The 70 Percent Ratio Theorem

A conceptual study of human expertise and AI assisted productivity

Concept proposed by Kalu Wu; AI-assisted conceptual analysis

25 October 2026 | Conceptual paper and proposed research protocol

Abstract

Kalu Wu proposes that a 70/30 ratio can guide productive living: approximately 30 percent of a day supports sleep, while professionals develop a strong specialist foundation, use artificial intelligence to assist with costly additional work, and redirect learning effort toward broader knowledge. This paper examines the proposal as a testable allocation framework rather than an established mathematical theorem. It distinguishes daily time allocation, proficiency, task coverage, and learning investment, which cannot be exchanged simply because they share a percentage. A selective review of primary research finds that generative AI can improve performance on particular tasks, but benefits vary across workers and tasks, and assistance can also reduce correctness or interfere with independent learning. A formal model therefore includes verification, rework, competency requirements, and limits on acceptable harm. Under restricted diminishing-return assumptions, an optimal allocation equates the marginal value of further specialization with the marginal value of breadth; the result does not identify 70 percent as universally optimal. Illustrative calculations show how different assumptions move the preferred allocation. A factorial study is proposed to compare depth-focused and foundation-plus-breadth training, with and without AI, while measuring complete-workflow performance and retained unaided ability. The defensible conclusion is conditional: preserve sufficient human expertise and recovery, delegate only where evidence supports reliable oversight, and reinvest verified savings in purposeful breadth. The exact ratio should remain an empirical question.

Keywords Human AI collaboration; expertise; generalist knowledge; diminishing returns; verification; time allocation; learning transfer

Status of this paper

This is a conceptual analysis, not a report of a completed experiment, a systematic review, or a peer-reviewed validation of a new theorem. The proposed name is retained to represent Kalu Wu's idea. Numerical examples are constructed illustrations, not observations. The analysis does not recommend reducing professional qualification requirements or treating sleep as time available for productivity cuts.

1 The question and the meaning of 70 percent

The motivating question is important: if AI changes the cost of producing specialist work, should people change how they invest in expertise? Kalu Wu's proposal connects this question to a wider philosophy of balance. Instead of pursuing every increment of narrow mastery, a professional might retain a capable core, obtain assistance for suitable tasks, and develop the knowledge needed to see problems from more than one discipline. The potential benefit is greater useful output per unit of effort, together with better judgment about the wider context of that output.

The strongest version of the proposal is not a fixed ceiling on human learning. It is a rule for examining opportunity costs. Further depth may be worthwhile, but it should be compared with the value of adjacent learning, recovery, and collaboration. Conversely, breadth should earn its place through improved understanding or performance rather than through the number of subjects encountered. This interpretation respects the original ambition while making it possible to test.

Four quantities that must remain separate

Quantity	A legitimate denominator	What 70 percent would mean
Daily time	A 24-hour day	16 hours 48 minutes awake
Proficiency	A defined assessment	70 percent of weighted score; not all knowledge
Task coverage	A specified workload	Tasks accounting for 70 percent of chosen weights
Learning allocation	Discretionary training time	Seven of ten learning hours in the specialty

A professional might solve 70 percent of a benchmark, cover tasks representing 70 percent of workload, or spend 70 percent of a training budget on the specialty. These are different statements. A benchmark score also depends on item difficulty and weighting; it is not a natural percentage of everything that could be known. Expertise has no stable, universally agreed denominator, and many fields expand faster than any individual can master them.

For this paper, the **70 percent mastery claim** is the strong conjecture that a specified intermediate level of unaided capability, combined with AI and broader learning, can match or outperform a more deeply trained alternative on a specified task portfolio. The **70/30 allocation heuristic** is the weaker, directly actionable suggestion to test a 70/30 split of a discretionary learning budget. Evidence for the second would not establish the first. Both require a comparison against deeply trained professionals who also have access to AI.

2 Sleep and the daily time budget

Thirty percent of 24 hours equals 7.2 hours, or 7 hours 12 minutes; the complementary 70 percent is 16 hours 48 minutes. This arithmetic makes the proposed ratio memorable. It does not establish a biological optimum. The American Academy of Sleep Medicine and Sleep Research Society recommend that adults aged 18 to 60 regularly obtain at least seven hours of sleep, while recognizing individual variation and the importance of quality, timing, and regularity. Their consensus does not prescribe exactly 7.2 hours. [1]

The sleep example is therefore a motivating analogy, not evidence for a learning ratio. A valid inference would require a mechanism linking the same numerical partition to both biological recovery and knowledge acquisition. No such mechanism is supplied here. Some people need more sleep than the example provides. Health conditions, age, schedules, and personal circumstances also make a universal daily partition inappropriate.

