disagreement $\sigma_{\text{pref}} \in \{0.05, 0.18, 0.32\}$, and embodied cost multipliers $M_{\text{phys}} \in [1.0\times, 6.0\times]$ (Table 4, Figures 2–3).

5. **Trade-offs, limitations, and falsification protocol (Sections 6–8).** We analyze the Pareto trade-off between market wage surplus and hazardous toil elimination, meta-analytic cases where AI alone beats human-AI teams [21], and why the $M_{\text{phys}} \approx 4.0\times$ robotics crossover is conditional on this budget model, concluding with six falsification tests (Table 5).

## 2 Task-based economics and human-agency foundations

Economic and systems analyses of automation increasingly model occupations as bundles of heterogeneous tasks rather than monolithic jobs [1, 3, 19]. A single healthcare, engineering, or logistics workflow interleaves high-disutility physical or clerical drudgery with high-agency diagnostic and creative judgment. Table 1 maps five core principles of disutility-weighted automation onto formal economic and human-factors constructs.

Table 1: From worker-centric automation principles to formal econometric and human-agent workflow constructs.

| Principle | Canonical construct | Protocol component | Operational function | Primary grounding |
|---|---|---|---|---|
| Target hazardous drudgery first | Task-based R&D direction [1] | Disutility bounty scorer ($\mathcal{D}_j$) | Ranks full-replacement R&D targets by physical/ergonomic hazard $h_j$ and noisy low-agency consensus $1 - \hat{v}_j$ | [10, 18, 19] |
| Preserve high-agency craft | Inalienable human-in-the-loop | Craft augmentation barrier ($\theta_{\text{craft}}$) | Restricts unattended replacement on tasks with $\hat{v}_j > \theta_{\text{craft}}$, routing R&D to steerable co-pilots | [5, 19, 21] |
| Prevent supervisory ironies | Shared-control responsibility allocation | Handover responsibility gate | Bounds unsteered handover state-gap errors when authority shifts between agent and human | [4, 6, 9] |
| Attribute joint contributions | Counterfactual marginal credit | Shapley contribution ledger | Quantifies human steering and domain judgment value across collaborative trajectories | [2, 14] |
| Compensate displaced toil workers | Redistributive transition policy | Transition dividend pool | Recycles a fraction of automated productivity gains into retraining and wage support | [1, 22, 23] |

Two empirical insights anchor Table 1. First, in a landmark systematic review and meta-analysis of 106 experimental studies (370 effect sizes), Vaccaro et al. [21] demonstrated that human-AI combinations do *not* universally outperform the better of the human or AI alone: on pure decision-making and prediction tasks, human-AI teams frequently underperform AI alone (H+AI <

max(H, AI)), whereas on open-ended creation and synthesis tasks, human-AI synergy is significantly positive. Second, Shao et al. [19] showed in WORKBank that workers partition tasks into an "Automation Red Light" zone (high intrinsic agency $v_j > \theta_{\text{craft}}$, where workers reject replacement) and an "Automation Green Light" zone (high monotony/hazard $h_j$, low agency $v_j$, where workers actively seek full delegation), though individual workers exhibit meaningful disagreement ($\sigma_{\text{pref}}$) around task means. Figure 1 illustrates how `DWAP-Bounty` operationalizes this partition.

