Research article

MechEcology 2026;5(1):e9
https://doi.org/10.23104/ME.e9

MechEcology

eISSN 2799-9297

The Constant and the Variable: An Ontological Evolution of Human–AI Interaction

Hyunghun Kim*

The MechEcology Academia, Seoul, Korea

Published: August 31, 2026

***Corresponding author**
Hyunghun Kim
The MechEcology Academia, 150, Mokdongdong-ro, Yangcheon-gu, Seoul 08014, Korea
Tel: +82-2-2643-2323
E-mail: relyonlord@gmail.com

ORCID
Hyunghun Kim
https://orcid.org/0000-0002-3013-075X

Competing interests
No potential conflict of interest relevant to this article was reported.

Funding sources
Not applicable.

Abstract

This paper develops a formal framework that connects two levels, the single episode and the institution, in order to describe a structural shift in human-AI collaboration. As AI systems lower the marginal cost of execution, the binding constraint on cognitive production moves upstream, from doing the work to specifying it clearly enough to be used under institutional conditions. At the micro level, we hold the decoding system fixed and study the user side. We separate what a specification is from what producing it costs. The completeness of a specification, its amplitude, is a property of the specification itself, defined independently of the environment. To be admissible, a specification must reach a minimum amplitude, a fixed decoding bar set by the system and unchanged by the environment. The central separation at this level is that the bar stays the same while the minimum cost of reaching it falls as instruction-environment quality improves. Evaluation is then organized by a three-gate structure. The first gate, specification admissibility, requires the specification to reach the minimum amplitude needed to be reliably decoded. The second gate, directional admissibility, defines the evaluable domain through the direction the user declares *ex ante*. The third gate, institutional quality sufficiency, applies acceptance standards on non-directional quality within that domain. Two limits fix what the environment and scale cannot do. Because the decoding bar is fixed, no improvement in instruction-environment quality removes the minimum amplitude the user must supply. Because direction is set by the user *ex ante*, no amount of execution scale substitutes for it. At the institutional level, we represent each institution's instruction support by the distribution of instruction-environment quality across its episodes. A comparative-statics bridge then carries an upward shift in this distribution, in the sense of first-order stochastic dominance, down to the micro level. An improvement in the environment can widen participation only by lowering cost, never by lowering the bar. Together, these locate human responsibility for the goal and for the minimum specification in the user, under a normative premise the paper states openly, even as execution scales and environments improve. We outline measurement and audit strategies for the core constructs and anchor the framework in two contrasting case observations, regulated clinical AI deployment and automated welfare determination.

Keywords: human-AI symbiosis; AI governance; instruction environment; institutional design; directional admissibility

Acknowledgements

This manuscript was edited for language and style with the help of GPT-5.5 (OpenAI), Claude Opus 4.8 (Anthropic), and Claude Fable 5 (Anthropic), which are large language models. These tools were also used as learning aids. They helped the author study the mathematical concepts behind the formulations presented in this paper and cross-check the steps of the derivations. In addition, the tools were used to check the internal consistency of the manuscript, including the consistent use of terms and symbols and the match between cross-references and the sections they point to. All mathematical expressions, analytical results, and the outcomes of these consistency checks were verified independently by the author, who takes full responsibility for the scientific validity and accuracy of the final content.

Availability of data and material

Not applicable.

Authors' contributions

The article is prepared by a single author.

Ethics approval

Not applicable.

1. Introduction: From Computation to Formulation

Recent work suggests that large language models (LLMs) could substantially reduce the time required for a meaningful share of work tasks, although this evidence concerns technical exposure rather than realized adoption or productivity effects (Eloundou et al., 2024). This paper therefore treats the decline in downstream execution cost as a motivating premise, not as a result derived from that study. Once a task is specified well enough, the premise is that downstream execution becomes cheaper and upstream formulation becomes more consequential. This premise is consistent with long-standing visions of human-computer symbiosis, in which humans set goals, formulate hypotheses, determine criteria, and evaluate results, while computers perform routinizable work (Licklider, 1960).

Research on AI-advised decision making shows that higher model accuracy is not sufficient for higher team performance, because outcomes also depend on how people understand and use the system (Bansal et al., 2019). Empirical reviews further show that human-AI decision making must be studied through the decision task, the form of AI assistance, and the evaluation used, rather than inferred from model capability alone (Lai et al., 2023). In data science, practitioners likewise describe a future in which automation and human expertise remain jointly necessary (Wang et al., 2019). Design guidance emphasizes understandable interaction, user control, explanation, and attention to social context (Amershi et al., 2019; Shneiderman, 2020). Human-AI teaming research documents that work increasingly requires people to collaborate with AI as well as with each other, and argues for a shift from technology-driven questions to a human-centered design agenda (Berretta et al., 2023). At the institutional level, existing audit practices do not yet resolve the governance challenges raised by LLMs (Mökander et al., 2024). Machine-behavior research therefore argues that AI systems should be studied together with the social environments, incentives, and constraints in which they operate (Rahwan et al., 2019).

Against this background, this paper takes an explicitly ontological stance on human-AI collaboration. Here, *ontological* is used in an operational sense. Rather than focusing only on workflow optimization or interface design, the paper sets out the kinds of entities, roles, and structural relations that make up a joint human-AI system of action. It then asks what positions humans and AI hold within it. In the same operational spirit, *teleology* refers throughout to the setting of goals, evaluative criteria, and normative constraints that fix what counts as success in a given episode. It does not refer to anything in a broader metaphysical sense. This operational reading is consistent with Floridi's account of current generative AI as agency without intelligence (Floridi, 2023). The stronger claim adopted here, that the system does not fix the governing goals, criteria, or constraints of the episode, is this paper's modeling proposal rather than a result established by Floridi. In this operational sense, teleology is supplied through human specification, and responsibility is anchored in the human declaration of goals and constraints.

To make this distinction operational rather than just a matter of words, the paper uses a deliberately mathematical style of analysis. The mathematical expressions are not meant mainly as tools for numerical prediction or optimization. They are conceptual devices.

They fix boundaries, hold distinctions steady, and make clear how roles, constraints, and responsibilities are assigned. They separate intent, execution, and evaluation, and treat each as an analytically distinct part of the interaction. By making these separations explicit, the formalism shows where control can be exercised, where ambiguity can enter, and where failure can arise.

This shift in attention upstream raises a question the diagnosis itself leaves open: how should the upstream burden be modeled? We separate two things that are easy to conflate. One is what a specification is, how complete and well-formed it is, which we will call its amplitude. The other is what producing such a specification costs the user in effort. These are distinct axes. A more complete specification is not the same as a more costly one, and the cost of producing the same specification can differ with the instruction-environment quality.

Throughout the paper, we hold the decoding system fixed and ask what the user faces. The parser, the validation rules, and the interpretation layer that turn a specification into usable intent are treated as given, not as objects that improve within the analysis. Holding them fixed has a direct consequence: the minimum amplitude a specification must reach to be reliably decoded, the decoding bar, is a fixed positive constant. The variation this paper studies is therefore not the system becoming more capable, but the instruction environment around the user becoming better, through templates, guided fields, examples, and clarification. All of these leave the bar untouched.

What, then, does a better instruction environment do? It lowers the cost of reaching the bar. The same minimum amplitude can be produced at lower cost when the environment is better. Myers et al. (2000) recommend tools with a low threshold and a high ceiling, where threshold means the difficulty of learning to use a tool. The present framework separates that pair into two objects. What a better environment lowers is the cost of reaching the bar, and what it does not move is the bar itself. In this reading the decoding bar is not Myers' threshold. Their threshold is a cost of reaching that bar, and the bar is what the cost is paid to reach.

Three commitments organize what follows. First, evaluation runs through a three-gate structure: the user's declared direction, not execution scale, fixes what kind of outcome counts as success. Second, each institution's instruction support is represented by the distribution of instruction-environment quality across its episodes. Third, a comparative-statics bridge connects shifts in that quality to the cost the user faces at the micro level. The paper develops these in turn and examines their implications for governance, responsibility, and the design of human-AI joint systems. Appendix A consolidates notation, definitions, and assumptions, and maps the object system in Appendix Fig. A1.

2. Specifications, Amplitude, and Cost

We model a single human-AI interaction and place the variation on the human side. Throughout, we hold the decoding system fixed. By the decoding system we mean the parser, the validation rules, and the interpretation layer that turn a specification into usable intent. We do not study the case where the system itself becomes more capable; that is a

separate question. Holding the system fixed does not freeze the user's surroundings. The instruction environment in which a user produces a specification can still vary: the templates, guided fields, examples, and clarification available to them. Much of our analysis turns on this variation. Fixing the system has one consequence that the rest of the paper rests on. The standard a specification must meet to be reliably decoded is set by the system, and so it does not change within the analysis. The variable of interest is what the user supplies, and what supplying it costs.