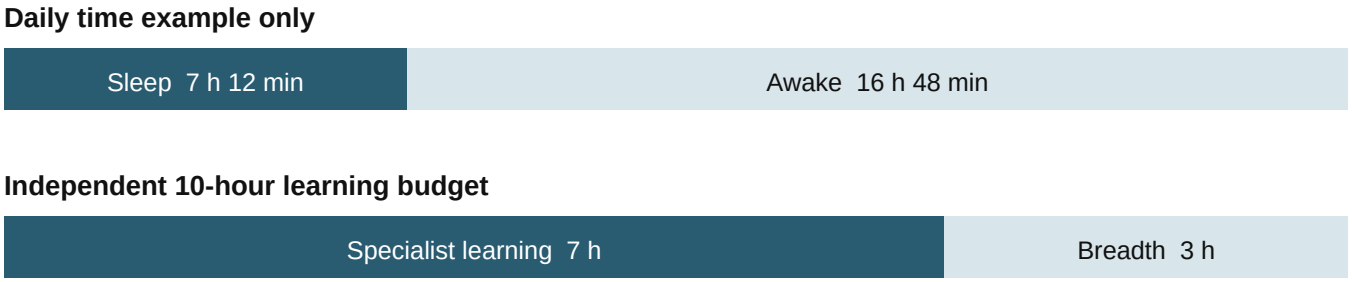


Figure 1. Separate denominators. The upper bar illustrates the arithmetic of a 24-hour day, not a sleep prescription. The lower bar shows a hypothetical 10-hour discretionary learning budget. The bars do not imply that waking hours are all working hours or that seven learning hours produce 70 percent proficiency.

Recovery is a constraint rather than a residual

A realistic daily budget includes sleep, paid work, domestic labor, caregiving, meals, movement, relationships, and unstructured time. Only part of the remainder can be allocated freely. Treating all waking time as productive labor ignores obligations and replaces a balanced-life proposal with continuous optimization. Time saved with AI may reasonably be used for recovery or relationships rather than additional output.

The practical order is to protect appropriate sleep and other non-negotiable commitments, identify an actual discretionary budget, and only then compare uses of that budget. Learning can be planned weekly when a daily ratio is unrealistic. A person with two free hours each week and a person with ten should not be judged by identical absolute learning targets. The ratio is a planning prompt, not a measure of personal worth or discipline.

3 What AI productivity evidence supports

Published studies support the possibility of substantial task-level gains, but they do not test a universal 70 percent human-competence threshold. Results are tied to particular tools, populations, workflows, and outcome measures. This distinction matters because faster drafting, higher rated quality, correct decisions, and durable expertise are not interchangeable outcomes.

Noy and Zhang randomly exposed half of 453 college-educated professionals to ChatGPT while they completed occupation-specific writing tasks. The assisted group took 40 percent less time on average and produced work rated 18 percent higher in quality. The experiment supports assistance on bounded writing tasks; it does not show that participants could safely omit 30 percent of their professional education. [2]

Brynjolfsson, Li, and Raymond studied a staggered deployment among 5,172 customer-support agents. In the final published paper, AI access increased issues resolved per hour by approximately 15 percent on average. Less-experienced and lower-skilled workers benefited more; the highest-skilled workers experienced small speed gains and small quality declines. This workplace evidence is valuable, but a nonrandom rollout at one firm is not a randomized comparison of alternative training depths. [3]

Dell'Acqua and colleagues randomized AI access among 758 consultants. Across 18 tasks within the tested model's capability frontier, assisted participants completed 12.2 percent more tasks and worked 25.1 percent more quickly. On a managerial task outside that frontier, correctness declined by about 19 percentage points when the AI groups were pooled against the control. The published 2026 study makes task selection central to any delegation policy. [4]

A contrasting randomized study by Becker and colleagues involved 16 experienced open-source developers completing 246 repository tasks. With early-2025 AI tools, completion took 19 percent longer, with a reported 95 percent confidence interval from 2 to 39 percent longer. This narrow result is not a verdict on all coding assistance. METR's February 2026 update reported that newer benefits were plausible but that selection and time-measurement problems made the follow-up estimates unreliable. [5, 6]

What can be inferred

The evidence supports testing AI-assisted workflows rather than assuming either benefit or harm. It does not support treating the most difficult 30 percent of a field as the part AI will necessarily perform best. Human difficulty and AI reliability can diverge. A task that is expensive for a person to learn may also be unusually hard to verify, dependent on local context, or outside the model's capabilities.

Nor can these effect sizes be pooled casually: they measure different outcomes under different conditions. A 40 percent reduction in task time is not a 40 percent increase in knowledge, and none of the studies establishes how saved time is subsequently invested in generalist learning. That reinvestment is a separate causal link that requires direct measurement.

4 Expertise and the learning problem

The proposal assumes that the marginal cost of further mastery often rises. This is plausible for some skills, but the shape and meaning of a learning curve require care. Ericsson, Krampe, and Tesch-Romer's influential work associated extended, effortful practice with expert performance. Its observations and retrospective practice estimates do not establish a universal number of hours, a single route to expertise, or a point at which practice becomes unnecessary. [7]

Learning-curve research also cautions against treating a convenient mathematical shape as a natural law. Heathcote, Brown, and Mewhort found that exponential functions fitted individual practice series better than power functions in the datasets they examined, and that averaging could favor apparent power laws. This concerns particular performance measures and practice settings, not a universal measure of mastery. [8]

The practical issue is which additional capability is being acquired. Improving the elegance of a routine report may have small marginal value. Learning to detect a rare but consequential error may have very large value. The last portion of a curriculum may also integrate earlier material, enabling a practitioner to recognize when a standard procedure does not apply. Such threshold effects can reverse the assumption that late-stage learning is merely expensive refinement.

