

The Labour That Output No Longer Requires

A Protocol for Measuring Required Human Time and the Saved Hours Question

Dennis Hedegreen

Hedegreen Research

Publication edition | Paper v1.0

Publication date: 27 September 2026.

Deposit DOI: [10.5281/zenodo.23002321](https://doi.org/10.5281/zenodo.23002321).

Licence: Creative Commons Attribution 4.0 International.

This edition incorporates the revised source-state and duration qualifications. No new independent coding exercise has been performed.

Abstract

Technological change is usually described as rising output per unit of labour. This paper reverses the perspective and asks how much **required human person-time** a defined output no longer needs. It uses the term **Labourfold** for the canonical fold ratio

$$LF = \frac{H_0^{req}(O, Q^*)}{H_1^{req}(O, Q^*)},$$

where comparator and target states deliver the same declared output and satisfy the same material service/admissibility rule. The arithmetic is conventional: at fixed output it is equivalent to familiar unit-labour and relative labour-productivity relationships. The proposed contribution is instead a cross-domain **claim protocol** that combines established work-measurement, functional-unit, quality, boundary and sensitivity disciplines into a codable structure that can return **canonical Labourfold**, a **proxy human-time ratio**, **no valid scalar**, or a **zero-human-time endpoint**. The protocol distinguishes state-conditioned required human person-time from realised, elapsed and scheduled time, and from under-identified human-time evidence; requires explicit output, quality, boundary, state and horizon definitions; follows material service functions across actor/vendor/customer shifts; and separates evidence acquisition, state realisation and causal identification. Two LLM held-out calibration runs matched all 16 pre-specified primary branches; later same-evidence LLM exercises exposed under-specification and triggered protocol repairs; and later same-evidence LLM coding sessions matched a proxy branch and the then-specified bounded R1/R2 classification on author-prepared source packets. A subsequent source recheck separated the modelled assembly timing from an earlier physical demonstration; the numerical interval is retained only as a conditional standard-time illustration. These exercises assess rule application to selected evidence, not independent source extraction, demonstrated time savings, or mature multi-domain or human-expert validation. The absolute counterpart, **Saved Hours**, denotes technical fixed-output release of required human person-time, not leisure, employment loss, payroll saving or any predetermined allocation.

Keywords: labour productivity; automation; working time; human labour; AI; productivity measurement; Saved Hours; Labourfold

1. The Simple Ratio

Suppose a defined output requires ten hours of human labour in one production state and two hours in another. Before asking why the change occurred, whether anyone loses a job, or where any released capacity goes, a smaller question can be answered first: **how much human person-time does the output require in one state relative to the other?**

The smallest useful representation is:

$$LF = \frac{H_0}{H_1}.$$

For the simple case, $LF = 10/2 = 5$. I call this ratio **Labourfold**. A Labourfold of 5 means that the comparator state requires five times as much human person-time as the target state for the production object being compared. When $LF > 1$, required human time is lower in the target state; when $LF = 1$, it is unchanged; when $0 < LF < 1$, it is higher. Nothing in the ratio yet says anything about cost, energy, employment, welfare, or desirability.

The arithmetic is not new. Let unit labour requirement be

$$u_i = \frac{H_i}{O},$$

and labour productivity be

$$p_i = \frac{O}{H_i}.$$

Holding output fixed,

$$LF = \frac{H_0}{H_1} = \frac{u_0}{u_1} = \frac{p_1}{p_0}.$$

Labourfold is therefore a fold-form presentation of familiar unit-labour and labour-productivity relationships, not a new production identity. Output per hour and its reciprocal, unit labour requirements, are longstanding productivity measures (U.S. Bureau of Labor Statistics, 1977; OECD, 2001). The historical antecedent is stronger than a shared equation. Jerome (1934) used the phrase “**labor-saving ratio**” and distinguished **operating, auxiliary, embodied, and indirectly required labour**, explicitly warning that the measured labour-saving effect changes with the labour scope used. Industrial work measurement already supplies standard-time methods; LCA and social LCA supply functional-unit and system-boundary discipline; worker-hour and MRIO approaches trace upstream labour; and automation economics already separates task displacement from downstream labour-market effects (IISE, n.d.; ISO, 2006; UNEP/SETAC, 2009; Acemoglu & Restrepo, 2019; Springer, Wulf & Zapp, 2025; Godé, Mair & Gómez-Baggethun, 2026).

The proposed contribution is therefore **integration and codification**, not primitive invention. Labourfold couples conventional human-time arithmetic to one claim architecture that declares the output/service object, material quality rule, human-labour boundary, production state and horizon; codes the evidential basis of human time; and permits the result to be canonical, proxy, refused, or a zero-human-time endpoint. It is not a replacement for productivity statistics, MTM/PMTS, LCA/S-LCA, MRIO labour-footprint methods, or task-based automation economics.

A compact comparison helps locate the contribution:

Tradition	What it already contributes	What Labourfold uses it for
Jerome / productivity measurement	labour-saving ratios, output adjustment, labour-scope dependence	historical and arithmetic foundation
Work measurement / MTM / PMTS	standard time for declared methods	R-class evidence when necessity is supported
LCA / S-LCA	functional units, system boundaries, worker-hour scaling	output/boundary discipline
MRIO / labour footprints	upstream and embodied labour tracing	B-X extended claims
Task-based automation economics	displacement, productivity and new-task effects	downstream interpretation, not the LF estimand
Specification-sensitivity methods (Steege et al., 2016; Simonsohn et al., 2020)	consequences of defensible analyst choices	Q* and boundary sensitivity

No individual row is claimed as novel. The methodological proposal is the combined decision/reporting column.

This discipline is needed because the ratio is easy to overstate. Saying “this technology has Labourfold 5” already hides the comparator. Change the output, operating conditions, boundary, or reference state and the number may change without any arithmetic error.

The first failure is simple. If the ten-hour state produces 100 units but the two-hour state produces only 10, $10/2 = 5$ remains mathematically correct but no longer describes a fivefold reduction in human time for the same output.

Labourfold is relational from the first line. It compares human-time requirements between specified production states; it is never a free-standing constant of a technology.

The equation therefore earns only one conclusion at this stage. If two states refer to the same production object and one requires ten human hours while the other requires two, the target state exhibits fivefold human-time compression relative to the comparator.

That immediately forces the first refinement: **ten hours and two hours for the same what?**

2. Same Output

The simple ratio silently assumes that the numerator and denominator refer to the same output. Once that assumption is made explicit, the equation becomes:

$$LF(O) = \frac{H_0(O)}{H_1(O)},$$

where O is the declared output or service object.

The requirement to define a common service basis also has established methodological precedent. Life-cycle assessment uses a **functional unit** as the quantified performance of a product system that serves as the reference for comparison, and social LCA commonly scales worker-hour activity variables to such functional units (ISO, 2006; UNEP/SETAC, 2009; Springer, Wulf & Zapp, 2025). Labourfold uses the same kind of comparability discipline for a narrower question: required human person-time across two production states.

The change looks minor. It is not. A large share of apparent productivity comparisons become ambiguous as soon as the output object is examined carefully.

Consider a simple quantity error. A workshop requires ten human hours to make 100 components under one process and five hours to make 50 under another. Reporting $LF = 2$ would confuse lower total labour with lower labour required for the same output. The natural correction is to hold the output quantity fixed or to convert both states to a defensible common unit.

But output is not always a single count. Suppose one production state jointly produces:

$$(100A, 100B)$$

and another produces:

$$(180A, 20B).$$

Both produce 200 items. Yet the output bundles are not equal. Adding the item counts erases the product mix, and treating both states as $O = 200$ can create a Labourfold claim where no fixed-output comparison exists.

For such cases, the output object may be better represented as a bundle or vector:

$$\mathbf{O} = (O_1, O_2, \dots, O_k).$$

The point is not to make the notation more impressive. It is to prevent the arithmetic from manufacturing equivalence. A scalar output unit should be used only where one common service unit is substantively defensible. A transport process, for example, may sometimes be compared in tonne-kilometres under a fixed service specification. A calculation task may be compared as completion of the same problem. A cooling process may be compared as removal of a defined amount of heat from a defined quantity of product under specified conditions. In other cases, the bundle must remain multidimensional.