Let $x \in X$ denote the specification the user provides, including the goal, the constraints, the context, and the criteria for success. Throughout, the user is a role, not a fixed person: whichever party supplies the specification and the declaration in the episode at hand. A specification has two features we keep separate throughout. The first is how complete and well-formed it is, as an object. We capture this with a single number, its amplitude, written $a(x)$. Amplitude does not record what the specification is about; it records how fully and clearly the specification is articulated, setting aside whether its content is correct or points in the right direction. A specification that states its goal, its constraints, and its success criteria in full has high amplitude; one that leaves them implicit or partial has low amplitude.

Formally, amplitude is a function

$$a: X \to \mathbb{R}_+,$$

assigning each specification a non-negative number $a(x)$. Two points matter for what follows. First, amplitude is a property of the specification itself, defined without reference to the environment in which it is later read. The same specification has the same amplitude whether the surrounding environment is rich or poor. Second, amplitude is a scalar, which is what lets us state admissibility as a threshold and compare specifications by how far they clear it. The single number does not record which element of a specification is missing; that loss is deliberate, since Gate 1 asks only whether the specification can be reliably decoded. What is missing shows up, if at all, in how the outcome fares at the later gates. Direction, what the specification aims at, is a separate matter, treated apart from amplitude.

Amplitude measures the specification, not the work the user does to produce it. The same specification can take more effort to produce in one setting than in another. We use effort informally for the user's own time and energy spent in producing the specification; the specification cost defined below is its formal counterpart. The setting supplies no part of this effort; it only determines how much effort a given specification takes. To capture this, we introduce a second quantity on the user side, the specification cost, together with the environment it depends on. We write $i > 0$ for instruction-environment quality, a strictly positive scalar measure of how well the user's surroundings support writing the specification. Those surroundings, the instruction environment, consist of the templates, guided fields, examples, and clarification mechanisms available while writing. Thus i summarizes, in a single ordered quantity, how much they reduce the cost that a given specification takes to produce. This concerns cost alone. A higher i makes a given specification cheaper to produce, and so lets the user state a given aim more fully. It does

not change the aim itself. What the user is asking for, and which features of the outcome are at stake, are set by the user and do not vary with i. As with amplitude, treating i as a scalar is what lets us compare settings by how high i is and state the cost relations that follow.

The specification cost c is the cost needed to produce a specification of amplitude a in instruction-environment quality i, when the user works efficiently. A user who works inefficiently may spend more; c records this efficient cost, so that the cost is well defined as a function of the amplitude achieved. We write

$$c = c(a, i),$$

and take it to be continuously differentiable, with

$$\frac{\partial c}{\partial a} > 0 \quad \text{and} \quad \frac{\partial c}{\partial i} < 0,$$

so that a more complete specification costs more to produce, while higher instruction-environment quality lowers the cost of producing any given amplitude. Cost and amplitude are distinct axes: amplitude records what the finished specification is, cost records what producing it took, and the two need not move together. Because cost rises strictly with amplitude at fixed instruction-environment quality, the relation can also be read in the other direction. For a given amount of cost, higher instruction-environment quality lets the user produce a more complete specification.

3. Admissibility and Its Cost

3.1 The decoding bar

A specification is not used directly; it is read by the system and turned into usable intent. We call this decoding. Whether decoding is reliable depends on how complete the specification is: a fuller specification leaves less to guess, a sparser one more. There is a level of completeness below which the system cannot reliably recover the intended meaning. We write $h_{\min}$ for this level, the decoding bar, and state the admissibility condition, as

$$a(x) \geq h_{\min}.$$

A specification meets this condition when its amplitude reaches the bar. We call such a specification admissible: structurally complete enough to be reliably decoded, which is a property of the specification, not of how much effort or execution follows it. Two features of $h_{\min}$ matter. First, it is strictly positive. A specification with no content cannot be reliably decoded into any particular intent, so the bar sits above zero: some minimum of completeness is always required. Second, it is constant in our analysis. Because the decoding system is held fixed (§2), the level of completeness it requires does not change as the user's environment changes. The bar is set by the system, and the system is what

we hold still. Instruction-environment quality, which varies, does not move the bar; what it moves is the cost of reaching it.

This already fixes a structural floor. Because the bar is a positive constant, every admissible specification requires the user to supply at least $h_{\min}$ of amplitude, whatever the environment. The environment lowers what that amplitude costs; the amplitude itself stays required. What the user must supply for an admissible specification is therefore bounded below by a fixed amount that the environment does not remove.

3.2 The entry cost

An admissible specification requires amplitude at least $h_{\min}$, and producing amplitude costs effort. The cheapest way to be admissible is to produce exactly $h_{\min}$: since cost rises with amplitude, any specification above the bar costs more than one at the bar. We write $c_{\min}(i) \coloneqq c(h_{\min}, i)$ for the cost of producing a specification at the bar in instruction-environment quality i. It is the least a user can spend and still be admissible, so we call it the entry cost.

Because $h_{\min}$ is constant, instruction-environment quality enters the entry cost only through the cost of producing that fixed amplitude. Evaluating the cost assumption $\partial c/\partial i < 0$ at $a = h_{\min}$ gives $\partial c_{\min}/\partial i < 0$.

The entry cost falls as instruction-environment quality rises. The bar stays exactly where it is, while the cost of clearing it goes down. This is the precise content of the claim that a better environment makes admissibility cheaper: the standard is unchanged, and what improves is the effort required to meet it.

The same fact can be read from the user's side. Fix any cost level $\bar{c}$ that the environment can realize, that is, a level in the range of $c(\cdot, i)$, and ask how much amplitude that cost buys. Because cost rises strictly with amplitude at a fixed environment, the cost function is invertible in amplitude. For each environment, a given cost corresponds to a single amplitude the user can produce, the amplitude at which cost equals $\bar{c}$. Write this as $a(\bar{c}, i)$. The function $a(\bar{c}, i)$ is the inverse of $c(a, i)$ in its first argument; we keep the letter a because its value is an amplitude, but it is not the amplitude map $a(x)$. Raising instruction-environment quality lowers the cost of every amplitude, so the amplitude that a fixed cost can reach goes up:

$$\partial a(\bar{c}, i)/\partial i > 0.$$

The two statements are one fact seen from two sides. Holding amplitude at the bar and asking for its cost gives $\partial c_{\min}/\partial i < 0$: the same admissible specification becomes cheaper. Holding cost fixed at $\bar{c}$ and asking for its amplitude gives $\partial a(\bar{c}, i)/\partial i > 0$: the same outlay produces a more complete specification. In both readings, the bar is fixed and only the cost relation shifts; what moves is never the standard, only the cost that meeting it requires.

This is what it means, in this framework, for a better environment to widen participation. Admissibility is unchanged, so the set of specifications that count as admissible is the same in a rich environment and a poor one. What changes is that clearing the fixed bar costs less, so a user spending any given cost reaches it in a wider range of

environments. Participation widens not because the standard drops, but because the cost of meeting it does.

4. The Evaluation Objects

§3 fixed the condition a specification must meet to be decoded, and what meeting it costs. That condition is about the specification alone. The rest of the framework is about what the system then produces, the outcome, and how it is judged. Judging an outcome takes more than a single number: it takes a way to say where the outcome points and how good it is, measured against what the user fixed in advance. This section builds those objects, the outcome, the user's *ex ante* criterion, and the two axes along which the outcome is read. Each is defined before it is used, so that what comes later rests only on objects already in hand.

4.1 The outcome

Where amplitude $a(x)$ measures the specification, the outcome is what the system produces from an admissible specification; §5.1 separates this from the bare output. We represent the outcome as a vector R in an evaluation space $\mathbb{R}^d$, and write $\|\cdot\|$ for the Euclidean norm and $\langle\cdot,\cdot\rangle$ for the inner product on $\mathbb{R}^d$. The coordinates of the evaluation space, and the units in which they are read, are fixed in advance and held fixed throughout. Representing the outcome as a vector, rather than a single number, is what lets us read two things off it separately: the direction it points and its magnitude on a given axis. How R is produced from the specification, through decoding and execution, is set out in Appendix B.1; here we need only that the outcome is a vector in $\mathbb{R}^d$.