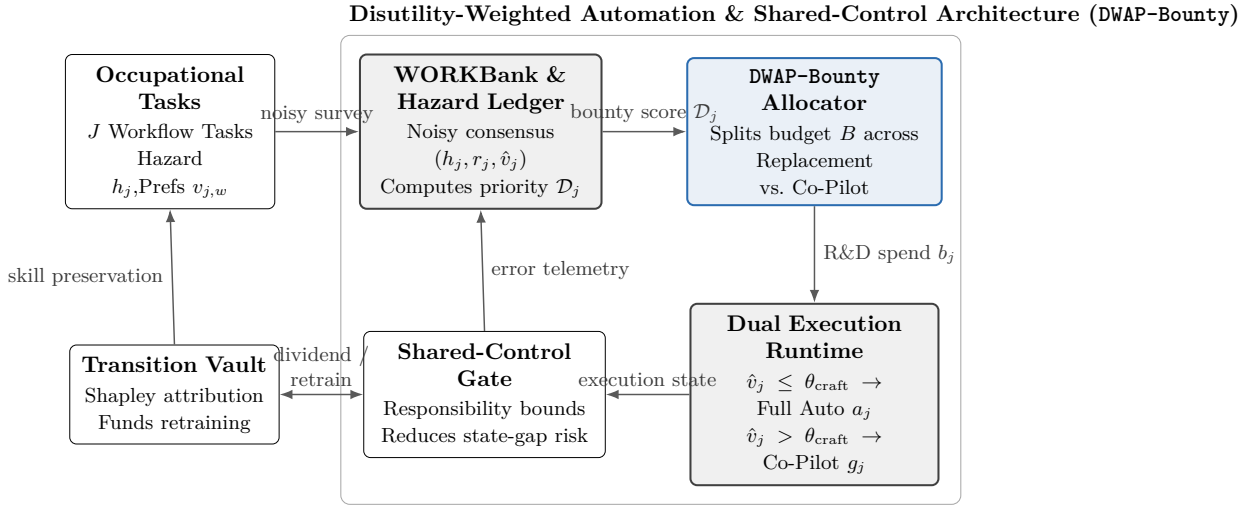



Figure 1: Architecture of Disutility-Weighted Automation Prioritization (`DWAP-Bounty`). Tasks are drawn from an illustrative 24-descriptor catalog characterized by occupational hazard $h_j$, routine index $r_j$ [3], and heterogeneous worker agency preferences $v_{j,w}$ aggregated into noisy planner estimates $\hat{v}_j$ [19]. The `DWAP-Bounty` allocator directs full-replacement R&D ($a_j$) toward high-disutility toil ($\hat{v}_j \leq \theta_{\text{craft}}$) while routing high-agency craft ($\hat{v}_j > \theta_{\text{craft}}$) to human-steered co-pilots ($g_j$) under shared-control responsibility bounds [5, 6].

## 3 Formal mathematical framework

Consider an occupational task portfolio comprising $J$ tasks indexed by $j \in \mathcal{J} = \{1, \ldots, J\}$, sampled from an illustrative 24-descriptor catalog anchored to O*NET, BLS, and WORKBank profiles [1, 3, 19] with physical-toil prevalence $\pi_{\text{toil}} \in (0, 1)$. Each task $j$ is characterized by:

1. **Occupational hazard and ergonomic strain** $h_j \in [0, 1]$ (e.g., confined-space tank cleaning 0.94, biohazard waste sorting 0.91, asbestos abatement 0.95);
2. **Routine task intensity** $r_j \in [0, 1]$ [3] (e.g., invoice OCR reconciliation 0.92, payroll tax ledger entry 0.90);
3. **Heterogeneous worker agency preferences and noisy consensus estimate:** around latent task agency $v_j \in [0, 1]$, individual workers $w \in \{1, \ldots, W\}$ hold preferences $v_{j,w} = \text{clip}(v_j + \epsilon_{j,w}, 0, 1)$ with disagreement $\epsilon_{j,w} \sim \mathcal{N}(0, \sigma^2_{\text{pref}})$, while planners observe only a noisy survey estimate $\hat{v}_j = \text{clip}(\frac{1}{W} \sum_{w=1}^{W} v_{j,w} + \eta_j, 0, 1)$ [19];
4. **Market wage weight** $w_j \in (0, 1]$ and **digital observability/feasibility** $f_j \in (0, 1]$;
5. **Unit R&D capital cost** $c_j = M_j(c_0 - c_1 f_j) > 0$, where $M_j = M_{\text{phys}} \geq 1$ for physical/embodied hazard-toil tasks [10, 16] and $M_j = 1$ for digital tasks.

### 3.1 Market wage-bill, feasibility, routine-biased, and constrained surplus baselines

Given a cumulative R&D budget $B > 0$, all policies share the *exact same* R&D conversion function $\eta(f_j) = 0.70{+}0.30 f_j$. Alongside the three classical greedy rules ($\pi_j^{\text{mkt}}$, $\pi_j^{\text{feas}}$, and routine-biased $\pi_j^{\text{rout}}$ [3]), we formulate a **Constrained Wage-Surplus Optimizer** ($\pi_j^{\text{csurp}}$) that maximizes expected net wage-bill surplus per unit R&D cost accounting for observed agency rework penalties ($1{-}0.45\hat{v}_j$) and co-pilot routing above $\theta_{\text{craft}}$:

$$\pi_j^{\text{mkt}} = \frac{w_j\,\eta(f_j)}{c_j}, \quad \pi_j^{\text{feas}} = \frac{f_j}{c_j}, \quad \pi_j^{\text{rout}} = \frac{r_j}{c_j}, \quad \pi_j^{\text{csurp}} = \frac{w_j\,\eta(f_j)\max\!\big(1-0.45\hat{v}_j,\ 0.32/\gamma_{\text{aug}}\big)}{c_j}. \tag{1}$$

### 3.2 Disutility-weighted bounty scoring and multi-objective Pareto frontier

`DWAP-Bounty` ranks tasks by a composite Disutility Priority Index $\mathcal{D}_j(\hat{v}_j)$ computed from observed survey consensus $\hat{v}_j$, and traces the full multi-objective Pareto frontier $\Pi(\lambda)$ for $\lambda \in [0,1]$ trading off disutility elimination against net market wage surplus under identical modality conversion:

$$\mathcal{D}_j(\hat{v}_j) = \frac{\alpha_h h_j + \alpha_v(1-\hat{v}_j) - \lambda_{\text{craft}}\max(0,\hat{v}_j - \theta_{\text{craft}})}{c_j}, \qquad \Pi_j(\lambda) = \lambda\,\mathcal{D}_j(\hat{v}_j){+}(1{-}\lambda)\,\frac{w_j\eta(f_j)(1-0.45\hat{v}_j)}{c_j}, \tag{2}$$