Prices do not solve this problem for Labourfold. If two output bundles are converted into revenue and then declared equivalent because their market value is equal, the comparison has changed objects. It may be a useful economic productivity calculation, but it is no longer a pure measure of human person-time required for the same defined output. Labourfold is intentionally unwilling to use money as a universal solvent for physical or service differences.

Shared production creates a second trap. If one human hour simultaneously supports products A and B , the same hour cannot be credited in full to each product and then counted twice.

Where the labour is genuinely joint and no defensible allocation exists, the cleanest Labourfold claim is often at the level of the smallest output bundle that actually requires the shared hour.

The output discipline is therefore stricter than “same number of things.” It asks whether the two states deliver the same declared production object. That object may be one unit, a service, a bundle, or a vector. The representation can be simple when reality is simple; it must remain explicit when reality is not.

This yields the second surviving rule:

**Labourfold compares human time for the same declared output object.
Equal totals are not enough when product mix or service content differs.**

With O held fixed, the ratio is more meaningful. Yet an obvious loophole remains. A system can use less human time by producing the same quantity **worse**.

So the next question is not merely whether the output is the same. It is whether the output still satisfies the same functional requirements.

3. Same Quality

A production system that halves human time by doubling the defect rate has not necessarily compressed the labour required for the same service. Nor has a medical workflow that saves reporting time by accepting materially worse diagnostic performance, a logistics system that becomes faster by allowing more damage, or an AI workflow that appears efficient only because humans stop checking errors that the service cannot tolerate.

The canonical Labourfold comparison therefore becomes:

$$LF = \frac{H_0(O, Q^*)}{H_1(O, Q^*)}.$$

Here Q^* denotes the declared quality or functional-admissibility rule associated with the output. Like O , it need not be a single number. Depending on the domain, Q^* may include accuracy, reliability, latency, safety, durability, error severity, service availability, regulatory admissibility, or another characteristic without which the two outputs are not functionally comparable.

The important phrase is **declared requirement**. Labourfold does not require realised quality to be numerically identical when the scientific object is a common service standard. Define an admissibility set

$$Q^* = \{q : q \text{ satisfies the predeclared material service requirements}\}.$$

Suppose the relevant service standard requires accuracy of at least 95 per cent and the new system achieves 99 per cent while using less human time. It would be perverse to degrade the new system until it performs exactly like the old one. Both states must lie inside the same Q^* ; a target may exceed the standard.

That flexibility creates a shopping risk, so the admissibility rule must be justified independently of the Labourfold/Saved-Hours result. The protocol therefore records whether Q^* is **anchored**—for example by regulation, contract, preregistration, a baseline service obligation, or a recoverable domain standard—or **researcher-specified**. Material quality dimensions cannot be dropped merely because they make the target state look better. If a material threshold, equivalence band, weighting rule, or multidimensional admissibility surface is researcher-specified, materially plausible alternatives under the **same frozen research question and output/service object** must be sensitivity-tested. If plausible alternatives change scalar eligibility, required human time, or the interpretation of Saved Hours, one unqualified scalar is not permitted; the analyst reports conditional results, a coherent sensitivity

family, or NS-Q. This is a constrained specification-sensitivity rule, not an invitation to enumerate every imaginable quality philosophy.

A further **constitutive-quality identity test** is needed because process completion alone can be misleading. The analyst asks whether there is a quality dimension without which the output would no longer count as the same functional service even if both workflows reach the same completion step. Diagnostic correctness, safety-critical failure probability, structural integrity, legal validity, preservation of a material biological or chemical property, and identity/authenticity can be constitutive depending on the domain. If such a dimension is material and the evidence does not establish a common admissibility rule, the protocol returns NS-Q rather than treating procedural completion as equivalence.

The quality condition also matters when apparent labour saving is purchased by hidden human assurance work. Consider a baseline process requiring 100 human hours. Raw automation appears to reduce the direct human requirement to 20 hours, suggesting:

$$LF = \frac{100}{20} = 5.$$

Now suppose the automated state introduces a safety failure mode that requires 30 additional hours of human verification to meet the same service requirement as the baseline. The relevant denominator is not 20. It is:

$$H_1 = 20 + 30 = 50,$$

so:

$$LF = \frac{100}{50} = 2.$$

The automation still compresses required human time, but by a factor of two rather than five. The verification hours are part of the human time required to produce the output **at the declared service rule**, not a downstream inconvenience to be mentioned after the productivity result.

This point is especially important in AI-assisted knowledge work, where generation can be extremely fast while verification, correction, moderation, legal review or safety oversight remains human. It is equally relevant outside AI: inspection, cleaning, calibration, testing, rework and quality assurance can all move labour away from the visibly productive step without making the labour disappear.

A chest-radiograph study (Ahn et al., 2022) illustrates the same discipline. AI assistance reduced reported median reading time while measured sensitivity improved and specificity showed no general penalty. The example is useful precisely because faster does not establish acceptable service quality by itself. Its headline median timing statistics are not converted here into a canonical Labourfold scalar, because a ratio of medians is not the same object as a ratio of compatible total or expected required person-time.

The third surviving rule is therefore:

Both states must satisfy the same predeclared material service/admissibility rule. Any human work required to restore or sustain that equivalence belongs inside the human-time requirement.

At this point the equation is compact enough to look finished. It is not. The most dangerous symbol is now neither O nor Q . It is H .

What, exactly, counts as a human hour?

4. What Is a Human Hour?

It is tempting to identify human labour with whatever a dataset records: paid hours, scheduled hours, keyboard activity, active task minutes, prompts sent, or elapsed time between start and completion. None of these is automatically the object required by the canonical Labourfold claim.

The canonical target is **state-relative required human person-time**. For a declared production state Z , boundary B , output/service object O , quality rule Q^* , and material horizon τ , let $\mathcal{C}(Z, B, O, Q^*, \tau)$ denote the material human-time components required by that declared state. Conceptually,

$$H^{req}(O, Q^* \mid Z, B, \tau) = \sum_{c \in \mathcal{C}(Z, B, O, Q^*, \tau)} \tau_c^{req}.$$

Every included component must pass the necessity test and, for a canonical scalar, have an R-class evidential basis. This object is not observed elapsed time merely because it occurred, scheduled or paid time merely because it was allocated, or a mathematical minimum across every alternative workflow that could exist. An auditable work-measurement standard such as MTM/PMTS may support R1/R2 for the **declared method/state**; it does not by itself prove a lower envelope across all alternative configurations. A separate engineering optimisation estimand may be defined when that is genuinely the research question.

The coding manual uses a simple hierarchy:

Code	Human-time basis	What may be inferred
R1	verified required person-time	canonical point estimate may be possible
R2	bounded required person-time	canonical interval/sensitivity claim may be possible
P1	realised engaged-time proxy	proxy ratio only
P2	elapsed task-time proxy	proxy ratio only
P3	scheduled/paid/staffing proxy	proxy ratio only unless necessity is independently established
U	materially under-identified	no scalar for the requested claim

A material component is R1 only when necessity is supported—for example by a removal or coverage test, a binding process dependency, a formal safety/regulatory requirement that defines the state, an auditable work-measurement standard, or repeated operational evidence showing failure below the requirement. R2 is used where required time is defensibly bounded but not known as a point. Observed human time does not become $H^{\wedge\{req\}}$ merely because it was measured precisely.

Necessity of an action and identification of its duration are separate evidential questions. An assembly may require ball insertion without that fact establishing an exact number of seconds. A standard-time claim must identify its reference method and pace, applicable worker/operating conditions, treatment of allowances and interruptions, and material uncertainty. R1 denotes a point estimate justified under those declared assumptions, not an error-free physical constant. If material duration uncertainty can be bounded, use R2; otherwise downgrade or refuse the requested required-time claim. A work-measurement label alone cannot establish these conditions.

Standby illustrates the distinction. Imagine an operator covering an eight-hour safety-critical service window. In the baseline the operator spends four hours actively intervening and four hours waiting while required to respond. Automation reduces active intervention to one hour, but the same response-time and safety requirement still binds the person for the full eight-hour window. Counting active minutes gives $4h \rightarrow 1h$, but the required person-time remains

$$H_0^{req} = 8h, \quad H_1^{req} = 8h,$$

so the canonical result is $LF = 1$. The operator is less busy; the production state has not released the operator's time.