4.2 The criterion and the two axes

An outcome is judged against a criterion that the user fixes before execution, not against the environment that produced it. From the specification x, a normative extraction map Γ builds this *ex ante* criterion,

$$x^\star := \Gamma(x) \in \mathbb{R}^d.$$

Because i does not affect the aim or which features are at stake (§2), the criterion $x^\star$ is immune to i. Γ reads the specification only through the aim and the features at stake, so its output inherits their independence of i. We record this as a named assumption, criterion invariance, rather than as a derived fact. The immunity is structural: if $x^\star$ moved with i, the same environment would both set the standard and meet it. What i affects is the cost of reaching admissibility, and through that cost which specification the user can afford to write; it touches neither the decoding rule nor the criterion. The criterion is the user's, fixed in advance. The outcome is read along two axes, carried by two subspaces of $\mathbb{R}^d$. The directional subspace T holds the dimensions the user takes to be at stake: what the outcome should be aiming at.

The quality subspace Q holds the non-directional standards the institution applies: how well-built the outcome must be, regardless of where it points. The two are orthogonal,

$$T \perp Q,$$

so direction and quality measure different things about R and cannot stand in for each other. This orthogonality is a design requirement on how evaluation is represented, not an empirical claim about existing schemes. The two subspaces need not span $\mathbb{R}^d$: components of an outcome outside both are read by no gate, and evaluation is deliberately confined to the declared and the institutional axes. The prior review of §5.2 is where this requirement on declarations is enforced. Let Π_T and Π_Q be the orthogonal projections of $\mathbb{R}^d$ onto T and onto Q.

For any outcome R, then, $\Pi_T(R)$ is its directional part and $\Pi_Q(R)$ is its quality part. The user's target direction is the criterion seen along the directional axis, normalized to unit length, $\boldsymbol{t} \coloneqq \Pi_T(x^\star)/\| \Pi_T(x^\star) \|$, $\| \boldsymbol{t} \| = 1$, for $\Pi_T(x^\star) \neq 0$. We assume the user's declaration carries a direction, so that $\Pi_T(x^\star) \neq 0$ holds and the target direction is well defined. Since $x^\star$ is fixed in advance and free of i, so is $\boldsymbol{t}$: the direction the user is aiming at is set before the outcome exists. Two thresholds complete the criterion. The user sets an alignment threshold $\rho \in (0,1]$, how closely the outcome's direction must match $\boldsymbol{t}$. We take ρ to be strictly positive, because a declaration that tolerated orthogonal or opposed outcomes would not require the outcome to point toward the target at all, and the declaration is assumed to carry a direction. When the directional subspace is one-dimensional, every positive value of the alignment threshold yields the same admissible region: the outcomes whose directional component points along the target. The threshold then reduces to an orientation check. When the directional subspace has dimension at least two, we read the threshold as a graded tolerance. The institution sets a quality bar $\theta_Q > 0$, how much quality magnitude the outcome must carry. The triple $(T, \boldsymbol{t}, \rho)$ is the user's; the pair (Q, θ_Q) is the institution's. The triple fixes when an outcome counts as pointing the right way. The outcome's direction is $\Pi_T(R)/\| \Pi_T(R) \|$, and it points the right way when its alignment with $\boldsymbol{t}$ reaches ρ. We call the outcomes that meet this the directionally admissible region,

$$A_\rho \coloneqq \left\{ R : \left\langle \frac{\Pi_T(R)}{\| \Pi_T(R) \|}, \boldsymbol{t} \right\rangle \geq \rho, \qquad \| \Pi_T(R) \| > 0 \right\}$$

Membership in A_ρ is a question of direction alone: it asks where R points, not how good it is. Quality is the other axis, read off the quality part $\Pi_Q(R)$. The coordinates of Q are set up by the institution so that each quality dimension is measured on a nonnegative scale, with larger values meaning better built; deficits appear as small magnitudes, not as negative ones. We record this as a named assumption, the nonnegative quality convention. Where direction compares an angle, quality is a matter of magnitude: how much $\Pi_Q(R)$ carries on the institution's standards, with larger meaning better-built. The institution fixes a level θ_Q this magnitude must reach,

$$\| \Pi_Q(R) \| \geq \theta_Q,$$

so an outcome meets the quality standard when its quality magnitude clears θ_Q. More fully, the institution may judge quality not by this aggregate alone but also by a floor on each dimension of Q that it evaluates, with safety-critical dimensions carrying higher floors. A high aggregate then cannot offset a low value on any single dimension. The main text uses the aggregate bar throughout; Appendix C.2 states the general form.

5. The Three Gates

§4 built the objects: the specification and its amplitude $a(x)$, the outcome R, the user's criterion $x^\star$, the user triple $(T, \boldsymbol{t}, \rho)$, and the institution pair (Q, θ_Q). An episode is judged by three conditions on these objects. We name them here and state each as a single condition. The three are checked in order, and each later one is defined only where the earlier one holds.

5.1 Gate 1: amplitude

Gate 1 is the admissibility condition of §3, named. A specification clears it when its amplitude reaches the decoding bar,

$$a(x) \geq h_{\min}.$$

Gate 1 is checked on the specification, before any outcome exists. Below the bar the system does not fall silent: it still produces an output. The two words are kept apart deliberately: an output is what the system produces, an outcome is what stands to be evaluated. What is missing is standing, not production. We write what the later gates receive as

$$\begin{cases} R, & \text{if } a(x) \geq h_{\min}, \\ \perp, & \text{if } a(x) < h_{\min}, \end{cases}$$

where R is the outcome of §4.1, now licensed: with the bar cleared, the specification is reliably decoded, and what the system produces stands as the admissible outcome the later gates read. $\perp$ marks the opposite state. It does not name the output, and it is not a second kind of outcome for the later gates to judge. It records that no reliable decoding stands behind the output, so there is no R for the later gates to read. This is why Gate 1 is first not only in order but in standing: the later gates judge R, and below the bar there is no R to judge, whatever the output.

Remark (the bar and the form of the output). Gate 1 places its condition on the specification, so $\perp$ does not track the form the output takes. Below the bar the system may ask for clarification, it may hedge, or it may return a polished and confident product. Consider a request to improve a document that says nothing about what counts as improvement. The system may ask what kind of improvement is wanted, or it may return a fluent revision. The two outputs look nothing alike, and both episodes stand at $\perp$,

because in both cases the specification left no intent to be reliably decoded. The polished case deserves emphasis. A confident output looks like an admissible outcome, but there is no decoded intent for it to be faithful to, so we cannot even ask whether it is faithful. Standing comes from decoding, not from appearance.

5.2 Gate 2: direction

Gate 2 is checked on the outcome R, once Gate 1 holds. It asks whether the outcome points where the user declared. The outcome clears it when its direction lies in the directionally admissible region of §4.2,

$$R \in A_\rho.$$

Gate 2 decides membership, not quality: an outcome outside A_ρ is not a poor result to be weighed on its merits, it is not a result of the declared kind at all. Nor is an outcome outside A_ρ the state that $\perp$ records. Directional failure presupposes that Gate 1 holds and an outcome R stands to be judged; $\perp$ marks the prior state, in which nothing stood. In deciding membership, Gate 2 is constitutive. It fixes what counts as an evaluable outcome, rather than rating outcomes already taken to be evaluable.

Remark (governance layer). Gate 2 takes the user's declaration $(T, \boldsymbol{t}, \rho)$ as given and asks only whether an outcome conforms to it, not whether the declaration is itself acceptable. A declaration can be well-formed and still conflict with the institution's substantive commitments, such as non-discrimination, patient safety, or due process. Judging the declaration itself is the task of a governance layer that reviews $(T, \boldsymbol{t}, \rho)$ before the episode enters operation, in forms such as ethics review, regulatory authorization, and institutional oversight. This layer sits above the operating gates. The framework treats it as a structural precondition and does not formalize its internal operation.

5.3 Gate 3: quality

Gate 3 is checked on the outcome, once Gate 2 holds. It asks whether the outcome, being of the declared kind, is well-built on the institution's standards. The outcome clears it when its quality magnitude reaches the bar of §4.2,

$$\| \Pi_Q(R) \| \geq \theta_Q.$$

Gate 3 decides sufficiency: an outcome in A_ρ that falls short of θ_Q is of the right kind but not good enough. This is the one gate that rates an outcome rather than admitting or excluding it. Because Gate 3 is asked only within A_ρ, an outcome that fails Gate 2 is never weighed against θ_Q: the quality question does not arise for a result that is not of the declared kind. The single bar θ_Q on the aggregate magnitude is the simplest case. In safety-critical settings the institution may strengthen Gate 3, either by reshaping the bar into a risk-adjusted form or by adding floors on individual quality dimensions. Appendix C develops both and shows that each keeps Gate 3 a scalar sufficiency test in the same position and ordering. In those variants the institution fixes further parameters

before the episode, the risk operator Ω with its threshold and the dimension floors F_k; the pair (Q, θ_Q) of §4.2 is its baseline standard.