where `DWAP-Bounty` corresponds to $\lambda = 0.85$ with craft protection threshold $\theta_{\text{craft}} = 0.62$. Crucially, all policies and Pareto frontier points share the *identical* R&D conversion law $\eta(f_j) = 0.70 + 0.30 f_j$:

$$(a_j, g_j) \;=\; \begin{cases} \Big(\min\!\big(1,\ \frac{b_j}{c_j}\eta(f_j)\big),\ 0\Big), & \text{if } \hat{v}_j \le \theta_{\text{craft}} \text{ (Full Toil Replacement)},\\ \Big(0,\ \min\!\big(1,\ \frac{b_j}{\gamma_{\text{aug}} c_j}\eta(f_j)\big)\Big), & \text{if } \hat{v}_j > \theta_{\text{craft}} \text{ (Steered Co-Pilot Augmentation)}, \end{cases} \tag{3}$$

where $a_j \in [0,1]$ is unattended replacement coverage, $g_j \in [0,1]$ is human-steered co-pilot coverage, and $\gamma_{\text{aug}} = 0.35$.

### 3.3 Non-circular multi-objective evaluation under worker disagreement

We evaluate every policy on four independent, policy-neutral outcome metrics across the heterogeneous worker population $\{v_{j,w}\}_{w=1}^{W}$ and latent task parameters $(h_j, v_j, w_j, f_j)$:

$$\mathcal{T}_{\text{elim}} = \frac{\sum_{j=1}^{J} h_j a_j}{\sum_{j=1}^{J} h_j}, \qquad \mathcal{A}_{\text{WB}} = \frac{1}{JW}\sum_{j=1}^{J}\sum_{w=1}^{W}\Big(\mathbb{I}[v_{j,w} \le 0.55]\,a_j + \mathbb{I}[v_{j,w} > 0.55](1-0.88a_j)(0.8$$

$$\mathcal{S}_{\text{mkt}} = \frac{\sum_{j=1}^{J} w_j(a_j + 0.32 g_j - 0.45 a_j v_j)}{\sum_{j=1}^{J} w_j}, \qquad \mathcal{E}_{\text{gap}} = \frac{1}{J}\sum_{j=1}^{J}\Big(a_j(0.28v_j + 0.12(1-f_j)) + 0.04 g_j(1-f_j)\Big). \tag{4}$$

## 4 System architecture and parameter provenance

To deploy Equations (1)–(4) in cooperative or enterprise R&D portfolios, `DWAP-Bounty` combines three operational stages (Table 2, Table 3):

1. **Stage 1: Participatory WORKBank Telemetry and Response-Time Preference Aggregation.** Stage 1 elicits worker ratings of task hazard $h_j$ and desired agency $v_{j,w}$ using the WORKBank protocol [19], resolving heterogeneous worker preferences ($\sigma_{\text{pref}}$) via response-time-weighted social choice [7] to form $\hat{v}_j$.

2. **Stage 2: Inalienable Human-in-the-Loop (I-HITL) and Shared-Control Gates.** For tasks above the craft barrier ($\hat{v}_j > \theta_{\text{craft}}$), Stage 2 enforces I-HITL governance [5] and shared-control responsibility allocation (`LiRA` [6]), ensuring that co-pilot tools assist without stripping human sign-off authority or inducing cognitive divergence [9].

3. **Stage 3: Counterfactual Shapley Attribution and Transition Dividends.** When physical or clerical toil tasks are automated ($a_j \to 1$), Stage 3 uses counterfactual Shapley credit assignment [14] and AI-Economist redistributive tax rules [22, 23] to fund paid worker retraining into high-agency roles [1].

Table 2: Readiness and deployment bottlenecks of `DWAP-Bounty` subsystems.

| Subsystem | Maturity | Supporting empirical & theoretical evidence | Realistic deployment bottleneck |
|---|---|---|---|
| WORKBank agency & hazard auditing | Field validated | 844 tasks audited across U.S. workers in WORKBank [19]; response-time alignment [7] | Worker preferences $v_{j,w}$ exhibit within-task disagreement ($\sigma_{\text{pref}}$), requiring local calibration |
| Task-conditioned co-pilot routing | Meta-analysis grounded | 106-study meta-analysis of human-AI synergy [21]; long-task horizon bounds [12] | Uncalibrated confidence causes humans to over-rely on flawed co-pilot suggestions [2, 8] |
| Shared-control & I-HITL gates | Prototype ready | Inalienable HITL [5]; learned responsibility in shared control [6]; cobot safety [10] | High-frequency handovers can still induce supervisory fatigue [4, 9] |
| Embodied robotics toil replacement | Hardware constrained | Open X-Embodiment RT-X datasets [16]; task displacement vs. reinstatement [1] | Conditional on our budget model, when physical R&D cost multiplier $M_{\text{phys}} \geq 4.0\times$, single-firm budgets cannot afford robotics R&D |