The converse matters. If ten minutes in each hour genuinely become unencumbered—no monitoring, response duty, or other required task—those intervals can be technically released even when they are fragmented and difficult to turn into a shorter shift. Fragmentation is then an allocation-feasibility issue, not a reason to put the minutes back into the technical requirement.

Person-time accounting also prevents double counting. Four production lines that once require four separate operator-hours may later be supervised by one operator for one common hour. The target contains one person-hour, not four. Two different people required simultaneously still contribute two person-hours. Shared task labels do not multiply time.

Knowledge work adds a measurement problem. A visible communication trace is not equivalent to cognitive labour: a one-word instruction can follow seconds or minutes of required reading and judgement. Prompt count, tokens, keystrokes, or response latency are therefore possible traces, not definitions of H . Deliberate off-interface reasoning may count; unrelated calendar time does not. Autonomous machine or agent runtime also does not become human labour when the person is genuinely free.

The canonical/proxy distinction is therefore part of the result rather than a footnote:

Canonical Labourfold requires commensurable R-class human-time bases. P-class evidence can support useful proxy human-time ratios, but those ratios are not silently relabelled as required-time Labourfold.

Once the human-time basis is coded, the next question is whose time belongs inside the claim.

5. Where Does the Process End?

A firm can reduce its own employee hours without reducing the human time required by the wider service. Self-service makes the point cleanly. If checkout work moves from a cashier to the customer, an organisational boundary may show a large internal labour saving while the service-system human requirement falls much less—or not at all.

Boundary-aware labour measurement is not new. Jerome (1934) already distinguished operating, auxiliary, embodied and indirectly required labour and showed that appraisal changes with the scope chosen. OECD productivity measurement warns that outsourcing can raise measured gross-output labour productivity when internal labour is replaced by intermediate inputs, while LCA/S-LCA and labour-footprint methods formalise broader system boundaries (OECD, 2001; UNEP/SETAC, 2009; Godé, Mair & Gómez-Baggethun, 2026). Labourfold therefore treats the boundary as part of the estimand rather than as an afterthought.

The coding protocol distinguishes three boundary classes:

- **B-P — process/organisational:** human person-time inside a declared local process, site, or organisation;
- **B-S — service-system human boundary:** the default for an unqualified claim of human-time release;
- **B-X — extended/upstream:** an explicitly declared wider boundary for recurring or embodied upstream labour.

B-S follows the **service function**, not the employer, legal entity, physical location, or technical layer. A material human function belongs in ordinary B-S when it operates, assures, restores, resets, or maintains the declared service under O/Q^* , is recurrently relevant over the declared horizon, and can be defensibly attributed to that service. This can include direct execution, exception handling, service-specific verification, required moderation, runtime/model operations, recurring maintenance/recovery, or customer/patient/citizen task-time that replaces provider task-time.

The anti-relocation rule is simple: moving the same material service function from employee to contractor, vendor, customer, remote operator, or another technical layer does not make the human time disappear. If a shared service-specific function belongs in B-S but its quantity cannot be defensibly allocated, the analyst uses the existing shared-labour/bundle rules rather than excluding it as “infrastructure.”

B-S also needs a stopping rule. Causal necessity alone is insufficient. General electricity-system labour, generic telecom backbone operation, ordinary datacentre facilities work, standard equipment manufacture, or generic office cleaning may be necessary for a service to exist, but they are not automatically part of the ordinary service-operation boundary. Those general-purpose upstream functions belong to B-X when the research question intentionally traces them.

For material boundary changes, the claim record therefore maps the function, service coupling, recurrence, actor/location in each state, attribution rule, and boundary decision. If two plausible B-S mappings under the same frozen service claim change the direction across $LF = 1$, move either state's human-time total materially, change scalar class, or alter the Saved-Hours interpretation, the analyst must resolve the mapping from evidence, report boundary-conditional results, or return NS-B for an unqualified claim.

A labour saving is not established merely because work leaves the visible organisation. The service function must be followed far enough to determine whether human labour was removed, relocated, or pushed into an explicitly extended upstream boundary.

With the human boundary declared, the next question is whether the compared production states are actually the same state except for the change under study.

6. The Production State Is More Than the Tool

Productivity discussions often attach an effect to a named technology: a tractor, a robot, a software product, an AI model. Labourfold needs a more careful object. Human-time requirements belong to **production states**, not to tool names in isolation.

Consider an AI service that begins as a chat assistant requiring manual copy-paste and testing, then gains repository access, terminal tools, an agent harness and automated tests. The product name—or even the base model—could remain unchanged while the human-time requirement changes sharply. The production state, not the brand name, is therefore the comparison object.

The opposite can also happen. Hold the software fixed and change the workload. Suppose an automated system handles easy cases in 0.1 human hours and hard cases in 2 human hours. With 90 easy and 10 hard cases, the target state requires:

$$H_1 = 90(0.1) + 10(2) = 29h.$$

If the same unchanged system later faces 10 easy and 90 hard cases:

$$H_1 = 10(0.1) + 90(2) = 181h.$$

The technology did not deteriorate. The case distribution changed. A Labourfold estimated in the first regime cannot simply be carried into the second.

Where useful, the production state can be described compactly as:

$$Z = (V, D, \Omega, W, M, \dots),$$

where V captures technology/version/configuration, D the workload or case distribution, Ω the operating environment, W the relevant worker population or composition, and M maturity or exposure where learning materially changes human time. This notation is provenance, not a demand to model every possible variable.

Maturity deserves particular care. Comparing a mature incumbent process with a week-old automation prototype can make automation look worse than its later steady state. Comparing an immature baseline with a highly tuned automated system can make it look better. There is no universal rule that both states must have the same calendar age. There is a rule that

the maturity condition must not be hidden when it materially affects H , and that projected mature performance cannot be presented as already observed.

Worker composition creates a similar issue. Automation may leave specialists handling only difficult exceptions, so time per human-handled case rises while total human time over the full case universe falls. The denominator must remain the common output object, and expert-only results should not be transported to other worker populations without evidence.

Version drift matters especially for hosted AI services, whose model weights, routing, system prompts, safety layers, interfaces, and tool access can change without a user receiving a clean scientific version identifier. This does not make Labourfold measurement impossible. A study can define a time-bounded black-box production state: the service as actually deployed, under documented settings, during a specified period. What it cannot do is assume that an observation from that state automatically remains valid for a successor system six months later merely because the brand name is unchanged.

Production-state discipline also prevents causal overreach. If H falls from ten hours to two between states Z_0 and Z_1 , Labourfold describes the difference:

$$LF_{0 \rightarrow 1} = 5.$$

It does not automatically identify which part of the change was caused by a machine, new workflow, better worker training, changed inputs, organisational redesign, or their interaction. Causal attribution requires a separate counterfactual argument. In some real systems the meaningful intervention is the entire package. Labourfold does not need to allocate causal credit in order to measure the state comparison.

The surviving rule is therefore:

Labourfold belongs to declared production states. Technology version, workflow, case distribution, operating environment, worker composition, and maturity must be visible when they materially affect the human-time requirement.

Once the state is properly defined, one more source of instability remains. The same state can behave differently at different scales and across different time horizons.

7. Scale and Time

A technology can reduce marginal human labour while adding fixed human overhead. In that case there is no reason to expect one scale-free Labourfold number.

Suppose the baseline human-time requirement is one hour per unit:

$$H_0(O) = O.$$

A new system requires 40 hours of setup but only 0.1 human hours per unit thereafter:

$$H_1(O) = 40 + 0.1O.$$

Then:

$$LF(O) = \frac{O}{40 + 0.1O}.$$

At ten units, the new state is much more human-labour intensive:

$$LF(10) = \frac{10}{41} \approx 0.24.$$

At 100 units:

$$LF(100) = \frac{100}{50} = 2.$$

At 1,000 units:

$$LF(1000) = \frac{1000}{140} \approx 7.14.$$

Nothing inconsistent has happened. The output scale changed. A headline such as “this technology has Labourfold 7” would conceal the output level at which the number is true.

Batch structure can matter just as much as total output. If the system requires five human setup hours per batch, producing 1,000 units in one batch is not human-time-equivalent to producing the same 1,000 units in 100 small batches. Timing, freshness, availability, product mix, or storage constraints may make the small-batch cadence part of the service. Extending the measurement horizon must not be used to make recurring setup work disappear by counting it once across an unrealistically aggregated output.