5.4 Joint admissibility

An episode is jointly admissible when the three conditions hold in order:

$$a(x) \geq h_{\min} \quad \text{(Gate 1; on the specification, before any outcome)},$$
$$R \in A_\rho \quad \text{(Gate 2; on the outcome, given Gate 1)},$$
$$\| \Pi_Q(R) \| \geq \theta_Q \quad \text{(Gate 3; on the outcome, given Gate 2)}.$$

The order is not a convenience. Gate 1 is checked before any outcome exists; Gate 2 and Gate 3 are checked on the outcome, first its direction, then its quality. The gates do not loop back: clearing a later one never reopens an earlier one. They sit at two layers, Gate 1 on the specification and Gate 2 and Gate 3 on the outcome, and are not merged into one. The user's $(T, \boldsymbol{t}, \rho)$ and the institution's (Q, θ_Q) are fixed before execution begins; Gate 2 applies the first, Gate 3 the second; and the orthogonality $T \perp Q$ keeps direction and quality from standing in for each other.

6. Institutions and Participation

The structure so far describes a single episode: one user, one specification, one outcome. Institutions are made of many episodes, and they differ in the instruction-environment quality they provide. Across an institution's episodes, the instruction-environment quality i is not one value but a spread, richer in some episodes and poorer in others. Let F summarize the distribution of instruction-environment quality by

$$F(\tau) \coloneqq \Pr(i \geq \tau), \quad \tau \in \mathbb{R}_+,$$

the share of episodes whose instruction-environment quality is at least τ. As the threshold τ rises, fewer episodes satisfy it, so F is non-increasing in τ. The question is how an upward shift in instruction-environment quality changes who can participate, given that the decoding bar $h_{\min}$ stays fixed.

Suppose instruction-environment quality shifts upward across episodes: high values of i become more common and low values become less common. This is a statement about the entire distribution of i, not about a single summary statistic, and we make it precise through first-order stochastic dominance. Let F' denote the distribution of instruction-environment quality after such a shift. Then, for any threshold τ, the share of episodes whose instruction-environment quality meets or exceeds τ is at least as large under F' as under F,

$$F'(\tau) \geq F(\tau), \quad \text{for all } \tau \in \mathbb{R}_+.$$

This is more demanding than a comparison of averages: at every quality level, the share of episodes at or above that level does not fall.

The decoding bar does not move with the distribution; what moves is the cost of reaching it. By §3.2 the entry cost $c_{\min}(i)$ is strictly decreasing in i, so higher instruction-environment quality makes admissibility cheaper. Fix a user who can spend up to a maximum cost $c_{\max}$, the most that user will spend in an episode. The user is admissible whenever the entry cost does not exceed this maximum,

$$c_{\min}(i) \leq c_{\max}.$$

In the notation of §3.2, the same condition can be read on the amplitude side: since cost rises strictly with amplitude, $c_{\min}(i) = c(h_{\min}, i) \leq c_{\max}$ holds exactly when $a(c_{\max}, i) \geq h_{\min}$. The user is admissible exactly when the amplitude their maximum cost can buy reaches the bar. Because $c_{\min}(i)$ decreases as i rises, the episodes satisfying $c_{\min}(i) \leq c_{\max}$ form an upward-closed set in i: whenever an episode qualifies, so does any episode with higher instruction-environment quality. An upward shift in the distribution of i therefore weakly increases the share of episodes in which the entry cost falls at or below $c_{\max}$. Equivalently,

$$\Pr\left(c_{\min}(i) \leq c_{\max} \mid F'\right) \geq \Pr(c_{\min}(i) \leq c_{\max} \mid F).$$

A user facing the same maximum cost $c_{\max}$ is admissible in at least as large a share of episodes under F' as under F.

This is §3.2 viewed at the population level. The argument rests on one mechanism: an upward shift in the distribution of instruction-environment quality produces at least as large a share of episodes in which the required entry cost falls below a fixed maximum user cost. No parametric functional form is required beyond the monotonicity and regularity assumptions of §2. The result relies only on the monotone relation

$$\frac{\partial c_{\min}}{\partial i} < 0,$$

together with a first-order stochastic improvement in the distribution of instruction-environment quality. Throughout, the decoding bar $h_{\min}$ remains fixed. Institutions do not widen participation by lowering the standard a specification must meet. They widen participation by lowering the cost of meeting that standard, so that the same maximum cost $c_{\max}$ is sufficient for admissibility in at least as large a share of episodes. What shifts is the cost, not the bar. Participation here refers to this episode-level feasibility for a given user; decisions about uptake lie outside the model.

7. Direction under Amplification

The three gates fix when an episode is admissible. They say nothing yet about scale: how the picture changes as the system executes more. This section adds that variable and asks a single question. As execution grows, which of the three gates does it relax, and which does it leave untouched? The answer is not symmetric, and the asymmetry is the

paper's main point: scale reaches one gate and not the other two, so growing execution cannot replace what the user supplies.

7.1 Execution scale

Through §5 the outcome R was a single fixed result. But the system can execute the same decoded intent at different magnitudes: more compute, more parallel attempts, longer runs, under a fixed execution protocol. We capture this with a single scalar, the execution scale $S > 0$, and from here write the outcome as $R(S)$ to mark that it depends on S. Larger S means more execution applied to the same specification. The three conditions of §5 are unchanged in form; what changes is that the outcome they read now moves with S. How $R(S)$ is produced from the specification and S is set out in Appendix B.1; here we need only how each gate responds as S grows.

7.2 What scale reaches

Take the three gates in turn and ask how each responds as S grows. Gate 1 does not move. Its condition, $a(x) \geq h_{\min}$, is about the specification, and the specification is fixed before any execution begins. Scale acts on what the system does with the decoded intent, not on the specification itself, so no amount of S changes whether the bar is cleared. Gate 1 is settled before S enters.

Gate 2 does not move either, for a different reason. Its condition, $R \in A_\rho$, is about direction: whether the outcome points where the user declared. Scale produces more of an outcome, but more of it is not a direction for it. A larger S can make the outcome bigger without aiming it at $\boldsymbol{t}$; an outcome pointed the wrong way stays misaimed no matter how large it grows. Direction is set by the user's declaration $(T, \boldsymbol{t}, \rho)$, and execution does not supply it. Appendix B.3 states the quantitative form of this claim. Under the regularities of B.2, the Gate 2 verdict is independent of S outside a band around ρ whose width is set by the audited error bound (B.3).

Gate 3 is the one gate scale reaches. Its condition is on quality magnitude: whether the outcome is well-built on the institution's standards. Magnitude is what execution adds. Applying more execution to the same decoded intent raises the quality the outcome carries, so the quantity Gate 3 measures, $\| \Pi_Q(R(S)) \|$, grows with S on the operating range of Appendix B.2, up to the audited relative error. An outcome that falls short at low scale can clear the bar at higher scale. One limit remains even here: scale lifts the magnitude only if the decoded intent carries a quality component to lift. If the decoded intent carries nothing on Q, more execution scales nothing, and Gate 3 stays unmet at every S.

7.3 Two limits, and where responsibility sits

§7.2 found a split: of the three gates, scale moves only Gate 3, and leaves Gate 1 and Gate 2 where they were. Two limits follow.

The first is on amplitude. Gate 1 holds before execution begins, and the bar $h_{\min}$ is a fixed positive constant. So, however large S grows, the user must still supply a specification that reaches $h_{\min}$: scale does not lower the bar. There is always a minimum

specification, and it is the user who supplies it. Growing execution does not relieve the user of it.

The second is on direction. Gate 2 holds or fails on where the outcome points, and scale adds magnitude, not direction. More execution makes the outcome larger, not more aligned; an outcome aimed where the user did not ask stays misaimed at every S. The direction is the user's declaration $(T, \boldsymbol{t}, \rho)$, fixed before any execution. No matter how much execution is added, it cannot replace the direction the user set.

The two limits land in the same place. What scale cannot reach, the user must supply: the minimum specification at Gate 1, and the direction at Gate 2. Within the operating range of Appendix B.2, scale lifts quality magnitude, and it still leaves both of these with the user. This is the sense in which amplification does not remove the user's responsibility but isolates it: as execution grows, the gates it cannot move are exactly the ones the user is answerable for. This step from structure to responsibility rests on one normative premise, stated openly: a party is answerable for what it alone sets and what no other part of the system can supply or revise. The framework adopts this premise rather than deriving it. Under it, the user is answerable for the declaration and the minimum specification, while the governance layer of §5.2 carries the distinct responsibility of testing the declaration before operation (§8.2, §9.2.2); responsibility for the declaration does not exhaust responsibility for the deployment. The quantitative form of the binding requirement, and how it behaves as S grows, is given in Appendix B.3.