Assisted performance is not retained competence

Bastani and colleagues conducted a randomized study of GPT-based assistance in high-school mathematics. Unrestricted assistance improved supported practice performance but harmed subsequent unaided exam performance; a tutor designed with learning guardrails differed. This educational result cannot be transferred directly to experienced professionals, but it demonstrates why the output produced with a tool and the ability retained without it must be measured separately. [9]

A professional therefore needs two scorecards: one for the quality of the human-AI system and another for retained human capabilities. The second should include problem framing, error detection, explanation, and appropriate escalation, not only the ability to reproduce a finished answer. AI can be used as a tutor, critic, simulator, or drafting assistant; those modes need not have the same learning effects.

Roediger and Karpicke's experiments found that retrieval testing produced better delayed retention than additional restudy under their experimental conditions. This supports testing an implementation that includes unaided recall and explanation, rather than assuming that reading polished AI output creates durable understanding. [10] It does not guarantee transfer to every workplace skill. Periodic independent work remains a useful diagnostic: can the person still reason through the central problem when the tool is unavailable or misleading?

5 Turning breadth into multidimensional understanding

Generalist knowledge is valuable only if it changes what a person can notice, explain, decide, or communicate. A specialist who learns basic statistics may better assess uncertain evidence. A software developer who studies organizational behavior may better understand why a technically sound system is not adopted. An art professional who studies materials, history, and consumer psychology may frame a question that a narrow technical workflow misses. These are hypothetical applications, not measured effects of the proposed ratio.

Breadth can be organized as a connected portfolio rather than a collection of unrelated subjects. One layer deepens understanding of the immediate work environment; another supplies transferable methods such as measurement or argument analysis; a third permits exploration whose usefulness is not yet known. A useful portfolio needs enough depth in each selected area to apply an idea correctly and recognize its limits.

Gick and Holyoak's experiments on analogical transfer showed that comparing multiple examples and abstracting their shared structure could help participants transfer a solution to a different problem. The research supports deliberate connection-making, not the claim that exposure to many topics automatically creates general intelligence. [11] In practice, a learner should explain both why an analogy is relevant and where it breaks down.

A proposed integration exercise

For each adjacent subject, choose an authentic problem from the core specialty. Describe the problem first using specialist concepts, then using the new subject's concepts. Identify one decision that changes, one useful additional observation, and one assumption that remains untested. Ask a knowledgeable reviewer to challenge the translation. Finally, attempt a new problem without AI assistance. This sequence creates evidence of integration rather than a record of content consumption.

For example, a designer learning statistics might distinguish preference surveys from experiments before claiming that a redesign caused higher sales. The value arises from choosing a better inference, not from replacing design expertise with statistical vocabulary. The same exercise can reveal that the adjacent subject is not useful enough to justify continued investment.

The proposal must also recognize costs: switching subjects can fragment attention, learning materials vary in quality, and the benefits of distant knowledge may arrive slowly. Broad exploration and uninterrupted specialist work can coexist when scheduled in blocks rather than interleaved constantly. A weekly 70/30 learning plan should be evaluated through completed reasoning tasks and retained understanding, not through equal enthusiasm for every new topic. Purposeful breadth complements a professional identity; it does not require abandoning depth.

6 A model of constrained human AI collaboration

Consider a professional allocating a fixed discretionary learning budget L. Let x be specialist-learning time and y = L - x be breadth-learning time. Let $q = D(x; q_0)$ denote a vector of specialist capabilities after training from baseline q_0, and $b = G(y; b_0)$ a vector of adjacent capabilities. The functions need not be linear. AI access a specifies the actual model, tools, data permissions, and workflow rather than a timeless level of intelligence.

For each task j, a policy chooses human execution, AI-assisted execution, or escalation to another expert. Define t_j as total human time, including instructions, review, correction, and coordination; v_j as the value of an acceptable result; and e_j as the probability of a materially erroneous final result after the complete workflow. All depend on capabilities and the chosen policy. Merely generating an answer is not completion.

One possible planning objective is to maximize expected net value over a fixed task portfolio and evaluation horizon:

$$U = \sum_j p_j [v_j (1 - e_j) - c\, t_j - h_j e_j] - C_{AI}$$

Here p_j is expected task frequency, c converts human time into the chosen value units, h_j represents additional consequences of error, and C_{AI} includes tool and implementation costs. This is an explicit decision model, not an empirically estimated equation. It requires agreed valuations; where consequences cannot responsibly be monetized, they belong in constraints rather than a financial trade-off.

Feasibility comes before optimization

Require each critical capability q_k to meet a task-specific threshold, each serious-error probability to remain below its approved limit, and each action to satisfy applicable authorization and professional requirements. Include a workable fallback when AI is unavailable. A high mean test score must not compensate for inability to recognize a dangerous exception. Some work remains infeasible for a particular person regardless of the average productivity gain.