The same rule applies across recurring temporal regimes. Suppose a quiet period has:

$$100h \rightarrow 25h, \quad LF_q = 4,$$

while a peak period has:

$$300h \rightarrow 375h, \quad LF_p = 0.8.$$

Averaging the two ratios gives:

$$\frac{4 + 0.8}{2} = 2.4,$$

which suggests strong overall compression. But the actual person-time totals are:

$$H_0 = 100 + 300 = 400h,$$

$$H_1 = 25 + 375 = 400h,$$

so over the complete two-regime horizon:

$$LF^\tau = \frac{400}{400} = 1.$$

The correct general rule is to aggregate compatible person-time ledgers first and then take the ratio:

$$LF^\tau = \frac{\sum_r H_{0,r}}{\sum_r H_{1,r}}.$$

Period-specific Labourfold ratios should not be averaged as though they were hours. Labourfold is dimensionless; human hours accumulate, and the ratio is recomputed from those accumulated hours.

Temporal boundaries create a further opportunity for illusion. Suppose a Monday process falls from 100 human hours to 20. If the Monday output now requires 60 hours of preprocessing

on Sunday, the end-to-end requirement is 80 hours, not 20. Similarly, deferred verification, repair, or maintenance can make a short measurement window look human-light by pushing work just outside the observation period.

A robust comparison therefore closes material stocks and liabilities across the window where they affect the claim: inventory, work in progress, verification backlog, maintenance backlog, or other deferred obligations. Saved human time cannot be created merely by moving required work into the previous or next period.

One-off transition work is different again. Integration, commissioning, migration, and technology-specific training can be real human costs without being recurring operating labour. If a system reduces operating labour from one hour per unit to 0.1 but requires 900 one-off human hours to install and learn, the operating Labourfold is still 10. Yet the cumulative human-time gain remains negative until enough output has been produced to repay the transition work. Those are two legitimate questions and should be reported separately rather than blending the transition into an arbitrary hidden “amortisation period.”

Scale and time therefore do not require a replacement equation. They require the claim to state the relevant output scale, batch/cadence, horizon, and recurring regimes where material.

The surviving rule is:

A production state can have a Labourfold function rather than one universal number. Aggregate compatible human hours before taking ratios, and do not manufacture Saved Hours by shifting required work across scale or time boundaries.

At this point the conceptual object is reasonably well specified. The remaining problem is epistemic: how do we know that the hours, states, and boundaries we have written down are actually supported by the evidence?

8. From Ratio to Protocol

At this point the arithmetic is easy enough to become dangerous. Two time figures can almost always be divided. The protocol’s main task is to decide **what the ratio may legitimately be called**.

The sequence is: freeze the comparator and target states; establish a common output 0; define the material service/admissibility rule Q^* ; declare a symmetric human boundary B; surface material scale, horizon, maturity, workload and version conditions; code the human-time basis on each side; and only then decide whether a scalar is admissible. The protocol has four outputs:

Protocol output	Condition
Canonical Labourfold	$0/Q^*$, boundary and state comparability pass; both sides use commensurable R-class required-person-time bases; $H_1^{req} > 0$
Proxy human-time ratio	comparability passes, but at least one side uses P-class realised, elapsed, scheduled or aggregate time rather than identified H^{req}
NO VALID SCALAR	a material output, quality, boundary, hour-basis, state, temporal, joint-labour, feasibility or under-identification failure remains unresolved
Zero-human-time endpoint	$H_1^{req} = 0$; report the target zero directly rather than an infinite ratio

For positive target required time:

$$LF = \frac{H_0^{req}(O, Q^*)}{H_1^{req}(O, Q^*)}.$$

If $H_1^{req} = 0$ and $H_0^{req} > 0$, the protocol reports H1=0, 100 per cent technical human-time release within the declared boundary, and Saved Hours equal to H0. It does not manufacture “infinite productivity.” Refusal is also coded: NS-0 for output mismatch, NS-Q for service/quality mismatch, NS-B for boundary mismatch, NS-H for incompatible human-time bases, NS-Z for irreparable state non-comparability, NS-T for temporal closure, NS-J for unresolved shared labour, NS-F for an infeasible requested target state, and NS-U for residual material under-identification.

Evidence provenance is recorded on separate axes because one label cannot say how time was obtained, whether the state was realised, and whether the comparison is causal. **Acquisition** is Direct / Reconstructed / Modelled; **state realisation** is Operational / Demonstrated / Conditional; and **causal status** is separately descriptive, randomised/credible, quasi-experimental, or unresolved. These axes do not determine canonical status by themselves. A modelled MTM standard can be R-class if it is a defensible work-measurement requirement, while a directly observed task timer can remain a P-class proxy.

The minimum coding record is deliberately compact, but conditional safeguards are mandatory when triggered:

Field	Minimum entry
Claim	research question; Z0 → Z1
Output	common 0 or defensible bundle
Service rule	Q* and its independent justification
Q* provenance	anchored / researcher-specified
Q* sensitivity	if researcher-specified and material: plausible alternatives tested + stability/conditional result
Boundary	B-P / B-S / B-X + description
B-S service-function map	if B-S and material functions move/change: function, service coupling, recurrence, actor/location, attribution, decision
Boundary sensitivity	if material: resolution, conditional result, or NS-B
State conditions	material scale, horizon, maturity, version/environment
Hour basis	R1 / R2 / P1 / P2 / P3 / U for each side
Person-time ledger	material components counted once
Provenance	acquisition + state realisation + causal status
Decision	canonical / proxy / zero endpoint / no scalar
Refusal codes	if applicable
Uncertainty	material numerical or specification caveat

8.1 Development tests and reproducibility status

The coding manual did not remain only a checklist. After a hostile external review argued that the proposed protocol was not operational enough to reproduce, its rules were frozen and tested in stages.

Test stage	Result	What it established
Synthetic held-out calibration, two sets	16/16 primary outcomes matched across two blind LLM coding runs	the decision tree could recover pre-specified canonical, proxy, refusal, boundary, stochastic, transition and zero-endpoint branches
First same-evidence real case — Di Matteo et al. (2026)	material disagreement on canonical vs proxy despite agreement on the human-time ledger	exposed under-specification around MTM/work-measurement evidence and state realisation; rules were revised rather than the disagreement hidden
Fresh microbiology case — Foglia et al. (2026)	two coders refused a preregistered scalar as NS-Q	showed that the quality gate could reject the project’s own expected result; motivated the constitutive-quality test
Fresh HbA1c workflow — Karlsson et al. (2025)	same-evidence LLM proxy-branch match: 85/39 ≈ 2.18, 46 min proxy difference	demonstrated same-evidence reproduction of a scalar-capable P-class claim under the frozen revised manual
Fresh Rzeppa assembly case — Calvo & Gil (2022); Cherubini et al. (2016)	Historical source-packet branch agreement: $LF \approx 1.112$ -1.299	sessions agreed on the supplied ledger; later source recheck requires Conditional target-state status and limits this to a standard-time illustration

The testing sequence shows that the rules can be applied to preselected evidence and that disagreement can expose under-specification. The exercises used separate LLM sessions, not independent human domain experts; M19 used separate DeepSeek sessions rather than different model families. The source packets were author-prepared summaries and extracts. In M20 they supplied the component times, local boundary and standby issue, while omitting

the earlier demonstration’s approximately 120-second cycle and implementation details discussed in §9.5. Agreement on that packet does not test independent discovery of all material evidence.

The Rzeppa correction preserves the historical responses as records of what those sessions received and concluded. It does not treat them as validation of the revised interpretation. The 16/16 synthetic result is a descriptive score on deliberately constructed development cases; it is not an accuracy estimate for unseen domains. No inter-rater reliability coefficient is inferred. A stronger future test would freeze the revised manual and have human/domain specialists independently extract and code complete original sources, with disagreements and exclusions retained.

The applications below illustrate useful proxy comparisons, a refusal, and the limits of a conditional engineering-standard calculation.

9. Worked Applications: Canonical, Proxy, or Refused?

The main case set spans one historical anchor, two contemporary AI studies, an explicit refusal, and one conditional manufacturing standard-time illustration. The ratios are not pooled or ranked across technologies; each claim remains tied to its own output, boundary, state and hour basis.