This asymmetry has a close counterpart in alignment research, where added capability does not secure the intended objective: an agent can retain competence while pursuing the wrong goal, and stronger optimization can improve a proxy objective while the intended outcome worsens (Di Langosco et al., 2022; Pan et al., 2022). Those results concern learned or supplied objectives in reinforcement-learning settings. Here, by contrast, the direction is the user's *ex ante* declaration in a single episode, and scale is execution magnitude applied to the decoded intent. Relocating the asymmetry to the human-AI episode, and making it structural through the gates, is what the present treatment adds.

8. Practical Implications

8.1 Training and inclusive design

As execution becomes automated, the scarce skill shifts upstream. What the system does with a specification is increasingly handled by the system; what still varies, and still decides the outcome, is the specification the user supplies. The skill that matters is therefore the construction of specifications. This means stating the goal, the constraints, and the success criteria completely enough to clear the decoding bar (Gate 1). It also means declaring the direction precisely enough that the outcome lands where the user intends (Gate 2). Two users of the same system can get very different results, not because one executes better, but because one supplies a specification that is more complete and more clearly aimed. For training in educational, professional, and public-service settings, the core skill is no longer fluent text production. It is, rather, the ability to say what the work is for, what must not happen, and how success will be judged.

The same threshold logic shapes who can participate. §6 showed that a better instruction-environment quality does not lower the fixed bar but lowers the entry cost of clearing it. Inclusive design works on that cost. Guided input fields, ambiguity checks, and standardized templates lower the entry cost $c_{\min}(i)$, so that a user spending a given amount reaches the bar in a larger share of episodes. This lever does not touch the standard. This is what separates inclusive design from dilutive design. Inclusive design lowers the cost a user faces in clearing the fixed bar. Dilutive design, by contrast, would weaken the standard itself by lowering the decoding bar at Gate 1, so that less complete specifications are admitted.

Consider two researchers using the same generative AI system to develop a research proposal for a human-subject study. One works at an institution that provides standardized Institutional Review Board (IRB) templates tailored to different categories of research, together with examples, checklists, and procedural guidance. The other works at an institution where the relevant information is available only through fragmented, less systematically organized materials. Suppose both face broadly similar IRB expectations; the difference lies only in instruction-environment quality i. At the institution with standardized templates, the researcher receives clearer guidance on formulating research questions, describing study procedures, specifying risks and benefits, and documenting safeguards. The proposal they can write at a given cost is more complete, that is, higher in amplitude $a(x)$; equivalently, the decoding bar is reached at a lower entry cost $c_{\min}(i)$. At the institution relying on fragmented materials, the researcher must spend more to identify what information is required, locate relevant guidance, and determine how it should be expressed. As a result, the same decoding bar $h_{\min}$ is reached only at a higher cost.

8.2 High-stakes domains: Gate 2 and the governance layer

In domains where errors carry serious consequences, such as healthcare, finance, employment, or public infrastructure, the gate structure points to one place. Gate 2 takes the user's declaration $(T, \boldsymbol{t}, \rho)$ as given and checks conformity to it. By the asymmetry of §7, growing execution scale S cannot supply or correct direction. It only enlarges whatever the declaration admits. A badly set declaration therefore scales into large harm, and the failure is not performance below a threshold. It is the declaration itself being set wrongly, so that outcomes the institution should never accept count as exactly what was asked for. And because operation checks conformity to the declaration, nothing inside operation revises the declaration itself. Responsibility sits at the moment of declaration, and correction must come before operation, not during it.

That prior position is the governance layer of §5.2, and it functions as the precondition on which Gate 2 rests. Gate 2 can take $(T, \boldsymbol{t}, \rho)$ as given only because a layer above it has already judged that this declaration may enter operation at all. For the layer to bear that weight, the declaration must be documented, auditable, and stable, and the procedures that produce and revise it must themselves bear scrutiny.

Consider a company that brings a generative AI system into a personnel management project, screening and evaluating its workforce. Seeking to push out its older staff, the

company declares a direction $(T, \boldsymbol{t}, \rho)$ that places age among the scored dimensions of T and weights $\boldsymbol{t}$ so that younger employees rank higher. The declaration makes discrimination the objective. Under the declaration, outcomes that disadvantage older employees are exactly what was asked for. An outcome conforming to it clears Gate 2, and by the asymmetry of §7 no execution scale turns that direction legitimate. What stops it sits above operation. The governance layer of §5.2 reviews the declared $(T, \boldsymbol{t}, \rho)$ before any episode runs, here through the company's compliance and legal review under age discrimination law. The review examines which dimensions T includes, what weight $\boldsymbol{t}$ places on each, and how much room ρ leaves. Finding that the declaration targets a characteristic the law protects, the layer does not permit it to enter operation and returns it for revision. The filtering happens at the declaration, before any episode runs, which is the only place it can happen.

One might object that the system itself, tuned toward safe behavior, will soften or refuse the age-based ranking the declaration asks for. Such behavior changes what is produced, not what counts. The declaration is fixed before execution, so the system can shape the outcome R, but it cannot alter T, $\boldsymbol{t}$, or ρ. The same declaration still targets the discriminatory outcomes; a softened or refused output may simply fail Gate 2 under it, and whatever does clear Gate 2 is credited as on target. Protection that depends on which point in A_ρ the system happens to produce is unstable across models, versions, and episodes. It is recorded nowhere in the declaration and, unless separately governed, answers to no prior review; it can supplement, but cannot replace, that review. The governance layer puts the protection where it can be reviewed: in the declaration itself.

The safeguard in high-stakes deployment is therefore joint. A well-set Gate 2 without an effective governance layer is fragile, since nothing inside the operation prevents the declaration from being set wrongly to begin with. A governance layer without Gate 2's anchored declaration has nothing structural to review. Together they form the safeguard. Either alone is insufficient. The risk diagnosis above applies where the layer is absent, weak, or has failed, and the asymmetry of §7 offers no compensation for that failure. The framework treats the layer's operation as a structural precondition and leaves the design of its review and accountability mechanisms to future work.

9. Realism, Implementation, and Research Seeding

9.1 Institutional correlates of framework constructs

A recurring concern with abstract models of human-AI interaction is whether they describe anything that exists outside theory. The present framework is built to avoid that gap: its core objects correspond to institutional structures already in operation, and each correspondence seeds an independent line of research.

First, instruction-environment quality i is not a latent psychological variable but the degree to which the surrounding materials supports the user in writing a specification. These supports include templates, guided fields, examples, ambiguity checks, and standardized vocabularies. They are already deployed in clinical documentation, regulatory compliance, procurement, and annotation pipelines. Treating i as a structured

institutional variable makes these design choices objects of formal study. And since an institution's episodes carry a distribution of i, the F of §6, institutional design becomes a question with a definite shape. The open questions are which interventions shift F upward, and how far the shift travels into entry costs and participation.

Second, the decoding bar $h_{\min}$ formalizes a familiar but under-theorized boundary: the minimum completeness below which a specification is not usable, whatever the system then produces. In the framework the bar belongs to the decoding system and does not move with the environment. What varies across organizations is the cost of reaching it, $c_{\min}(i)$, borne in practice as the effort whose required amount templates and checks either reduce or leave with the user. Two empirical objects follow. One is the bar itself, located, for a given system, at the point where specifications begin to be reliably decoded. The other is the entry cost, measured across environments as the effort a usable specification requires under different scaffolding regimes. The framework's separation is itself testable: holding the system fixed, improvements in the instruction-environment quality should appear in measured costs, not in where the bar sits (§3, §6).

Third, the threshold form of participation yields a sharp prediction: usability should turn on clearing the bar rather than rise smoothly with better instructions. Modeling participation through $a(x) \geq h_{\min}$ renders this discontinuity directly. It also invites work on scaling and on the risks that arise when execution capacity grows faster than the scaffolding that supports admissible specification. The two grow on different engines: execution capacity grows with the era, while the scaffolding grows only where an institution builds it. When the first outruns the second, the share of episodes a user can afford stands still, and a badly set declaration is carried at a larger scale (§6, §8).

Fourth, the three-gate structure of §5 corresponds to sequential evaluation practices in regulated settings. A medical diagnostic system is categorized within a condition class before its accuracy is assessed. A financial report is verified for required disclosures before substantive review. A compliance submission is checked for inclusion in the regulatory domain before its content is examined. Failure at a prior gate terminates evaluation rather than producing an unfavorable judgment within it, and each gate is verified at its natural layer, which makes each empirically tractable through existing audit and regulatory mechanisms.