Thresholds are local to the workflow. Independent tests, expert review, or narrower task scope may sometimes enable safe work with less individual knowledge. Other contexts demand extensive independent expertise. The model therefore does not assert that one capability floor is universally necessary or sufficient.

The unconstrained objective alone is inadequate: it could reward many minor successes despite a small chance of an unacceptable failure. Verification must also be substantive. Asking the same model repeatedly is not independent confirmation, and a second model may share errors. Reliable checks may require authoritative records, calculations, tests, physical measurement, or a qualified reviewer. [12, 13]

7 A conditional allocation proposition

The full model can be difficult to estimate. A restricted case clarifies the opportunity-cost argument. Hold the task portfolio, AI configuration, and oversight policy fixed. Suppose feasible specialist allocations form a nonempty interval [m, L]. Let S(x) be the net contribution of specialist learning, including its effect on reliable AI oversight, and B(L - x) the contribution of breadth. Assume both functions are twice differentiable with strictly negative second derivatives on the relevant domain, and that their contributions are additive.

$$F(x) = S(x) + B(L - x), \text{ with } m \le x \le L$$

Proposition. Under these assumptions, F has a unique maximizer. If F'(m) > 0 and F'(L) < 0, that maximizer is interior and satisfies:

$$S'(x^*) = B'(L - x^*)$$

Proof. Continuity on the compact feasible interval gives existence. Strict concavity gives uniqueness. Differentiation yields F'(x) = S'(x) - B'(L - x) and F''(x) = S''(x) + B''(L - x) < 0. The derivative is strictly decreasing; the stated endpoint signs imply a unique zero, which is the maximum. If F'(m) ≤ 0, the maximum is m; if F'(L) ≥ 0, it is L. These are the boundary alternatives.

The result says to compare marginal net benefits, not percentages. An interior 70/30 allocation is optimal in this restricted model only if x* = 0.70L is feasible and the two marginal values happen to be equal there. A binding floor could select 70 percent without marginal equality. Nothing in the proof selects that number universally. Moreover, x/L measures learning-time allocation, while D(x) measures capability: the proposition does not prove 70 percent mastery is sufficient.

Why AI need not reduce optimal specialization

AI could lower the value of manually learning a routine execution technique. It could also increase the value of deeper knowledge that enables better questions, harder projects, or reliable verification. Either change can dominate. A positive interaction between depth and breadth would violate the additive assumption, while prerequisite chains can violate concavity. Multiple local optima or strong boundary solutions are then possible, and the proposition no longer settles the choice.

This conditional result is deliberately narrower than a theorem about daily life. Its value is to identify what must be measured: marginal gains, oversight costs, constraints, and interactions. The framework becomes useful when it invites those measurements rather than treating its memorable ratio as their substitute.

8 Illustrative sensitivity and diminishing returns

Consider L = 10 normalized learning units, with S(x) = a ln(1 + x) and B(10 - x) = b ln(11 - x), where a and b are positive weights. The units and weights are invented for illustration; they have not been fitted to workers. With a feasible interval [m, 10], the optimum is the following unconstrained value clipped to that interval:

$$x_0 = (11a - b) / (a + b)$$

Constructed scenario	a	b	Minimum x	Optimal depth share
Equal contribution weights	1	1	3	50%
Higher specialist contribution	2	1	3	70%
Much higher specialist contribution	4	1	3	86%
Same weights as 70% case with binding floor	2	1	8	80%

The 70 percent result occurs because the example selects a specialist weight twice the breadth weight. Raising the breadth weight shifts the allocation toward breadth; increasing the specialist or verification contribution shifts it toward depth. A binding capability constraint can override either preference. No row is a recommended allocation for an actual profession.