Case	Time evidence	Hour basis	Protocol output	Main qualifier
USDA wheat, 1915–19 → 1979–83	98/7 = 14.0×	historical aggregate unit-labour statistic	Proxy historical human-time ratio	reconstructed national series; no single-technology causal attribution demonstrated/randomised; 12-test service rule only
GitHub Copilot HTTP server	160.89/71.17 ≈ 2.261×	completion-conditioned elapsed task time	Proxy task-time ratio	randomised field study; state-relative necessity and component durations not established
METR experienced OSS developers	1/1.19 ≈ 0.840×	realised active implementation time	Proxy realised-time ratio	milk output and other state dimensions differ
Automatic milking	tempting ≈2× labour/cow-day	output/service mismatch	NO VALID SCALAR — NS-0	Modelled acquisition; target timing configuration
Rzeppa joint assembly	715.8 / [643.5,551.0] TMU	MTM-standard ledger, conditional on its duration assumptions	Conditional standard-time illustration: 1.112-1.299	Conditional; not a verified real-cell required-time bound

9.1 USDA wheat: a historical ratio, not an identified required-time frontier

USDA *Agricultural Statistics 1984*, Table 571, reports **98 labour-hours per 100 bushels of wheat** for 1915–19 and **7 hours per 100 bushels** for 1979–83 (U.S. Department of Agriculture, 1984). The arithmetic ratio is

$$\frac{98}{7} = 14.$$

The fixed physical output unit makes this a useful historical anchor, but the hour basis is an aggregate unit-labour statistic without a component-level demonstration of state-relative required time. The protocol therefore reports a **14× historical human-time proxy**, not a canonical R-class Labourfold. It also does not attribute the change to one machine: machinery, agronomy, yields, organisation, farm structure and other state dimensions co-evolved.

9.2 GitHub Copilot: a completion-time proxy

Peng et al. (2023) randomly assigned 95 professional programmers to implement the same JavaScript HTTP server with or without Copilot. Completion was the first commit passing

all **12 tests**. Among successful completers, mean elapsed task time was **160.89 minutes** without Copilot and **71.17 minutes** with it:

$$\frac{160.89}{71.17} \approx 2.261.$$

The service rule is deliberately narrow: the assignment plus the provided tests. Broader security, performance and maintainability were not established. The time basis is also completion-conditioned elapsed task time, coded P2 rather than verified required person-time. The defensible output is therefore a **2.261× proxy task-time ratio** for this study state, not “Labourfold of programming.”

9.3 METR: below one remains below one

Becker et al. (2025) studied 16 experienced open-source developers working on 246 real repository issues. Issues were defined before treatment assignment and randomised to AI-allowed or AI-disallowed conditions. Developers tracked active implementation time, excluding passive waiting for review, with screen recordings used for validation and analysis.

The reported main estimate was that AI-allowed issues took **19% longer**. Expressed in the comparator direction used here:

$$\frac{H_{noAI}^{realised}}{H_{AI}^{realised}} \approx \frac{1}{1.19} \approx 0.840.$$

The protocol preserves the below-one result rather than treating it as an inconvenient exception. Active realised implementation time is P1: strong evidence about what developers actually spent, but not a component-level identification of required time under the declared state. The result is a **0.840× realised-time proxy ratio** for this worker/task/tool state.

9.4 Automatic milking: refusal is coded

Oudshoorn et al. (2012) compared nine Danish organic farms using automatic milking systems with nine using conventional systems. Labour per cow-day was about half under automatic milking, making a scalar near 2 tempting. The protocol refuses it for the requested fixed-output human-time claim.

The published denominator is labour **per cow-day**, not labour per fixed milk output. Milk yield differs materially—**9021 kg ECM per cow-year** under AMS versus **7664 kg** under CMS—and grazing and other production-state indicators also differ. The publication does not provide the decomposition needed to standardise both systems to one common output/service bundle without additional modelling. The primary result is therefore:

NO VALID SCALAR — NS-O.

Additional quality or state codes may become relevant under a differently specified claim, but the output mismatch already blocks this scalar as published.

9.5 Rzeppa assembly: a conditional standard-time illustration

Calvo and Gil (2022) analyse manual and collaborative assembly of a Rzeppa homokinetic joint using Methods-Time Measurement. Their table supports a standard-time comparison. Cherubini et al. (2016) separately demonstrate physical collaborative assembly. Those two sources must not be combined as though the particular timing configuration in the 2022 model had been physically verified by the 2016 experiment.

The declared object is one correctly completed joint assembly in a local assembly-cell boundary, excluding downstream drivetrain endurance and whole-factory effects. Conditional on the MTM method, pace and duration assumptions being applicable, the manual ledger is:

$$H_{0,\text{std}} = 715.8 \text{ TMU} = 25.7688 \text{ s.}$$

The collaborative human components in Table 3 sum to:

$$262.7 + 23.2 + 5(47.8) + 26.1 = 551.0 \text{ TMU} = 19.836 \text{ s.}$$

The modelled complete cycle totals 643.5 TMU, or 23.166 seconds. Counting none or all of its remaining 92.5 TMU as required human standby produces this scenario envelope:

$$H_{1,\text{std}} \in [551.0, 643.5] \text{ TMU.}$$

Its implied ratios and differences are:

$$\frac{715.8}{643.5} \approx 1.112, \quad \frac{715.8}{551.0} \approx 1.299,$$

$$S_{\text{std}} \in [2.6028, 5.9328] \text{ human-seconds per joint.}$$

These numbers are a **conditional standard-time illustration**, not measured savings or a confidence interval. The interval varies only the standby treatment while holding the standard-time ledger fixed. It does not bound all uncertainty in pace, interruptions, recovery, allowances or applicability of the model to a physical cell. A prose value of 653.5 TMU conflicts with the source's table total and its 23.166-second conversion; the table-consistent 643.5 is used with that discrepancy disclosed.

Cherubini et al. (2016, §8.3) report approximately 120 seconds for the illustrated physical assembly, operation below 0.1 m/s, keyboard notification to resume robot motion, and delays including a dropped ball. The 2022 analysis uses a different timing construction, describing robot motion at a human-like speed subject to a 0.25 m/s ceiling. The earlier study demonstrates physical feasibility of collaborative assembly, but does not verify the later model's exact duration, recovery assumptions or full human-time ledger. Conversely, its 120-second elapsed cycle cannot simply replace required human time in the denominator.

The time-estimate acquisition is therefore **Modelled**, and the target configuration at the claimed standard-time performance is **Conditional**. The existence of a related demonstrated cell is supporting feasibility evidence, not a transfer of Demonstrated status to the timing claim. An empirical R1/R2 classification requires a documented match between the timing model and the actual declared state, with material human actions and duration uncertainties included. Until then, the interval above remains an explicitly conditional illustration rather than a verified bound on required time in the demonstrated cell.

The following filled claim card shows how the protocol changes the interpretation of the same arithmetic:

Claim-card field	Application to this evidence
Question	What person-time comparison follows from the declared MTM ledger, and what is established for a physical cell?
Output and quality	One correctly completed assembly; narrow assembly admissibility, without downstream endurance claims
Boundary	B-P, the stated assembly operation; no factory-wide or service-system saving claimed
Time evidence	715.8 TMU manual; 551.0 TMU human components in the collaborative model
Unresolved component	Required availability during the model's 92.5 TMU remainder
Duration qualification	Standard-time applicability and material interruptions/allowances require verification for empirical use
Acquisition / target realisation	Modelled / Conditional at the claimed timing configuration
Permitted statement	Conditional standard-time ratio 1.112-1.299, with the assumptions above
Unsupported statement	The demonstrated cell has been shown to release 2.60-5.93 human-seconds per joint
Evidence needed	Matched configuration, complete human-time observations or justified standards, coverage requirements and comparable service quality

M20's two LLM sessions agreed on the original prepared packet and its R1/R2 interpretation. That result is retained as historical packet-conditioned agreement. The packet did not expose the full timing discrepancy. No new coding run has tested this revised claim card, and the original agreement is not an independent validation of it.

9.6 What the cases and tests establish

The examples show that time ratios can remain useful with explicit proxy labels, that a below-one finding need not be discarded, and that output mismatch can justify refusing a scalar. The Rzeppa example additionally shows why a mathematically correct engineering interval must remain conditional when evidence for state realisation and duration applicability is incomplete.