Table 3: Provenance of simulation parameters in `experiment.py`. All five policies and Pareto frontier points share identical R&D conversion efficiency $\eta(f_j)$ and handover risk functions.

| Parameter | Value / Range | Provenance Type | Empirical or Theoretical Grounding |
|---|---|---|---|
| Illustrative 24-task catalog ($J$=60) | 8 hazard, 8 clerical, 8 craft descriptors | Illustrative O*NET / WORKBank catalog | Named task profiles (`A1_ConfinedTankClean`..`C8_SurgicalDecisionLead`) with prevalence $\pi_{\text{toil}} \in \{0.15, 0.33, 0.50\}$ [3, 19] |
| Preference disagreement $\sigma_{\text{pref}}$ | 0.18 (swept $\{0.05, 0.18, 0.32\}$) | Worker heterogeneity sweep | Models inter-worker disagreement $v_{j,w}$ around task agency $v_j$ and noisy survey consensus $\hat{v}_j$ ($W$=12) [19] |
| R&D budget $B$ & Pareto $\lambda$ | $B \in \{15\%, 30\%, 45\%, 60\%\}$, $\lambda \in [0, 1]$ | Parameterized budget & frontier sweep | Spans early adoption (15%) to mature deployment (60%) and traces the multi-objective Pareto frontier |
| R&D conversion $\eta(f_j)$ | $0.70 + 0.30 f_j$ (equal across all policies) | Equalized environment law | Removes policy-specific R&D efficiency bonuses across all baselines |
| Physical cost multiplier $M_{\text{phys}}$ | 1.5× (swept 1.0 × –6.0×) | Conditional cost-ratio sweep | Explores the conditional robotics cost crossover within this budget model [16] |

# 5 Empirical evaluation, Pareto frontier, and sensitivity analysis

We evaluate `DWAP-Bounty` in a paired Monte Carlo benchmark (`experiment.py`): $N = 500$ paired 60-task portfolio runs across five policies and four budgets $B \in \{15\%, 30\%, 45\%, 60\%\}$ at $M_{\text{phys}} = 1.5\times, \pi_{\text{toil}} = 0.33, \sigma_{\text{pref}} = 0.18$, plus $N = 300$ runs per weight $\lambda \in \{0.0, 0.25, 0.50, 0.75, 1.0\}$ along the multi-objective Pareto frontier, and $N = 250$ paired runs across each level of $\pi_{\text{toil}} \in \{0.15, 0.33, 0.50\}$, $\sigma_{\text{pref}} \in \{0.05, 0.18, 0.32\}$, and $M_{\text{phys}} \in \{1.0\times, 1.5\times, 2.5\times, 4.0\times, 6.0\times\}$.

## 5.1 Evaluated policies under paired portfolios and noisy agency estimates

In every trial, all five policies observe the exact same paired task portfolio and noisy survey estimates $\hat{v}_j$: **Market Wage Greedy** ($\pi_j^{\text{mkt}}$), **Feasibility-First** ($\pi_j^{\text{feas}}$), **Routine-Biased Replacement** ($\pi_j^{\text{rout}}$ [3]), **Constrained Surplus Optimizer** ($\pi_j^{\text{csurp}}$, which maximizes net market wage surplus while routing $\hat{v}_j > 0.62$ to co-pilots), and **`DWAP-Bounty`** (Equations 2–3).

Table 4: Equalized-efficiency paired comparison across simulated 60-task R&D portfolios ($N = 500$ trials per policy at $B = 45\%, M_{\text{phys}} = 1.5\times, \pi_{\text{toil}} = 0.33, \sigma_{\text{pref}} = 0.18$, plus Toil Elimination scaling at $B = 15\%$ and 60%). Values report means and 95% confidence intervals `[lo, hi]`.

| R&D Policy | Toil Elim 15% | Toil Elim 45% | Toil Elim 60% | Worker Align (%) | Market Surplus (%) | Gap Err (%) |
|---|---|---|---|---|---|---|
| Market Wage Greedy | 7.14 [7.07, 7.22] | 44.80 [44.76, 44.84] | 61.91 [61.87, 61.95] | 39.11 [39.06, 39.16] | **53.41 [53.38, 53.44]** | 12.10 |
| Feasibility-First | 22.63 [22.55, 22.71] | 44.78 [44.74, 44.83] | 62.32 [62.28, 62.36] | 39.23 [39.18, 39.28] | 53.10 [53.06, 53.14] | 12.05 |
| Routine-Biased (Autor) | 23.57 [23.53, 23.61] | 58.87 [58.80, 58.95] | 74.19 [74.10, 74.28] | 73.06 [72.92, 73.19] | 37.23 [37.15, 37.30] | 4.77 |
| Constrained Surplus | 11.65 [11.61, 11.69] | 51.23 [51.19, 51.28] | 68.15 [68.10, 68.20] | 73.73 [73.68, 73.79] | 48.29 [48.26, 48.31] | 4.51 |
| **`DWAP-Bounty`** | **24.04 [24.01, 24.08]** | **60.49 [60.45, 60.53]** | **76.89 [76.84, 76.93]** | **74.29 [74.23, 74.35]** | 37.26 [37.22, 37.30] | **4.50** |