Fifth, the governance layer of §5.2 corresponds to review mechanisms that already operate before deployment in regulated settings. Ethics committees vet protocols before clinical AI enters use. Regulatory bodies examine intended-use statements before market authorization. Boards, in turn, assess whether a proposed deployment's evaluable domain aligns with the institution's substantive commitments. Treating this prior review as a distinct meta-level on the acceptability of the declaration $(T, \boldsymbol{t}, \rho)$ opens research on several questions. These include how such review should be constituted, what standards of documentation and auditability it should enforce, and how its effectiveness can be evaluated independently of downstream system performance. The direction is salient because, by the asymmetry of §7, no operating-gate correction revises an unacceptable declaration; the corrective burden sits on prior review (§8.2).

Finally, *ex ante* evaluation criteria reflect existing practice rather than a theoretical ideal. In regulated environments, criteria are declared in advance as predefined acceptance criteria and scoring rubrics, and outcomes are assessed against them rather than used to redefine them. This is the independence that the criterion of §4.2 requires. The scaling regularities on which the bounds of Appendix B.3 rest are likewise auditable. On a reference sample of admissible episodes, deviations from the nominal scaling relations give institutions an empirical check on the assumed regularities. The audit quantities are stated in Appendix B.4. This opens research in AI governance, accountability, and the political economy of evaluation standards.

Because the core constructs map onto existing institutional mechanisms, researchers can extend, contest, or reinterpret specific components without adopting the entire framework. The contribution is thus not only a closed theoretical argument but the seeding of a broader research ecology around institutional decoding (how institutions make reliable decoding affordable), threshold governance, and admissible human-AI collaboration.

9.2 Institutional anchoring: two contrasting case observations

To anchor the three-gate structure in concrete institutional settings, we present two stylized case observations. These are not empirical validations but illustrative reconstructions, showing how the gates of §5, in their order, map onto distinctions already operative in regulated practice.

9.2.1 Case 1: regulated clinical AI deployment

Deployment of AI-enabled radiology devices under Software as a Medical Device regulation offers an illustrative mapping of the three gates. The regulatory frame includes premarket pathways such as 510(k), the AI/ML Software as a Medical Device (SaMD) Action Plan (U.S. Food and Drug Administration, 2021), the guidance on predetermined change control plans (U.S. Food and Drug Administration, 2025), and the clinical-evaluation principles for SaMD (International Medical Device Regulators Forum, 2017). The last of these asks three questions in order: whether there is a valid clinical association between the output and the targeted clinical condition, whether the software correctly processes input data to generate accurate, reliable, and precise output, and whether use of that output achieves the intended purpose.

Gate 1 has its regulatory counterpart in analytical validation. The requirement is that the software correctly and reliably processes the input it is given, which presupposes that the input is well enough formed to be processed. The guidance does not say what a well formed input contains, and each deploying institution settles that for itself. A radiology service that implements the requirement therefore fixes the imaging protocol, the acquisition parameters, and the clinical question before a study is sent to the device, because the device was validated only for inputs of that kind. These are institutional implementations of a regulatory requirement rather than named regulatory criteria, and the decoding bar is our name for the level at which they are set.

Gate 2 has its counterpart in the intended use, the targeted clinical condition, the target population, and the conditions of use. A device cleared for one purpose is not evaluated on another, and outcomes outside that purpose fall outside the evaluation entirely. The declaring user here is the deploying entity, the manufacturer or the clinical institution acting as principal. Reading the declared triple onto these regulatory objects is our interpretation and not a regulatory equivalence.

Gate 3 has its counterpart in clinical validation against acceptance criteria that are fixed in advance, using measures such as sensitivity and specificity. The guidance on predetermined change control plans asks for such criteria to be predefined and for evidence that any modification preserves safety and effectiveness across the intended-use population. These standards are set at the level of the device and its reference population rather than for a single episode, so the per-episode quality bar is an analytical counterpart rather than a regulatory term.

Above these gates, premarket review operates before deployment. The regulator examines the intended use before the device enters the market, and this is the working counterpart of the governance layer whose failure defines Case 2. Ethics review plays a comparable part in research use, although the cited device guidance does not establish it as a general condition of market authorization.

9.2.2 Case 2: the Australian Robodebt scheme

The Australian Robodebt Scheme, implemented from 2015 and continued through 2019, later became the subject of the Royal Commission into the Robodebt Scheme (2023). It provides the contrast. Where Case 1 presents an illustrative mapping of declaration, review, and operating evaluation, Robodebt shows what can follow when an unlawful design enters operation without effective prior legal review. The lesson the framework draws from it concerns the governance layer.

Two points about the Scheme are supported by the Commission, and the framework reads them in a definite order. The first concerns the declaration. As the declaring user, the deploying agency had to fix $(T, \boldsymbol{t}, \rho)$ before operation. In the framework's reconstruction, the lawful target was the recovery of genuine overpayments established under social security law. The Scheme instead treated employer-reported income, averaged across fortnights and often unconfirmed by the employer or recipient, as a sufficient basis for raising debts. The Commission found that this method did not accord with the statutory use of actual fortnightly income and that the Scheme was "devised without regard to the social security law" (Royal Commission into the Robodebt Scheme, 2023, Executive Summary). The description of this failure as a wrongly set declaration is the paper's interpretive mapping, not the Commission's terminology.

The Commission also documented repeated legal warnings that were not acted on. Internal advice in 2014 identified inconsistency between income averaging and the legislative framework. Draft external advice in August 2018 concluded that averaging employer-reported income was not permissible and should have prompted immediate suspension or further advice from the Solicitor-General, but it was neither finalized nor acted on (Royal Commission into the Robodebt Scheme, 2023, Executive Summary). In

the framework's terms, this supports the narrower claim that the governance layer, here prior legal review, failed before operation. The further proposition that no operating gate could have corrected the declaration is a deduction from the framework, not a finding made by the Commission.

The second supported point concerns the later contestation episode. Recipients bore the onus of contradicting asserted debts by reconstructing actual earnings for periods extending as far back as five years, although official guidance had said that payslips needed to be retained for only six months. The Commission also documented difficulty obtaining records, upload problems, enforced reliance on an online portal, limited digital access and literacy, and information that was difficult to understand (Royal Commission into the Robodebt Scheme, 2023, Executive Summary; Chapter 10). In the framework's reading, these conditions raised the entry cost $c_{\min}(i)$ of an admissible contestation and could leave the attainable amplitude $a(x)$ below $h_{\min}$. The Commission did not estimate these formal quantities or show that the entry cost exceeded every affected person's maximum. The Gate 1 description is therefore an interpretive reconstruction of documented evidentiary and access barriers.

These are the two conclusions the framework can draw with qualified confidence: an unlawful design that should have been stopped by prior legal review, and a later contestation process burdened by severe evidentiary and access barriers. The Commission's broader judgment that the Scheme was neither fair nor legal concerns substantive law and administration beyond what the framework itself establishes. The framework's claims here are narrower and explicitly interpretive.

9.2.3 Interpretive note

The contrast between the two cases fixes the framework's diagnostic use. Case 1 shows the declaration, the review above it, and the operating gates working as distinct layers. Case 2 shows what follows when a declaration is set wrongly and the governance layer above it does not test the declaration before operation. The review that should have been a precondition arrived only afterward, in the form of the Royal Commission. It offered diagnosis rather than prevention. The framework locates the failure there, at the declaration and the review owed to it, not in the operating gates, which can only enforce the declaration they are given. Because the framework was developed independently of both cases, this retrospective reading is constructive rather than predictive: it shows that the framework's distinctions track the institutional distinctions that regulators and inquiries themselves draw. Prospective validation through deployment studies remains future work.

10. Conclusion

This paper presented a compact lens for a shift in human-AI collaboration. As the cost of execution falls, the binding constraint on cognitive production moves upstream to the specification the user supplies and to the institutional conditions under which it can be supplied.

At the level of a single episode, the framework separates the objects that this shift makes load-bearing. A specification x carries an amplitude $a(x)$, its completeness. The decoding bar $h_{\min}$ is a fact about the decoding system and does not move with the instruction-environment quality. What the instruction-environment quality i moves is the cost, with the entry cost $c_{\min}(i)$ falling as i rises (§3). Evaluation rests on a criterion fixed before execution, $x^{\star} = \Gamma(x)$, independent of the instruction-environment quality. It also rests on two orthogonal axes, the user's directional subspace T with target $\boldsymbol{t}$ and the institution's quality subspace Q with bar θ_Q (§4). Three gates then run in order (§5). Gate 1 is checked on the specification, $a(x) \geq h_{\min}$, with $\perp$ marking that below the bar the system's output has no reliable decoding behind it. Gate 2 is checked on the outcome, $R \in A_\rho$, and is constitutive: it fixes what counts as an evaluable outcome at all. Gate 3 is checked on the outcome within A_ρ, $\| \Pi_Q(R) \| \geq \theta_Q$, and decides sufficiency. The gates carry two kinds of primacy that do not compete. Gate 1 is first in order and in standing, since below the bar no admissible outcome arises. Gate 2 is constitutive of evaluability on the outcome. Above the operating gates sits the governance layer, which reviews the declaration $(T,\ \boldsymbol{t},\ \rho)$ before any episode runs (§5.2).