The development exercises test selected rule applications and reveal opportunities for repair. They do not establish human-expert agreement, independent source-extraction reliability, broad external validity or universal identification of required time. This revision itself illustrates that agreement among LLM sessions can coexist with a material omission in their shared source packet. Future validation should test the revised manual against complete sources and preserve coding disagreement before adjudication.

10. What Labourfold Does Not Mean

Suppose a comparison supports $LF = 5$. The target state requires one-fifth as much human person-time for the declared output, quality, boundary, state and horizon. That is useful, but it is much narrower than the interpretations a large number can invite.

Labourfold is **not labour cost**. Ten generalist hours at €20 replaced by three specialist hours at €80 gives $LF \approx 3.33$ while direct labour cost rises from €200 to €240. Human time and its price are different objects. It is also **not total resource efficiency**: a process can save human hours while using more electricity, compute, capital, water or materials. Those require separate ledgers.

It is **not speed, workload, welfare, employment, or causal attribution**. Human time can fall while lead time rises because autonomous runs or queues sit between interventions; five intense hours can replace ten moderate hours without making LF a welfare index; and a fixed-output state can require fewer hours while output growth, new tasks or institutional choices leave aggregate employment unchanged or higher. A historical state comparison also does not identify how much of the change came from machinery, organisation, inputs, or their interaction.

Finally, Labourfold does not determine who receives the technical release. A productivity gain contains no rule saying whether released capacity becomes lower prices, higher profits, higher wages, shorter schedules, more output, redundancy, new tasks, or public capacity.

Labourfold is a human-time requirement comparison, not a universal efficiency, welfare, employment or distribution index.

11. Saved Hours and the Allocation Boundary

The multiplicative ratio has an absolute counterpart:

$$S = H_0^{req}(O, Q^*) - H_1^{req}(O, Q^*).$$

For canonical R-class evidence, S is **Saved Hours**: technical fixed-output release of required human person-time. For P-class evidence, the corresponding difference is reported as a proxy difference rather than silently upgraded.

If a defined output falls from 100 required human hours to 20,

$$LF = 5, \quad S = 80h.$$

That does not mean society works 80 hours less. If output expands fivefold while the per-unit requirement falls fivefold, aggregate required labour can remain unchanged. Demand, new tasks, quality expansion, induced activity and institutional choice sit downstream of the fixed-output measurement.

Nor does $S = 80h$ imply that every incumbent can simply receive eight hours back. The same technical release can support very different feasible staffing patterns:

$$10 \times 40h \rightarrow 10 \times 32h$$

or

$$10 \times 40h \rightarrow 8 \times 40h.$$

Both use 320 hours after a 400-hour baseline, but the feasible human outcomes differ because of coverage, role indivisibility, skills, scheduling and institutional constraints. Labourfold therefore stops before a full allocation theory. It establishes the technical human-time change; allocation feasibility and actual allocation are separate empirical and normative objects.

The framework remains technology-neutral. Tractors, refrigeration, robot arms, software, AI assistants, organisational redesign and other production changes can all be asked the same prior question: **what happened to the human person-time required by the declared service, and what evidence supports that claim?**

12. Return to the Equation

The paper began with a ratio that looked almost too simple to deserve a name:

$$LF = \frac{H_0}{H_1}.$$

The arithmetic was never the difficult part. The difficulty was deciding when the two numbers describe a legitimate human-time comparison.

The states must refer to the same declared output and a common material service/admissibility rule. The hour basis must remain visible rather than silently substituting paid, scheduled, elapsed or observed time for required person-time. Shared time must be counted once; the human boundary must be explicit; material differences in workflow, maturity, case mix, version, scale and temporal window must be exposed; constitutive quality cannot be replaced by mere process completion; researcher-specified quality rules must be sensitivity-tested when plausible alternatives matter; and service-system claims must follow material human functions across actor or vendor relocation without recursively absorbing general-purpose upstream infrastructure.

After those constraints, the canonical expression remains:

$$LF = \frac{H_0^{req}(O, Q^*)}{H_1^{req}(O, Q^*)}.$$

Its persistence is not evidence of mathematical novelty. Labourfold is a comparative form of established unit-labour/productivity arithmetic. The contribution proposed here is the **integrated claim architecture around it**: a codable way to state, audit, narrow, sensitivity-test

or refuse comparative human-time claims and then hand the absolute difference to a separate allocation question.

The protocol has undergone limited adversarial development testing. Two synthetic sets yielded 16/16 pre-specified branch matches. Prepared real-source packets produced disagreement and repair, a quality-based refusal, and later agreement on a proxy case and an assembly ledger. Earlier reviews revised the infimum definition, quality sensitivity and service-system boundary rules. The present source recheck qualifies the assembly example because a demonstrated physical cell and a modelled timing configuration had been linked too closely. Historical agreement is preserved without presenting it as validation of the revised interpretation. These exercises support development and rule-application checks; independent extraction from full sources and human/domain-expert validation remain open.

The range of protocol outputs is intentionally broader than a scalar. Compatible positive R-class evidence can support canonical LF; P-class evidence supports a labelled proxy ratio; a zero target requirement is reported directly rather than as infinity; and materially unresolved comparison failures return **no valid scalar** with a refusal code.

The absolute counterpart,

$$S = H_0^{req}(O, Q^*) - H_1^{req}(O, Q^*),$$

names technical fixed-output human-time release, not its destination.

The empirical examples span early twentieth-century agriculture, contemporary software/AI studies, laboratory workflows and human-robot assembly. They are not offered as a comprehensive history of productivity. They demonstrate that the same prior question can be asked across very different technologies: **how much human labour did a defined output require in one production state, how much does it require in another, and what evidence justifies saying that the difference is real?**

That is the object measured here: **the human labour that output no longer requires.**

Research-process disclosure

This manuscript was developed with intensive assistance from large language models used as research and drafting tools. LLM assistance was used for adversarial stress-testing of definitions, generation and checking of counterexamples, mathematical and accounting checks, source triage, manuscript-structure iteration and prose drafting. The author selected the research question, accepted or rejected conceptual revisions, determined the definitions and evidentiary boundaries, chose the retained empirical claims, verified the final source set, and retains responsibility for the argument and final text. This revision was prepared with coding-agent assistance and a targeted original-source recheck; that check exposed the assembly-state qualification described in §9.5 and is not a complete audit of every reference.

The protocol-testing sequence used separate LLM sessions under frozen manuals, author-prepared source packets and records described in the archive as preregistered. Session separation does not establish independent source selection or statistical independence. These exercises are reported as synthetic calibration and packet-conditioned rule-application agreement. The revision does not establish an externally timestamped preregistration or human-expert inter-rater validation. No new LLM coding exercise was performed for this revision. Manuscript-production speed or LLM output volume is not used as Labourfold evidence; any claim about AI-assisted research time would require its own comparator, output/service rule, human-time ledger and evidence protocol.

References

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2), 3–30. <https://doi.org/10.1257/jep.33.2.3>