### 5.2 Quantitative findings, Pareto frontier, and multi-axis sensitivity

Table 4 and Figures 2–3 establish four empirical findings:

1. **Multi-objective Pareto frontier between market wage surplus and hazardous toil elimination.** Tracing the Pareto frontier $\Pi(\lambda)$ at $B = 45\%$ across $\lambda \in \{0.0, 0.25, 0.50, 0.75, 1.0\}$ under identical R&D conversion efficiency and co-pilot handover rules reveals the exact structural trade-off. At $\lambda = 0.0$ (pure constrained wage-surplus optimization) and Constrained Surplus Optimizer (`constrained_surplus_opt`), the planner achieves high market wage surplus (48.81% and 48.29% [48.26, 48.31]) and low handover error (4.51%), but eliminates only 50.74%–51.23% [51.19, 51.28] of hazardous toil (and only 11.65% [11.61, 11.69] at $B = 15\%$) because low-wage hazardous toil is deferred behind higher-wage routine clerical work. Increasing $\lambda \to 0.50, 0.75$, and 0.85 (`DWAP-Bounty`) shifts R&D along the Pareto frontier: hazardous toil elimination rises to 59.78%, 60.52%, and 60.49% [60.45, 60.53] (and 24.04% [24.01, 24.08] at $B = 15\%$) with 74.29% [74.23, 74.35] worker-agency alignment and 4.50% [4.49, 4.51] handover error, trading off market wage surplus (37.26% [37.22, 37.30]). Unconstrained Market Wage Greedy achieves 53.41% [53.38, 53.44] wage surplus but collapses to 44.80% [44.76, 44.84] toil elimination, 39.11% [39.06, 39.16] worker-agency alignment, 12.10% [12.10, 12.11] handover error, and 0.00% [0.00, 0.76] dual-target success (0/500 vs. 100.00% [99.24, 100.00] for `DWAP-Bounty`).

2. **Sensitivity to physical-toil archetype prevalence ($\pi_{\text{toil}} \in \{0.15, 0.33, 0.50\}$).** Varying the catalog share of physical hazard toil at $B = 45\%$ confirms that `DWAP-Bounty` consistently leads in toil elimination across all prevalence regimes: at low physical-toil prevalence ($\pi_{\text{toil}} = 0.15$), `DWAP-Bounty` eliminates 76.16% [76.09, 76.24] of hazardous toil (86.42% [86.33, 86.50] alignment) vs. 63.67% [63.58, 63.75] for Constrained Surplus and 48.90% [48.81, 48.99] for Market Wage Greedy; at high physical-toil prevalence ($\pi_{\text{toil}} = 0.50$, where half the catalog requires 1.5× costlier embodied R&D), `DWAP-Bounty` eliminates 50.06% [50.01, 50.12] of hazardous toil (57.60% [51.40, 63.57] dual-target success, 144/250) vs. 48.46% [48.38, 48.54] for Routine-Biased, 43.91% [43.86, 43.97] for Constrained Surplus, and 39.81% [39.76, 39.86] for Market Wage Greedy (0.00% [0.00, 1.51] dual-target success).

3. **Sensitivity to inter-worker preference disagreement ($\sigma_{\text{pref}} \in \{0.05, 0.18, 0.32\}$).** When individual workers strongly agree on task agency ($\sigma_{\text{pref}} = 0.05$), `DWAP-Bounty` achieves 77.03% [77.00, 77.06] worker-agency alignment. As inter-worker preference disagreement widens to $\sigma_{\text{pref}} = 0.18$ and $\sigma_{\text{pref}} = 0.32$—so any single task-level modality decision inevitably dissatisfies a minority of workers—realized worker-agency alignment degrades gracefully to 74.34% [74.26, 74.42] and 68.26% [68.13, 68.39], while hazardous toil elimination remains invariant at 60.55% [60.48, 60.61] and 60.41% [60.34, 60.47].