At the level of an institution, the same cost mechanism reads as a statement about participation. An institution carries a distribution F of instruction-environment quality across its episodes. A first-order upward shift in F makes low entry costs more common, so the same user, spending no more, clears the same bar in at least as large a share of episodes (§6). Institutions widen participation by lowering the cost of meeting standards, not by lowering the standards.

The central implication is governance-oriented. Execution scale reaches exactly one gate (§7). Gate 1 is settled before scale enters. Gate 2 is unmoved by it: for a fixed declaration, no value of S, no level of $a(x)$, and no improvement in i modifies the directionally admissible region itself. Outside the audit margins of Appendix B.3 the verdict $R(S) \in A_\rho$ is likewise independent of S. Only Gate 3's quality magnitude grows with S, up to the audited error band. Two limits follow. There is always a minimum specification, bounded below by a fixed bar that neither scale nor instruction-environment quality moves, and there is always a direction, which execution cannot supply. Both sit with the user under the responsibility premise of §7.3, and growing execution does not dilute this responsibility but isolates it. In high-stakes deployment the practical consequence is a joint safeguard: a declaration anchored at Gate 2 and a governance layer that reviews it before operation (§5.2, §8.2).

These distinctions are not a theoretical imposition. Regulated clinical AI deployment illustrates the three gates at their natural layers. The Robodebt scheme shows what that layering prevents: a declaration set wrongly and never tested by the review owed to it before it entered operation (§9.2). The framework formalizes distinctions that institutional practice already draws, often implicitly, when it diagnoses success and failure in AI-mediated work.

The order of exposition has not been incidental. The extended treatment of cost and instruction-environment quality was needed to fix exactly what i and S reach and what they do not. They reach the cost of admissibility and the quality an outcome carries. They

reach neither the bar nor the constitutive layer at which the evaluable domain is declared. The irreducibility of that declaration becomes visible only once the reach of everything else has been stated in full. What remains outside that reach is not a residual detail but the place where the human contribution is structurally anchored. And because the framework's constructs map onto measurable institutional objects (§9.1), its thresholds and comparative statics are open to test rather than left as rhetorical claims.

"The aspects of things that are most important for us are hidden because of their simplicity and familiarity."

Ludwig Wittgenstein, Philosophical Investigations, §129 (Wittgenstein, 1958).

Anticipated Objections and Responses

The framework advanced in this paper invites, and indeed anticipates, a number of serious objections, and we are grateful for the occasion to address them. What follows is not offered as a defensive postscript but as an attempt to engage, with the seriousness they deserve, the concerns that a careful and fair-minded reader might raise. Some touch the conceptual foundations; others the formal apparatus or the scope of application. One further objection of a practical kind, namely that the system itself may supply what a declaration omits, arises inside deployment and was answered where it arises (§8.2). For the objections below, which are structural and conceptual, we try to say plainly where our commitments are strong, where they are deliberately bounded, and where further work is needed.

On the charge of excessive formalism

A natural first response to this paper, and one we expect many readers to share, is that its mathematical apparatus is out of proportion to the phenomena it addresses. One might argue that the essential insight, that specification quality matters more than execution capacity in AI-mediated work, is already widely appreciated in applied AI research. On this view, the insight does not require an amplitude with a decoding bar, a cost function, three gate conditions, projections onto two subspaces, and a stochastic-dominance argument. We concede at once that this objection is not without merit. The intuition itself is not novel, and the reader is entitled to ask whether the formalization adds value beyond what plain language could achieve.

Our reply rests on a distinction between intuition and structural precision. Many claims in the human-AI interaction literature remain, through no fault of their authors, at the level of heuristic observation: that prompts matter, that context shapes output, that alignment is important. These statements are directionally correct but analytically incomplete. They do not say where control is exercised, what determines participation, or how institutional conditions mediate the crossing of thresholds. We do not offer the formalism to impress; we offer it to discipline. By fixing definitions, separating what a specification is from what producing it costs, and deriving comparative statics from explicit assumptions, the framework makes its empirical claims falsifiable, which informal treatments typically do not. It is precisely because the core intuition is so widely shared that we believe its formal articulation earns its keep. Shared intuitions, left unstructured, tend to proliferate interpretations that are mutually inconsistent. We would rather be refutable than vague.

On the abstraction of the specification and the outcome

A reader working in the situated-action tradition would object that a scalar representation cannot capture how meaning is actually produced, since understanding in real interaction emerges from circumstances as they arise, from mutual interpretation, and from repair when interpretation fails. The same objection extends to the outcome side.

Outcomes are not literally vectors, conformity to a goal is not literally a cosine, and a single threshold may appear too thin to carry normative judgment.

We would ask such readers to regard these objects as operational proxies rather than as cognitive models, and we hope to show that the claims carrying weight do not depend on their fine structure. The scalar $a(x)$ captures the degree to which a specification suffices for reliable decoding, not the semantic content of an intention; direction and quality are handled separately, through the criterion $x^\star$ and the two axes. On the outcome side, what the argument uses is modest. It uses only that membership and sufficiency are different questions, with $(T \perp Q)$ keeping the two evaluated components from standing in for each other, that the criterion is fixed before the outcome (§4.2), that the gates run in order (§5), and that scale reaches only quality magnitude (§7). The separation itself depends neither on dimension counts nor on the cosine form in particular, though the graded reading of the alignment threshold turns on dimension (§4.2); A_ρ stands in for any operationalization of a declared direction that is fixed *ex ante*. As for the apparent arbitrariness of ρ, we would respectfully suggest that it is not a defect. The threshold ρ is declared, not derived, and that is what makes it the user's to own and the governance layer's to review (§5.2, §8.2). We readily grant that a richer model of pragmatics and situated meaning would enrich the framework, and we would welcome such work. In the meantime, we ask that $a(x)$ be read as a placeholder awaiting disciplinary refinement, not as a final word on the nature of specification.

On iteration and revision

Many readers may fairly object that the model looks static where practice is iterative: real use is dialogic, users refine their specifications across turns, and an *ex ante* declaration may seem to misdescribe the work.

Our reply is that iteration is a sequence of episodes, and that the declaration constraint binds each episode, not the user's learning across them. Nothing in the framework forbids revising $(T, \boldsymbol{t}, \rho)$ between episodes, and we would not wish it to. A revised declaration enters the next episode *ex ante* and passes through the same procedures of review and revision (§5.2, §8.2). Revision is welcome. The one thing the framework does not allow is fixing the criterion after seeing the outcome it is meant to judge. The criterion for each episode must be set before that episode's outcome, and repeating the work across episodes does not change this (§4.2).

On the separation of direction and quality

Many readers may object that direction and quality cannot, in practice, be cleanly separated: what counts as well-built often depends on what the work is for, so $T \perp Q$ may strike the reader as unrealistic. We acknowledge that many existing evaluation schemes do, in fact, entangle the two.

Our reply is that the orthogonality is offered as a design requirement on evaluation, not as an empirical claim that every existing scheme satisfies it. It asks that evaluation be built so that the two questions cannot substitute for each other. Where criteria entangle, we do not deny the entanglement; rather, we ask that it be recognized as a defect.

On the human anchoring of direction

Perhaps the most philosophically ambitious objection, and the one we have weighed most carefully, challenges the paper's foundational commitment that the setting of ends, values, and normative constraints remains anchored in human agents. Bostrom's orthogonality thesis holds that intelligence and final goals may vary independently across possible artificial agents, while his instrumental-convergence thesis allows sufficiently capable agents to pursue common intermediate goals in service of many different final goals (Bostrom, 2012). A critic might therefore argue that advanced AI could develop or approximate goal-setting capacities that make the human-anchoring assumption obsolete.

We do not presume to deny the theoretical possibility. We claim less: that under current and foreseeable architectures, deployed systems do not generate intrinsic ends; they optimize objectives supplied externally, and the framework is constructed for this operating regime. Our own machinery states the structural version of the point. A system can add content to the outcome, and it routinely generates subordinate goals, decomposing tasks and structuring execution. Within an episode, however, it cannot add a dimension to T, weight to $\boldsymbol{t}$, or strictness to ρ (§8.2). So what stays with humans is not every goal-related activity, but the first step: choosing the ends. The subordinate goals an AI generates are valid only because they serve those human-set ends, and whoever sets the ends is the one answerable for them on review (§5.2).