- Ahn, J. S., Ebrahimian, S., McDermott, S., et al. (2022). Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. *JAMA Network Open*, 5(8), e2229289. <https://doi.org/10.1001/jamanetworkopen.2022.29289>
- Becker, J., Rush, N., Barnes, B., & Rein, D. (2025). *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*. METR. https://metr.org/Early_2025_AI_Experienced_OS_Devs_Study-paper.pdf
- Calvo, R., & Gil, P. (2022). Evaluation of collaborative robot sustainable integration in manufacturing assembly by using process time savings. *Materials*, 15(2), 611. <https://doi.org/10.3390/ma15020611>
- Cherubini, A., Passama, R., Crosnier, A., Lasnier, A., & Fraisse, P. (2016). Collaborative manufacturing with physical human-robot interaction. *Robotics and Computer-Integrated Manufacturing*, 40, 1–13. <https://doi.org/10.1016/j.rcim.2015.12.007>
- Cieplinski, A., D'Alessandro, S., & Guarnieri, P. (2021). Environmental impacts of productivity-led working time reduction. *Ecological Economics*, 179, 106822. <https://doi.org/10.1016/j.ecolecon.2020.106822>
- Di Matteo, A. S., Capponi, M., Mastrogiacomo, L., & Franceschini, F. (2026). Productivity assessment in multi-human multi-robot manufacturing teams: insights and proposal for novel KPIs. *The International Journal of Advanced Manufacturing Technology*, 144, 617–639. <https://doi.org/10.1007/s00170-026-17947-4>
- Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B*, 16(2), 175–185. <https://doi.org/10.1111/j.2517-6161.1954.tb00159.x>
- Foglia, E., Garagiola, E., Ferrario, L., Adorisio, E., Gagliardi, G., & Lenoci, F. D. (2026). Total laboratory automation for microbiology activities: An HTA analysis in an Italian laboratory setting. *ClinicoEconomics and Outcomes Research*, 18. <https://doi.org/10.2147/CEOR.S597703>
- Godé, L., Mair, S., & Gómez-Baggethun, E. (2026). Labour productivity gains or offshoring? Implications for post-growth proposals on the future of work. *Ecological Economics*, 239, 108778. <https://doi.org/10.1016/j.ecolecon.2025.108778>
- Institute of Industrial and Systems Engineers (IISE). (n.d.). *Z94.17 Work Design and Measurement terminology*. <https://www.iise.org/Details.aspx?id=2604>
- International Organization for Standardization (ISO). (2006). *ISO 14040: Environmental management—Life cycle assessment—Principles and framework*.
- Jerome, H. (1934). *Mechanization in Industry*. National Bureau of Economic Research. <https://www.nber.org/books-and-chapters/mechanization-industry>
- Life Cycle Initiative. (2020). *Guidelines for Social Life Cycle Assessment of Products and Organizations 2020*. United Nations Environment Programme.
- Karlsson, M., Garpefjord, B., Ström, M., Fredolini, C., & von Horn, H. (2025). B-029 Development of an automated laboratory process for quantitative dried blood microsampling to improve core laboratory workflow. *Clinical Chemistry*, 71(Supplement 1), hvaf086.427. <https://doi.org/10.1093/clinchem/hvaf086.427>
- OECD. (2001). *Measuring Productivity—OECD Manual: Measurement of Aggregate and Industry-level Productivity Growth*. OECD Publishing. <https://doi.org/10.1787/9789264194519-en>
- Oudshoorn, F. W., Kristensen, T., van der Zijpp, A. J., & de Boer, I. J. M. (2012). Sustainability evaluation of automatic and conventional milking systems on organic dairy farms in Denmark. *NJAS - Wageningen Journal of Life Sciences*, 59(1–2), 25–33. <https://doi.org/10.1016/j.njas.2011.05.003>
- Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*. arXiv:2302.06590. <https://arxiv.org/abs/2302.06590>

- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4, 1208–1214. <https://doi.org/10.1038/s41562-020-0912-z>
- Springer, S. K., Wulf, C., & Zapp, P. (2025). PSILCA database for social life cycle assessment: worker hours vs. raw values approach. *The International Journal of Life Cycle Assessment*, 30, 1078–1098. <https://doi.org/10.1007/s11367-024-02348-0>
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- UNEP/SETAC. (2009). *Guidelines for Social Life Cycle Assessment of Products*. United Nations Environment Programme / Society of Environmental Toxicology and Chemistry.
- U.S. Bureau of Labor Statistics. (1977). *Productivity and the Economy*. Bulletin 1926. U.S. Department of Labor. <https://fraser.stlouisfed.org/title/productivity-economy-4824/fulltext>
- U.S. Department of Agriculture. (1984). *Agricultural Statistics 1984*, Table 571, “Selected crops: Labor-hours per unit of production and related factors, United States, indicated periods, 1915–83.” <https://esmis.nal.usda.gov/sites/default/release-files/j3860694x/c247dw13d/1c18dj64x/Agstat-04-23-1984.pdf>

Technical appendix

This appendix preserves the operational rules and edge cases behind the compressed manuscript. The main text presents only the minimum claim architecture; the fuller gate catalogue remains an audit trail rather than mandatory syntax.

A1. Canonical objects and protocol outputs

For a common output/service object 0 and common material service/admissibility rule Q*:

$$LF = \frac{H_0^{req}(O, Q^*)}{H_1^{req}(O, Q^*)}, \quad S = H_0^{req}(O, Q^*) - H_1^{req}(O, Q^*).$$

H^{req} is state-relative required human person-time within a declared boundary. LF is reported as canonical only when the comparison passes the output, quality, boundary, state/horizon and human-time-basis tests. The protocol can instead return:

1. **Canonical Labourfold** — compatible R-class required-time bases;
2. **Proxy human-time ratio** — comparison is otherwise admissible but at least one side is P-class;
3. **NO VALID SCALAR** — a material comparison condition remains unresolved;
4. **Zero-human-time endpoint** — $H_1=0$; report zero target requirement and technical release rather than infinite LF.

For $H_1>0$, calculation follows only after classification. The protocol therefore treats division as the last step, not the first.

A2. Required human person-time

For a declared production state Z, boundary B, output/service object 0, quality rule Q*, and material horizon τ , let $C(Z, B, 0, Q^*, \tau)$ denote the material human-time components required by that declared state. Conceptually:

$$H^{req}(O, Q^* \mid Z, B, \tau) = \sum_{c \in C(Z, B, 0, Q^*, \tau)} \tau_c^{req}.$$

Every included component must pass the necessity test. For a canonical claim it must be supported on an R-class basis. This is not observed elapsed/scheduled time by default and is not a minimum across every alternative workflow. MTM/PMTS can support R1/R2 for the declared method/state without identifying a global lower envelope.

Hour-basis classes

- **R1 — verified required.** Supported by a removal/coverage test, binding process dependency, formal legal/safety/contractual obligation, auditable work-measurement standard such as MTM/PMTS where appropriate, or repeated operational failure below the requirement.
- **R2 — bounded required.** Required time is identified only within defensible lower/upper bounds. A common example is an R1 active-task ledger plus an unresolved standby/availability interval.
- **P1 — realised engaged-time proxy.** Actual human engagement is measured but necessity under the declared state/specification is not established.
- **P2 — elapsed task-time proxy.** A timer mixes activity with possible waiting/idle and does not identify required person-time.
- **P3 — scheduled/paid/staffing proxy.** Roster or payroll time is available without evidence that all of it is required.
- **U — under-identified.** The record cannot support a sufficiently commensurable human-time basis for the requested scalar.

An R-class hour basis and evidence acquisition are separate. A modelled MTM standard can be R1/R2; a directly observed elapsed timer can remain P2. Necessity of an action does not identify its duration. R-class standard-time estimates require declared method/pace assumptions, applicability to worker and operating conditions, treatment of allowances and interruptions, and material uncertainty. Model-based point estimates remain conditional on those assumptions; the MTM label alone is insufficient.

Standby, concurrency and single-entry accounting

Required availability counts when the declared service requires a person to remain encumbered and ready. Genuinely free machine/agent runtime does not. Overlapping task labels are counted as the union of required intervals for each person; one person-hour cannot become several merely because it supports multiple labelled tasks. Two people required simultaneously contribute two person-hours.

Recurring maintenance, calibration, incident response, recertification, mandatory audit, safety sign-off and direct orchestration belong in operating H when materially required. One-off commissioning, migration, integration and adoption training are transition labour and enter only an explicitly cumulative/extended claim.

A3. Output and constitutive quality

0 may be a scalar unit, functional unit, task, standardised case mix or output bundle/vector. Equal counts do not establish equivalent output when product mix or service content differs. Price/revenue should not be used merely to force unlike outputs into one physical/service comparison.

Both states must satisfy a common Q^* . Permitted modes include:

- **Q-A:** common admissibility floor;
- **Q-E:** equivalence/non-inferiority band when the question requires it;
- **Q-V:** multidimensional service vector.

The target may outperform the comparator once both are inside the same admissible service set. Record Q^* provenance as **anchored** or **researcher-specified** where material. Researcher-specified thresholds, equivalence bands, weights or admissibility mappings require sensitivity to materially plausible alternatives under the same frozen research question and output/service object. If such alternatives change scalar eligibility or materially change required time/Saved Hours, one unqualified scalar is prohibited.

Constitutive-quality identity test

Before accepting Q^* , ask whether omitting a quality dimension would make the output cease to be the same functional service even if the workflow reaches the same completion step. Diagnostic correctness, safety-critical failure probability, structural integrity, legal validity, preservation of a material biological/chemical property, and identity/authenticity can be constitutive. If such a dimension remains materially unresolved, standardise/model it under a defensible service rule or return NS-Q.