4. **Model-conditioned embodied robotics cost crossover ($M_{\text{phys}} \in [1.0\times, 6.0\times]$).** Sweeping the physical-to-digital R&D cost multiplier $M_{\text{phys}}$ at $B = 45\%$ maps the budget boundary within our stylized linear-saturation cost model. At $M_{\text{phys}} = 1.0\times$ and $1.5\times$, `DWAP-Bounty` eliminates 74.90% [74.83, 74.97] and 60.50% [60.43, 60.57] of hazardous toil (100.00% [98.49, 100.00] dual-target success). At $M_{\text{phys}} = 2.5\times$, toil elimination falls to 48.89% [48.83, 48.95] (still exceeding Routine-Biased at 45.53% and Market Wage Greedy at 42.86%). Conditional on this budget and cost-normalization specification ($B = 45\%$, division by $c_j$ in Eq. 2), a crossover occurs around $M_{\text{phys}} \approx 4.0\times$: at $4.0\times$, `DWAP-Bounty` (41.93% [41.87, 41.99]) converges to Market Wage Greedy (41.77% [41.72, 41.82]), and at $M_{\text{phys}} = 6.0\times$, `DWAP-Bounty` eliminates less total toil (38.05% [37.99, 38.11]) than Routine-Biased (40.95% [40.89, 41.01]) while main-

taining higher worker-agency alignment (62.38% [62.30, 62.46] vs. 37.73% [37.61, 37.86]). We emphasize that this 4.0× crossover is a conditional property of the $B = 45\%$ cost-normalized objective rather than a universal industrial constant.

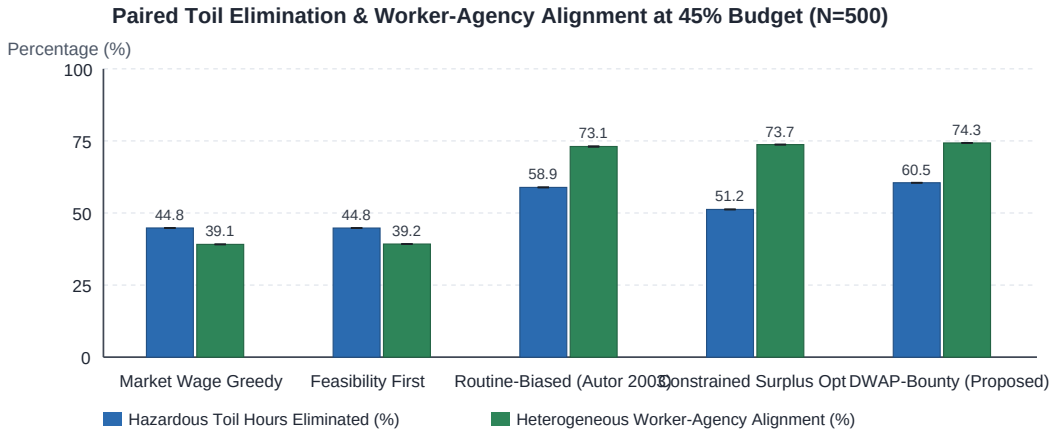



Figure 2: Equalized-efficiency paired comparison of Hazardous Toil Hours Eliminated and Heterogeneous Worker-Agency Alignment at $B = 45\%, M_{\text{phys}} = 1.5\times, \sigma_{\text{pref}} = 0.18$ ($N = 500$ portfolios per policy; error bars denote 95% confidence intervals).

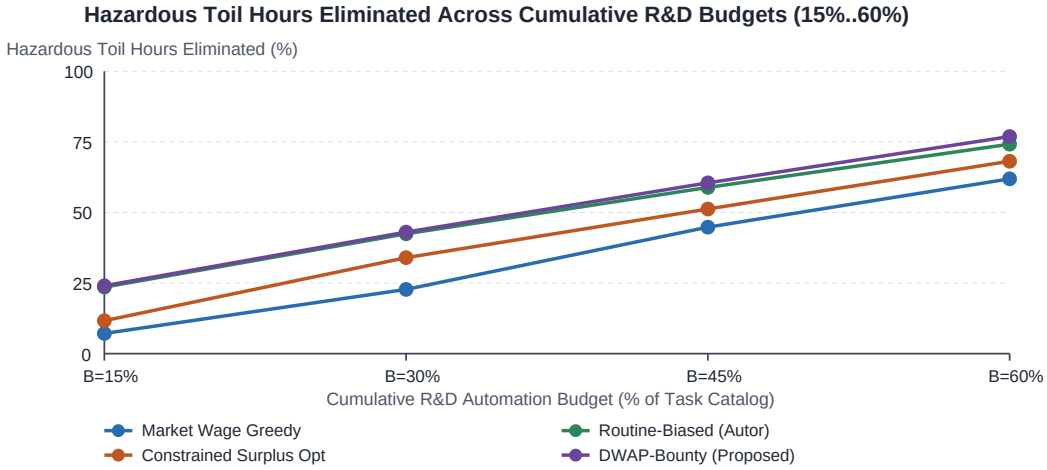



Figure 3: Scaling of Hazardous Toil Hours Eliminated across cumulative R&D budgets $B \in \{15\%, 30\%, 45\%, 60\%\}$ under equalized R&D conversion efficiency.

## 6 Economic trade-offs and operational boundary conditions

We highlight four empirical and theoretical counter-arguments that bound disutility-weighted automation:

1. **When human-in-the-loop augmentation underperforms solo AI (Arghal et al. [2], Vaccaro et al. [21]).** In their 106-study meta-analysis, Vaccaro et al. [21] prove that on

structured decision and prediction tasks, human-AI combinations perform worse on average than the AI alone because humans introduce miscalibrated overrides [2, 8]. Where safety or diagnostic accuracy strictly outweighs practitioner craft enjoyment $v_j$, governance must prioritize error minimization over agency preservation.