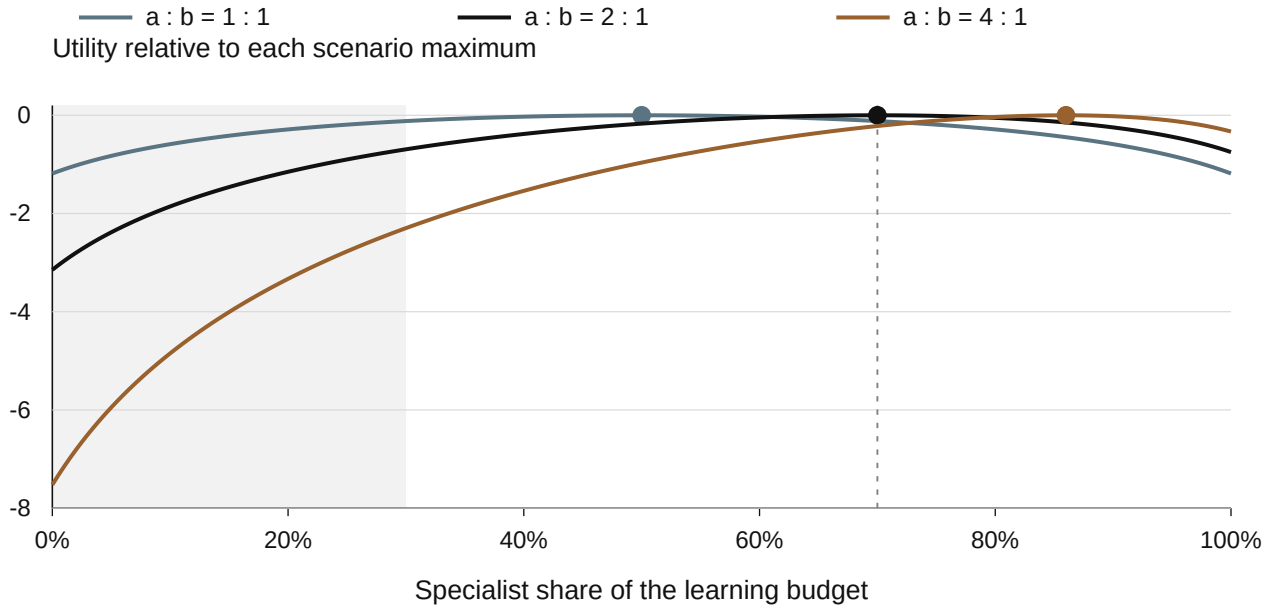


Figure 2. The preferred ratio depends on assumptions. Curves show constructed log-utility values relative to their own maxima. Points mark the 50, 70, and 86 percent optima for the first three scenarios. Gray shading is below their 30 percent floor. The dotted line is a candidate, not an estimated universal optimum.

9 Why the final increment can be expensive or essential

A second illustration separates learning time from performance. Assume a bounded benchmark score q(t) = 1 - exp(-kt), starting at zero and approaching a fixed ceiling. With the invented rate k = 0.02 per hour, reaching q = 0.70 takes approximately 60.2 hours; q = 0.90 takes 115.1 hours; and q = 0.99 takes 230.3 hours. Exact 100 percent is approached asymptotically rather than reached in finite time.

These values demonstrate one possible diminishing-return pattern: moving from 70 to 90 percent takes about 54.9 additional hours. They do not measure a real profession or show that the extra hours are wasted. If the remaining errors are costly, those hours can have greater value than the initial improvement. A different learning function, baseline, curriculum, or test would produce different numbers. The curve is a modeling assumption, not a biological law. [8]

Task coverage is not missing expertise

Suppose a practitioner reliably handles routine cases representing 70 percent of observed workload. The remaining 30 percent may contain unusual requests, ambiguous evidence, or rare high-consequence cases. Its contribution to risk need not be 30 percent. Conversely, an AI tool might assist with repetitive work in the familiar 70 percent while being least useful on exceptional cases. Delegation should follow measured task suitability, not the complement of a person's test score.

For each candidate task, compare the entire assisted workflow with a suitable alternative. An illustrative two-hour task that takes 30 minutes to draft with AI, 45 minutes to verify, and 30 minutes to repair saves only 15 minutes before coordination and tool costs. If qualified review takes another 30 minutes, the assisted process is slower. If the output cannot be checked adequately, even a large time saving is not sufficient to justify use.

This accounting makes visible a verification paradox: a person may need some of the knowledge they hoped to avoid learning in order to recognize mistakes. The appropriate response is not automatically to learn everything. It may be to narrow the task, use independently testable outputs, consult an expert, or deepen the particular capability that limits reliable oversight. The question becomes which knowledge changes the decision boundary, rather than how to stop at an attractive percentage.

10 Application across different kinds of work

The following cases illustrate how the framework could guide decisions. They are not claims that any occupation has a measured 70 percent sufficiency threshold. Each example retains human responsibility for task selection and final acceptance.

Business analysis and communication

A professional who understands the business question, data definitions, and evidential limitations might use AI for draft explanations, alternative layouts, or preliminary code. Saved time could support learning experimental design or stakeholder communication. The necessary foundation includes checking the data and the inference, not simply recognizing fluent prose. A plausible but invalid causal claim can survive stylistic review. The appropriate comparison is an equally well-supported analyst using the same AI, not an unassisted analyst working under an artificial disadvantage.

Software development

A developer might delegate boilerplate or request candidate tests while retaining responsibility for architecture, security assumptions, and integration. Automated tests help, but their coverage and correctness still matter; a passing test suite is not proof that all important requirements have been met. Breadth in user research or organizational processes may improve problem selection. For an unfamiliar, consequential subsystem, deeper specialist study or expert review may be more valuable than additional breadth. The observed benefit should include review and downstream defects, not only lines of code produced.

Licensed and safety critical practice

In medicine, law, engineering, and other regulated or high-consequence work, this framework cannot waive qualification, supervision, or competence requirements. Applicable rules must be checked in the relevant jurisdiction and institution. AI may assist with an approved administrative or information-processing task, but no numerical ratio licenses independent practice. The human capability needed for safe escalation or emergency response can be non-negotiable even when the relevant event is rare. Any study of such work should begin in simulation, not by lowering safeguards for real clients or patients.