Human verification, correction or assurance required to keep the target inside Q^* belongs in H1.

A4. Human-labour boundary

The boundary is part of the estimand:

- **B-P:** process/site/organisational boundary;
- **B-S:** service-system human boundary; default for an unqualified claim of human-time release;
- **B-X:** explicitly extended/upstream boundary.

B-S follows the **service function**, not the employer, vendor, customer, site or technical layer. A material function belongs in B-S when it operates/assures/restores/resets/maintains O/Q*, is recurrently relevant over the horizon, and can be attributed to the service.

For materially changed, shifted, introduced or removed functions, the claim record maps:

1. function;
2. service coupling;
3. recurrence/horizon relevance;
4. actor/location in each state;
5. attribution rule;
6. boundary decision;
7. materiality consequence if uncertain.

Actor relocation does not erase the function. A shared B-S function that cannot be materially allocated uses the existing NS-J / bundle treatment rather than being silently excluded.

Ordinary B-S stops before general-purpose upstream infrastructure whose human operation is not materially service-specific at the declared claim level. Causal necessity alone is insufficient. Electricity-system labour, generic backbone/network operation, general-purpose datacentre/facility operation, standard equipment manufacture and generic building services are B-X unless the claim deliberately extends upstream.

If plausible B-S mappings under the same frozen claim materially change LF direction, either state's total, scalar class or Saved-Hours interpretation, resolve from evidence, report boundary-conditional results, or return NS-B for an unqualified claim.

A5. Production state, scale and temporal closure

Labourfold belongs to production states, not intrinsic technology names. Where useful:

$$Z = (V, D, \Omega, W, M, \dots),$$

covering technology/version/configuration, workload/case distribution, operating environment, worker population and maturity/exposure. Different states do not themselves trigger refusal; every comparison contains differences. NS-Z is reserved for differences that make the requested estimand non-comparable or non-transportable and cannot be repaired by narrowing, conditioning or standardisation. Concurrent changes can block technology-only causality without blocking a descriptive state comparison.

With fixed human overhead:

$$H_i(O) = F_i + m_i O, \quad LF(O) = \frac{F_0 + m_0 O}{F_1 + m_1 O}.$$

Scale, batch size, cadence and maturity should therefore be exposed when material. For stochastic systems, aggregate or estimate compatible expected required person-time before division:

$$LF^E = \frac{\mathbb{E}[H_0^{req}]}{\mathbb{E}[H_1^{req}]}.$$

Across recurring temporal regimes:

$$LF^\tau = \frac{\sum_r H_{0,r}}{\sum_r H_{1,r}}.$$

Do not average LF ratios. Required preprocessing, inventory creation, WIP, deferred verification and maintenance liabilities must be closed symmetrically so released time cannot be manufactured by moving work outside the measurement window.

A6. Evidence provenance, uncertainty and refusal

Evidence provenance uses orthogonal axes:

- **Acquisition:** Direct / Reconstructed / Modelled;
- **State realisation:** Operational / Demonstrated / Conditional;
- **Causal status:** Descriptive / Randomised or otherwise credible / Quasi-experimental / unresolved.

A related physical demonstration cannot establish Demonstrated status for a different modelled timing configuration without a documented state match. When that bridge is missing, the claimed target remains Conditional. These axes are not a validity score. A direct P1 observation is not automatically canonical, and a modelled work-measurement standard is not automatically a proxy.

Refusal codes make non-scalar decisions auditable:

- NS-0 output mismatch;
- NS-Q quality/service mismatch;
- NS-B boundary mismatch;
- NS-H incompatible/under-identified human-time bases;
- NS-Z irreparable state non-comparability;
- NS-T temporal closure failure;
- NS-J unresolved joint/shared labour;
- NS-F infeasible requested target state;
- NS-U residual material under-identification.

A refusal stops the scalar, not the audit record: all still-inferable claim fields should remain documented.

Numerical uncertainty is separate from evidence provenance. Ratio inference should preserve numerator/denominator uncertainty and covariance; a small denominator can make ordinary confidence intervals unstable (Fieller, 1954). Specification sensitivity should distinguish alternative operationalisations of one claim from genuinely different estimands such as different boundaries or comparators.

A7. Aggregation, chaining and interpretation

Across compatible components:

$$LF_{bundle} = \frac{\sum_i H_{0,i}}{\sum_i H_{1,i}}.$$

Do not average unrelated LF ratios into an “average Labourfold of AI” or another heterogeneous technology class.

For genuinely compatible sequential states:

$$LF_{0 \rightarrow n} = \prod_{k=0}^{n-1} LF_{k \rightarrow k+1},$$

while absolute required-time differences add. Cancellation requires the same intermediate human-time object; chaining does not assign causal credit to individual technologies.

LF is not labour cost, lead time, total resource efficiency, workload/welfare, employment, causal contribution or distributive justice. Those require parallel ledgers or downstream analysis.

A8. Minimum claim card

Each independently coded claim should preserve:

1. research question and $Z_0 \rightarrow Z_1$;
2. 0;
3. Q mode and explicit Q*;
4. Q* provenance — anchored / researcher-specified;
5. if researcher-specified and material, the Q* sensitivity branch/result;
6. B-P/B-S/B-X boundary;
7. if B-S and material functions changed/shifted/appeared/disappeared, the service-function map;
8. if boundary classification is materially uncertain, the boundary-sensitivity resolution/conditional result/NS-B;
9. material scale/horizon/maturity/version conditions;
10. H0/H1 basis (R1/R2/P1/P2/P3/U);
11. component person-time ledger;
12. acquisition, state-realisation and causal axes;
13. active refusal/uncertainty/sensitivity issues;
14. output class and numeric result only where permitted.

Conditional Q* and B-S fields are mandatory when triggered; they are not optional merely because the core claim card is compact.

A9. Reproducibility and validation record

The protocol's operationality was tested after external criticism that a claim card was not yet a reproducible protocol.

- **M15-M16 synthetic calibration:** two blind LLM coding runs on two separate held-out sets recovered all **16/16 pre-specified primary outcomes**, spanning canonical scalar, R2 bounds, proxies, refusals, standby, boundary transfer, stochasticity, transition labour and the zero endpoint.
- **M17 same real evidence:** two coders recovered the same human-time ledger and numeric envelope but disagreed on canonical versus proxy classification. The disagreement was preserved and motivated Manual v1.1, clarifying MTM/work-measurement, R/P classes and state realisation.
- **M18 fresh real evidence:** two coders independently rejected the project's preregistered microbiology scalar as NS-Q, identifying diagnostic correctness as a missing constitutive quality dimension. The preregistration failed; the quality gate succeeded. This motivated Manual v1.2.
- **M19 frozen v1.2:** two independent DeepSeek sessions produced the same-evidence LLM proxy-branch match for the HbA1c workflow: 85 versus 39 reported hands-on minutes per 96-sample batch, ratio ≈ 2.18 and 46-minute proxy difference. Minor P1/P2 and provenance-label differences did not change the output class or magnitude.
- **M20 frozen v1.2 (historical):** two separate LLM sessions agreed on the supplied Rzeppa source packet, its R1/R2 classification and interval 1.112–1.299. Subsequent original-source checking identified an incomplete bridge between the 2022 standard-time model and the 2016 physical demonstration. The latter reports approximately 120 seconds and implementation conditions not reproduced in the former's 23.166-second model. The numerical standard-time illustration remains conditional; the original coding agreement does not validate that revised interpretation.

These exercises assess rule application to curated evidence, not independent source extraction or population-level reliability. No new coding run or reliability coefficient is supplied with this revision. Human/domain-specialist validation on complete original sources remains open.

A10. Version status

Archived Manual v1.5 and historical coding responses remain frozen records. The accompanying v1.6 manual strengthens duration-evidence and state-matching requirements and has not been independently validated. Revised manuscript §9.5 gives a worked conditional claim card. Earlier PASS labels in the archive do not certify this revision.

A11. Post-M22 repair record

External review after M21 identified three material protocol issues: an over-strong infimum definition, residual researcher-specified Q^* discretion, and an ambiguous B-S stopping rule. M23–M28 adjudicated and regression-tested these, producing Manual v1.5. M29 narrowed novelty claims to integration/codification; M30 narrowed the subtitle and allocation scope. These gates add protocol stress testing, not human-expert validation.