2. **Pareto trade-off against constrained and unconstrained wage-bill savings (Acemoglu and Restrepo [1]).** As shown in Table 4 and our Pareto frontier analysis, Market Wage Greedy and Constrained Surplus Optimization generate +16.15 and +11.03 percentage points higher short-term market wage surplus than `DWAP-Bounty` at $B = 45\%$ (53.41% and 48.29% vs. 37.26%). In an unregulated market without fiscal incentives or Pigouvian taxes on uncompensated displacement [23], firms that prioritize high-wage tasks can under-price cooperatives that prioritize low-wage hazardous toil.
3. **Embodied robotics capital barriers (Kwa et al. [12], Open X-Embodiment Collaboration [16]).** As demonstrated in our conditional $M_{\text{phys}} \in [1.0\times, 6.0\times]$ sweep, when physical manipulation R&D costs several multiples of digital software engineering, cost-normalized prioritization defers physical toil unless cooperatives pool R&D budgets.
4. **Intra-occupational worker preference disagreement (Echenique et al. [7], Shao et al. [19]).** As shown in our $\sigma_{\text{pref}} = 0.32$ sweep, when workers within the same role disagree on whether a task is drudgery or valued craft, task-level R&D allocation loses up to 8.77 percentage points of alignment unless paired with worker-level role rotation.

# 7 Limitations

- **Illustrative 24-descriptor task catalog.** The 24-descriptor catalog and 60-task portfolio simulation in `experiment.py` are illustrative parametric models grounded in qualitative O*NET and WORKBank task profiles [3, 19]; they do not constitute an econometric general-equilibrium calibration of sector-wide wage dynamics [1].
- **Model-conditioned cost crossover.** The $M_{\text{phys}} \approx 4.0\times$ embodied-cost crossover depends on the $B = 45\%$ budget level, linear conversion $\eta(f_j)$, and $1/c_j$ normalization in Equation (2), and should not be interpreted as a general industrial constant.
- **Static task boundaries.** Automating one node in a workflow alters the cognitive load and agency value of adjacent tasks [11, 20].

# 8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical tests designed to falsify the core claims of `DWAP-Bounty`.

Table 5: Six pre-registerable empirical tests that could falsify the `DWAP-Bounty` framework.

| Protocol | Hypothesis | Design and primary metrics | Falsified if |
|---|---|---|---|
| P1. WORKBank capital steering | `DWAP-Bounty` directs $> 70\%$ of full-replacement R&D to `HAS-1/2` toil tasks [19] | Audit R&D allocation across a 200-task enterprise backlog rated by 150 workers on hazard $h_j$ and agency $v_{j,w}$ | $> 25\%$ of full-replacement R&D budget is allocated to tasks with median worker rating `HAS-4` or `HAS-5` |
| P2. Shared-control error reduction | I-HITL and shared-control gates [5, 6] reduce handover errors vs. unattended replacement | Evaluate 200 multi-step workflows comparing unattended replacement vs. `DWAP-Bounty` co-pilots on high-$v_j$ tasks | Handover state-gap error rate under `DWAP-Bounty` fails to drop by $\geq 40\%$ relative to Market Wage Greedy |
| P3. Reverse-centaur workload audit | Protecting $\hat{v}_j > \theta_{\text{craft}}$ prevents human demotion into passive cleanup loops [4] | 12-week field pilot in 8 engineering cooperatives tracking time spent on unsteered exception cleanup | Fraction of human work hours spent on post-hoc agent cleanup increases relative to pre-deployment baseline |
| P4. Human-AI co-pilot synergy | Task-conditioned co-pilots on Group-C creation tasks beat solo AI [21] | Benchmark 300 Group-C tasks across Solo Human, Solo Agent, and `DWAP-Bounty` Co-Pilot | Co-pilot modality underperforms Solo Agent by $> 5\%$ on domain quality rubrics or degrades skill retention [20] |
| P5. Pooled-bounty robotics scaling | Multi-cooperative R&D pooling restores $> 55\%$ toil elimination under high physical R&D cost ($M_{\text{phys}} = 4.0\times$) | Compare single-cooperative ($B = 15\%$) vs. 3-cooperative pooled ($B = 45\%$) `DWAP-Bounty` under $M_{\text{phys}} = 4.0\times$ | Pooled R&D fails to improve physical toil elimination by $\geq 25$ percentage points over single-firm allocation |
| P6. Preference disagreement robustness | Response-time preference aggregation [7] preserves $\geq 65\%$ worker alignment under $\sigma_{\text{pref}} \leq 0.30$ | Measure worker-agency alignment across 400 workers in 4 occupational sectors with high preference variance | Realized worker-agency alignment falls below 65% or rank correlation $\rho$ across sub-cohorts falls below 0.70 |