Frontier science and theoretical physics

AI can suggest derivations, search strategies, or computational approaches, but a plausible argument is not a verified result. Novel research may require expertise precisely where available examples and familiar patterns are insufficient. A physicist must still check assumptions, mathematical steps, limiting cases, and consistency with evidence. Here, concentrated specialist work may be the highest-value use of time; breadth becomes productive when it supports a concrete connection between fields. The proposal should encourage a better research portfolio, not imply that foundational mathematics or difficult reasoning can simply be omitted.

Across these cases, a sensible human role is to frame, judge, and remain accountable for work while using appropriate assistance. Accountability requires real capacity, adequate time, and the authority to reject an output. A nominal human approval step without those conditions does not make the workflow reliable.

11 Objections and counterexamples

There may be no measurable 100 percent. A profession contains changing knowledge, tacit judgment, practical execution, and social responsibilities. A percentage makes sense only relative to an operational benchmark or curriculum. Even then, equal scores can conceal different weaknesses. The strong mastery claim should therefore be tested with capability profiles and mandatory critical items, not an aggregate score alone.

Expertise can be complementary to AI. Deep specialists may use tools to attempt work that was previously impossible or too costly. If AI raises the return to exceptional judgment, the preferred training investment may move toward greater depth. The proposition in Section 7 accommodates this through S, but a simple substitution story misses it.

Automation can redirect attention badly. In a simulated flight-monitoring experiment, Skitka, Mosier, and Burdick observed both missed events and incorrect actions associated with an imperfect automated aid. This older study does not quantify current generative-AI risk, but it establishes that assistance can change human error patterns. [12] The proposed framework must test reliance and anomaly recognition, not only the nominal reliability of the tool.

Breadth may become superficiality. A person can accumulate terminology without acquiring a useful explanatory model. A broad learner may also be less effective than a team whose members retain deep, complementary expertise. The right unit of optimization might be the team, with some people specializing deeply and others integrating across boundaries. An individually attractive allocation need not be organizationally optimal.

The most important error can be rare. Frequency-weighted coverage can underweight exceptional situations. Tail risks, correlated AI failures, and errors that remain hidden until deployment require separate evaluation. Passing routine benchmarks does not establish safety under distribution shift, unusual inputs, or adversarial conditions. NIST's generative-AI risk profile highlights risks including confabulation and human-AI configuration, reinforcing the need for context-specific controls. [13]

A short-term gain may conceal a long-term loss. Immediate output can rise while unaided capabilities decay or the profession's future expert pipeline weakens. If every novice avoids difficult practice because AI completes it, there may eventually be fewer people able to validate new methods or manage unfamiliar failures. This is a plausible mechanism requiring longitudinal research, not a demonstrated consequence of every AI deployment.

The saved time may never become learning time. Employers may increase output targets, coordination may expand, or workers may reasonably choose leisure. Reinvestment cannot be inferred from faster task completion. The framework succeeds on its own terms only if the intended benefits actually occur without unacceptable losses elsewhere. It must allow the evidence to reject both a particular ratio and the broader strategy in a particular setting.

12 Testable hypotheses and study design

A first study should test whether the allocation policy works under defined conditions, not try to confirm the label. The following hypotheses are conditional and preregisterable. None is presented as a finding.

H1 Workflow efficiency. Among participants meeting predefined competency requirements, AI assistance will improve accepted output per total labor hour on a validated, AI-suitable task portfolio. The effect may reverse on deliberately challenging or unsuitable tasks.

H2 Reinvestment. A foundation-plus-breadth policy with AI will improve performance on new cross-domain problems relative to depth-focused training with AI, without breaching prespecified quality and safety margins on core tasks.

H3 Oversight. The value of assistance will vary with baseline verification capability and independently rated task verifiability. Higher assisted output alone will not necessarily predict retained unaided ability.

H4 Allocation heterogeneity. Preferred learning allocations will differ across task portfolios and participants; a fixed 70/30 policy will not reliably dominate all neighboring allocations.

An initial factorial experiment

Recruit already-qualified professionals from two or more low-consequence work settings. Use baseline assessments to establish eligibility and characterize experience, specialist strengths, AI familiarity, and verification ability. Randomize participants, or intact teams where spillovers are likely, to the following four conditions with equal total training and work budgets.

Discretionary learning policy	No AI assistance	AI assistance
Depth focused	Arm A	Arm B
70% specialty and 30% purposeful breadth	Arm C	Arm D

Both training policies must maintain required core capabilities. The allocation applies only to a discretionary development budget, not to reducing required qualifications. All groups receive equivalent assessment preparation and non-AI resources; AI groups receive the same tool version, access, and task-relevant onboarding. Randomization is stratified by occupation and baseline capability. Evaluators remain blind to assignment where feasible.

An illustrative schedule is two baseline weeks, eight intervention weeks, and a delayed assessment four weeks later. These durations are proposed planning choices, not evidence-based optimal intervals. The design tests policy bundles and their interaction with AI. It cannot by itself establish that exactly 70 percent is optimal or isolate the causal effect of breadth separately from reduced specialist training.