# 9 Conclusion

When R&D capital is allocated solely by nominal wage-bill savings or digital feasibility, automation disproportionately displaces high-agency human craft while neglecting high-hazard toil. Evaluating five policies and the multi-objective Pareto frontier across an illustrative 24-descriptor task catalog under strictly equalized R&D efficiency, handover risk, and inter-worker preference disagreement ($\sigma_{\text{pref}} = 0.18$), we show that `DWAP-Bounty` eliminates $60.49\%$ $[60.45, 60.53]$ of hazardous toil hours and achieves $74.29\%$ $[74.23, 74.35]$ worker-agency alignment at $B = 45\%$ (compared to $51.23\%$ toil elimination and $48.29\%$ market surplus under Constrained Surplus Optimization, and $44.80\%$ toil elimination and $53.41\%$ market surplus under Market Wage Greedy), while mapping sensitivity across archetype prevalence $\pi_{\text{toil}}$, worker disagreement $\sigma_{\text{pref}}$, and conditional embodied robotics cost ratios $M_{\text{phys}}$.

## Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

## References


[1] Daron Acemoglu and Pascual Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. The American Economic Review, 108(6):1488–1542, 2018. URL https://doi.org/10.1257/aer.20160696.

[2] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL https://arxiv.org/abs/2610.09371.

[3] David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. The Quarterly Journal of Economics, 118(4):1279–1333, 2003. URL https://doi.org/10.1162/003355303322552801.

[4] Lisanne Bainbridge. Ironies of automation. Automatica, 19(6):775–779, 1983. URL https://doi.org/10.1016/0005-1098(83)90046-8.

[5] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL https://arxiv.org/abs/2610.35835.

[6] Xiaoyang Cao, Jingqi Li, Zhe Fu, and Alexandre M. Bayen. Who bears the burden? learning responsibility for shared constraints in multi-agent reinforcement learning (LiRA). arXiv preprint arXiv:2610.07491, 2026. URL https://arxiv.org/abs/2610.07491.

[7] Federico Echenique, Alireza Fallah, Baihe Huang, and Michael I. Jordan. Response time enhances alignment with heterogeneous preferences, 2026. URL https://arxiv.org/abs/2605.06987.

[8] Elena Eleftheriou, George Pallis, and Marios Constantinides. Confidence without competence in AI-assisted knowledge work. In *Proceedings of CHIWORK '26*, 2026. URL https://arxiv.org/abs/2604.09444.

[9] Netanel Eliav. The cognitive divergence: AI context windows, human attention decline, and the delegation feedback loop. arXiv preprint arXiv:2603.26707, 2026. URL https://arxiv.org/abs/2603.26707.

[10] International Organization for Standardization. ISO/TS 15066:2016 robots and robotic devices — collaborative robots. ISO Technical Specification, 2016. URL https://www.iso.org/standard/62996.html.

[11] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task, 2025. URL https://arxiv.org/abs/2506.08872.

[12] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long tasks. arXiv preprint arXiv:2503.14499, 2025. URL https://arxiv.org/abs/2503.14499.

[13] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3713778.

[14] Mingxuan Li, Kai-Zhan Lee, and Elias Bareinboim. Counterfactual Shapley credit assignment. Reinforcement Learning Journal (RLJ) / arXiv:2607.16999, 2026. URL https://arxiv.org/abs/2607.16999.

[15] J. C. R. Licklider. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1:4–11, 1960.

[16] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. IEEE International Conference on Robotics and Automation (ICRA), 2024. URL https://arxiv.org/abs/2310.08864.

[17] Sina Rismanchian, Hasan Uzun, Jeffrey Matayoshi, Eric Cosyn, and Eyad Kurd-Misto. Faster completion, less learning: Generative AI reduced study time on math problems and the knowledge they build, 2026. URL https://arxiv.org/abs/2605.21629.

[18] Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000.

[19] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with AI agents: Auditing automation and augmentation potential across the U.S. workforce (WORKBank). arXiv preprint arXiv:2506.06576, 2025. URL https://arxiv.org/abs/2506.06576.

[20] Judy Hanwen Shen and Alex Tamkin. How AI impacts skill formation, 2026. URL https://arxiv.org/abs/2601.20245.

[21] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8:2293–2303, 2024. URL https://arxiv.org/abs/2405.06087.

[22] Hitoshi Yamada. Universal basic income with time-decaying currency: Structural effects on essential labor and long-term formation. arXiv preprint arXiv:2602.18714, 2026. URL https://arxiv.org/abs/2602.18714.

[23] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The AI economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science Advances, 8(18):eabk2607, 2022. URL `https://doi.org/10.1126/sciadv.abk2607`.

[24] Jiayin Zhi, Harsh Kumar, and Mina Lee. Investigating the effects of LLM use on critical thinking under time constraints: Access timing and time availability. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2026. doi: 10.1145/3772318.3791796.