13 Outcomes analysis and safeguards

Use **quality-qualified output per total labor hour** as the primary efficiency endpoint: count completed tasks only when they satisfy a prespecified, blinded acceptance rubric, and include prompting, checking, correction, and escalation time in the denominator. Count supervisor or expert-review labor as well as participant labor. Report task difficulty and case mix so that greater throughput cannot be achieved by selecting easier work.

Quality and serious errors require separate outcomes rather than being hidden in one productivity index. A policy is successful only if it improves the prespecified efficiency or transfer outcome while meeting agreed noninferiority margins for core quality and retained unaided capability, and respecting a separate serious-error limit. Margins must be justified by domain experts and stakeholders before outcomes are examined; failure to detect a difference is not proof of equivalence.

Secondary measures include delayed unaided tests, performance on genuinely new transfer tasks, calibration between confidence and correctness, detection of seeded AI errors, fatigue, perceived workload, and observed learning-time allocation. Record actual minutes spent on breadth rather than relying on intentions. Log tool versions and substantive usage, while minimizing collection of sensitive work content. A production gain and a retention loss should both be reported, even if they complicate the conclusion.

Analysis and extension

Preregister the primary contrast between foundation-plus-breadth with AI and depth-focused training with AI. Estimate intention-to-treat effects and the allocation-by-AI interaction, with baseline adjustment and uncertainty intervals. Account for clustering at the randomization level and repeated tasks within people. Report attrition, contamination, treatment adherence, and missing-data sensitivity analyses. Limit confirmatory subgroup tests and adjust for multiple primary comparisons. Determine sample size by prospective power analysis using pilot variance and cluster correlations, not an invented universal participant count.

A second-stage study can randomize several permissible allocations, such as 50/50, 70/30, and 90/10, while retaining the same competency safeguards. Fit a prespecified response curve and compare the uncertainty around its optimum. Those points may be too sparse to locate a sharp optimum; additional allocations should be planned if needed. To test the mastery claim directly, separately measure validated capability profiles and task outcomes. Never equate assigned training time with achieved competence.

Participation requires informed consent and appropriate ethics review. Begin with synthetic, de-identified, or safely sandboxed tasks; no real-world safety-critical work should be delegated merely to test the hypothesis. Define independent escalation and stopping procedures for serious errors, privacy incidents, or unacceptable capability loss. No arm should restrict sleep. The study described here is a proposal, not an authorized or completed experiment, and it supplies no outcome data.

14 A practical implementation protocol

An individual or organization can use the framework before a large trial, provided the exercise remains low risk and does not substitute for professional standards. The aim is to find a better local allocation, not to prove the proposed ratio correct.

First establish the baseline. Choose a small, recurring task family and record complete-workflow time, acceptance quality, error types, and review effort. Specify what good performance means and which mistakes are unacceptable. Keep the initial comparison narrow enough that task mix does not change unnoticed. Record current unaided ability on the capabilities needed to supervise the workflow.

Then set the human floor. Identify knowledge that is essential for framing the problem, checking evidence, recognizing exceptions, and deciding when to escalate. Preserve required training and access to qualified help. If a person cannot tell whether the workflow is failing, improve the checks or the person's capabilities before expanding delegation.

Test assistance rather than assuming it. Compare matched tasks with and without the tool where practical, or use a randomized sequence for a personal low-risk experiment. Include retries and external review. Treat an initial improvement as provisional because novelty, task choice, and familiarity can distort short comparisons. Do not send confidential information to an unapproved system.

Allocate verified savings deliberately. If a real saving remains after oversight, reserve a feasible part of it for an adjacent learning goal and preserve a part for recovery or contingencies. A candidate 70/30 specialist-to-breadth split can be applied to the discretionary learning budget, not to all waking time. Document why the selected adjacent subject should improve a concrete decision or capability.

Make learning demonstrable. Pair study with an unaided explanation, a problem solved in a new context, and feedback from an appropriate source. Use AI to generate practice or critique a solution when suitable, but retain independent attempts. Protect blocks of focused specialist work when the problem requires sustained reasoning rather than frequent switching.

Review and revise. Recheck the workflow after meaningful tool changes, shifts in task difficulty, or observed errors. Reduce or stop delegation when review costs erase savings, important errors rise, or unaided capability weakens. Increase specialist study when it improves oversight or opens higher-value work. Reduce breadth when it remains disconnected from useful understanding. Maintain a record of what changed and why.

The resulting plan may be 70/30, 90/10, 50/50, or a changing sequence across a career. The framework should help a person choose and revise that plan, not penalize them for departing from its name. Its most important operational rule is that claimed productivity gains must survive complete-workflow accounting and independent quality checks.

15 Limitations and conclusion

This paper offers a selective, mechanism-focused synthesis rather than an exhaustive literature review. Sources were chosen for direct relevance to AI productivity, learning, transfer, sleep, and oversight, with preference for original studies and official guidance. The reviewed evidence spans different populations and technologies; it does not support a pooled effect estimate. Some findings come from short experiments or narrow deployments. Rapid changes in tools require fresh local evaluation rather than treating any cited result as a permanent capability forecast.

The formal model is intentionally simplified. It abstracts from motivation, bargaining power, career incentives, unequal access to tools, and the difficulty of measuring judgment. The illustrative utility weights and learning rates are not empirical estimates. Real learning may be discontinuous, task values uncertain, and specialist knowledge complementary to breadth. The conditional proposition should not be applied when its assumptions are implausible without first changing the model.

The proposed study would still face contamination, participant selection, imperfect blinding, and measurement burdens. An eight-week intervention may miss career-scale effects, especially the development of rare expertise and long-term dependence on assistance. The strong claim that a person needs only 70 percent of a specialty remains unresolved until the relevant capability scale, population, task portfolio, and comparator are defined. A favorable trial in one setting would not establish a universal law.

Conclusion

Kalu Wu's central insight is that AI changes the opportunity cost of how humans learn and work. It is reasonable to ask whether some effort devoted to incremental specialization could produce more value when redirected toward broader understanding. That question deserves serious study. The accompanying sleep example expresses a useful concern for balance, but it does not determine the correct ratio for learning or work.

The defensible framework preserves strong human foundations, uses AI selectively where the complete workflow is reliable, and reinvests genuine savings in chosen capabilities or recovery. Deep expertise remains necessary in many settings and may become more valuable when it enables better use of AI. Breadth becomes productive when it improves real reasoning across domains. Neither should be reduced to a fixed share of an undefined whole.

Accordingly, the 70 Percent Ratio Theorem is best treated at present as a named, testable allocation hypothesis. Seventy percent is a candidate to investigate, not a scientifically established stopping point. The principle supported by this analysis is conditional optimization under human-capability, quality, and safety constraints. Its success should be judged by accepted work, retained judgment, broader understanding, and sustainable effort, not by whether a schedule preserves the original number.

Disclosure. Concept proposed by Kalu Wu. This manuscript uses AI-assisted research, synthesis, mathematical illustration, and document preparation. It reports no original participant data and claims no institutional affiliation, external peer review, or established priority for the underlying ideas.

References

The numbered references link to publisher records, research papers, or official guidance. Publication versions were checked for this conceptual synthesis on 25 October 2026. All figures and numerical scenarios in this paper are original explanatory constructions; no third-party figure is reproduced.

[1] Watson, N. F., Badr, M. S., Belenky, G., Bliwise, D. L., Buxton, O. M., Buysse, D., Dinges, D. F., Gangwisch, J., Grandner, M. A., Kushida, C., Malhotra, R. K., Martin, J. L., Patel, S. R., Quan, S. F., and Tasali, E. (2015). Recommended amount of sleep for a healthy adult: A joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. Journal of Clinical Sleep Medicine, 11(6), 591-592. https://doi.org/10.5664/jcsm.4758

[2] Noy, S., and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187-192. https://doi.org/10.1126/science.adh2586

[3] Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889-942. https://doi.org/10.1093/qje/qjae044

[4] Dell'Acqua, F., McFowland, E., III, Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Organization Science. Published online 11 March 2026. https://doi.org/10.1287/orsc.2025.21838

[5] Becker, J., Rush, N., Barnes, E., and Rein, D. (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. Research preprint, arXiv:2507.09089. https://arxiv.org/abs/2507.09089

[6] Becker, J., Rush, N., Cunningham, T., Rein, D., and Mahamud, K. (2026, February 24). We are changing our developer productivity experiment design. METR research update. https://metr.org/blog/2026-02-24-uplift-update/

[7] Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100(3), 363-406. https://doi.org/10.1037/0033-295X.100.3.363

References continued

[8] Heathcote, A., Brown, S., and Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. Psychonomic Bulletin and Review, 7(2), 185-207. https://doi.org/10.3758/BF03212979

[9] Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

[10] Roediger, H. L., III, and Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

[11] Gick, M. L., and Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15(1), 1-38. https://doi.org/10.1016/0010-0285(83)90002-6

[12] Skitka, L. J., Mosier, K. L., and Burdick, M. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991-1006. https://doi.org/10.1006/ijhc.1999.0252

[13] Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., and Roberts, K. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile. NIST AI 600-1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Reproducibility of the illustrative calculations

For Section 8, evaluate $F(x) = a \ln(1 + x) + b \ln(11 - x)$ on the stated feasible interval and use $x^* = \max(m, \min(10, (11a - b)/(a + b)))$. In Figure 2, each curve is displayed as $F(x) - F(x^*)$ so that its maximum is zero; these are arbitrary utility differences, not percentages or observed productivity changes.

For Section 9, solve $q = 1 - \exp(-0.02t)$ as $t = -\ln(1 - q)/0.02$. Substitution of $q = 0.70$, 0.90, and 0.99 gives 60.1986, 115.1293, and 230.2585 hours respectively. Rounding produces the values in the text. These calculations demonstrate implications of the selected functions only.

The study protocol in Sections 12 and 13 would require a separate preregistration, ethics review where applicable, finalized instruments, and a prospective analysis plan before recruitment. This manuscript is not that registration and contains no implied approval to conduct the study.