Should the empirical landscape change in ways that confer autonomous end setting on AI systems, we would be the first to grant that the framework's allocation of responsibility requires fundamental revision. Until then, treating the human declaration as the source of direction is a modeling choice that reflects the actual structure of deployment, not an assertion that machine end setting is impossible. A further commitment motivates the paper, though the framework does not depend on it. On this commitment, the anchoring of human responsibility should be preserved as a permanent structural principle, not treated as a temporary feature of current AI architectures. The framework's main claims still hold under the weaker, operational reading. The stronger, normative reading we leave, respectfully, to the reader.

On power, politics, and distribution

A sociotechnical critic might argue that treating institutional conditions as parameters abstracts away the social actors, institutional settings, and contested values that shape the system (Selbst et al., 2019). A related objection from work on automated welfare administration is that such systems fall hardest on those least able to contest them, so that questions of political economy and distribution cannot be treated as parameters at all (Eubanks, 2018). Who designs the instruction environments, whose specifications count as legitimate, and how standards are set are questions of power and distribution. On this view, the framework may depoliticize a political process.

We acknowledge the force of this concern, and we answer it by pointing to where these questions live inside the framework. Instruction-environment quality makes visible that the conditions of legibility are not a natural given but a product of organizational and governance choices. §6 makes the distributive question statable, since the distribution of i across an institution's episodes sets the entry cost and so determines who can participate.

The distinction between inclusive and dilutive design (§8.1) separates widening participation from weakening standards. The political questions proper, who exercises the governance layer, how it is constituted, and whose voices it represents, attach to that layer (§5.2). There, the acceptability of a declaration against the institution's commitments, including non-discrimination and the protection of disadvantaged groups, is decided before operation (§8.2). The asymmetry of §7 does not insulate the user from accountability for what is declared; it locates that accountability at the prior review. We do not develop a theory of how those choices are made, or in whose interest. We name this as a real limitation, offered as an invitation to political economy and science and technology studies rather than as an oversight.

The foregoing objections are, in our estimation, among the strongest that can be raised against the paper's central claims, and we are indebted to them. That they can be articulated precisely is itself a consequence of the formal structure, for a theory that cannot be clearly opposed says very little. Our ambition has not been to foreclose debate but to offer a structure well defined enough that debate can proceed on shared ground. We will count the paper successful if its critics find it worth refuting.

In Practice

Generative AI is being deployed into high-stakes decision-making faster than institutional, regulatory, and organizational vocabularies can absorb the change. The question is no longer whether AI systems will enter benefits adjudication, clinical documentation, compliance workflows, educational assessment, and public-service delivery. Both case observations of §9.2 concern automated decision systems entering such settings, one where the prior review held and one where it failed. In the first, the declaration and the review above it are enforced as distinct layers. In the second, a declaration was set wrongly and the review owed to it never tested it before operation. The question is whether the institutions receiving these systems can say, with enough precision, what is being deployed, what counts as an admissible use of it, and where responsibility sits when it fails. The framework developed in this paper is offered as one contribution to that articulation, addressed to two classes of institutional decision-makers whose work has come to depend on AI-mediated systems.

For policymakers and governance bodies, the gate structure gives a formal account of why execution scale does not substitute for what the user supplies. By the asymmetry of §7, scale raises only the quality magnitude at Gate 3. It moves neither the bar at Gate 1 nor the direction at Gate 2. Governance designed around output magnitude alone, measuring success by throughput, coverage, or response time, therefore leaves the constitutive question untouched: which outcomes were admitted into evaluation in the first place, and against what declaration. The framework locates the audit point upstream, at the *ex ante* declaration that fixes the evaluable domain, and at the governance layer that reviews this declaration before any episode runs (§5.2, §8.2). A confident output is not yet an admissible outcome; standing comes from reliable decoding at Gate 1 (§5.1). And an admissible outcome is not yet an authorized one: the declaration fixes the evaluable domain, and the governance layer licenses that declaration to enter operation by testing it before any episode runs (§5.2, §8.2). Where that prior review is absent, there is nothing for evaluation to stand on. The Robodebt reconstruction records what follows when a wrongly set declaration enters operation unreviewed (§9.2.2).

For organizations deploying AI-mediated cognitive work, the separation of the bar from the cost provides a diagnostic for where to invest. The decoding bar belongs to the system and does not move with the surroundings; what an organization controls is the entry cost its people face in clearing it (§3). Many organizations experience a familiar pattern: AI tools are introduced, throughput rises, and yet results disappoint as correction cycles and rework accumulate. The framework offers one diagnosis: an under-invested instruction environment. Episodes run on specifications too thin to be reliably decoded, so what the system produces, however fluent, has no decoded intent behind it. The remedy is not more execution capacity, which moves only Gate 3. It is investment in the instruction environment, the templates, guided fields, examples, and clarification that lower the cost of an admissible specification. It is also investment in training that treats the construction of specifications, stating the goal, the constraints, and the success criteria, as the core upstream skill (§8.1).

These are not applications the paper performs. They are vocabularies the paper makes available. And because the constructs they name correspond to measurable institutional objects, each vocabulary comes with an empirical handle. Those objects are the bar located for a given system, the entry cost measured across environments, and the distribution of instruction-environment quality observed across an institution's episodes (§9.1).

The paper deliberately refrains from several things. It offers neither a predictive model of AI capability trajectories, nor specific policy prescriptions, nor advocacy for or against any particular technology. These omissions are intentional. The purpose is diagnostic and structural. Its first aim is to identify the variables that determine whether human-AI collaboration produces outcomes that are not merely large, but admissible, directed, and accountable. Its second aim is to locate the responsibility that stays with the user, for the minimum specification and for the direction, even as execution scales and environments improve. In an era when public discourse about AI oscillates between uncritical enthusiasm and categorical alarm, the framework is offered as a contribution to a conversation that will require many voices, many frameworks, and sustained patience.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). Glasgow, Scotland.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7, No. 1, pp. 2-11). AAAI Press.

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., & Kluge, A. (2023). Defining human-AI teaming the human-centered way: A scoping review and network analysis. *Frontiers in Artificial Intelligence*, *6*, 1250725.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2), 71-85.

Di Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., & Krueger, D. (2022). Goal misgeneralization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 12004-12019). PMLR.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, *384*(6702), 1306-1308.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 1-7.

International Medical Device Regulators Forum. (2017). *Software as a medical device (SaMD): Clinical evaluation* (IMDRF/SaMD WG/N41FINAL: 2017). https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1369-1385). Association for Computing Machinery.

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4-11.

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. *AI and Ethics, 4*(4), 1085-1115.

Myers, B. A., Hudson, S. E., & Pausch, R. (2000). Past, present, and future of user interface software tools. *ACM Transactions on Computer-Human Interaction, 7*(1), 3-28.

Pan, A., Bhatia, K., & Steinhardt, J. (2022). *The effects of reward misspecification: Mapping and mitigating misaligned models.* International Conference on Learning Representations (ICLR) 2022, Online Conference.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477-486.

Royal Commission into the Robodebt Scheme. (2023). *Report of the royal commission into the robodebt scheme*. Commonwealth of Australia. https://robodebt.royalcommission.gov.au/publications/report

Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59-68). Association for Computing Machinery.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, *36*(6), 495-504.

U.S. Food and Drug Administration. (2021). *Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan*. U.S. Food & Drug. https://www.fda.gov/media/145022/download

U.S. Food and Drug Administration. (2025, August 18). *Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions*. U.S. Food & Drug. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence

Wang, D., Weisz, J. D., Muller, M., Ram, A., & Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., & Gray, A. (2019). Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-24.

Wittgenstein, L. (1958). *Philosophical investigations* (G. E. M. Anscombe, Trans.; 2nd ed.). Blackwell (Original work published 1953).

Appendix A. Notation and Key Assumptions

Notational convention. Throughout the paper, $\mathbb{R}_+$ denotes the non-negative reals $[0, \infty)$, except where a symbol is stated as strictly positive $(i > 0,\ h_{\min} > 0,\ \theta_Q > 0,\ \theta_Q^{\Omega} > 0,\ S > 0,\ \lambda_T > 0,\ \lambda_Q > 0,\ \lambda_{Q_k} > 0,\ F_k > 0)$. On the evaluation space $\mathbb{R}^d$, $\|\cdot\|$ is the Euclidean norm and $\langle\cdot,\cdot\rangle$ the inner product (§4.1). τ is a generic threshold variable. The symbol $\perp$ is used in two ways: between subspaces it marks orthogonality (§4.2), and on its own it is the absence marker of §5.1. The letter F is likewise used in two ways: as the distribution of §6 and as the dimension floors of Appendix C.2.

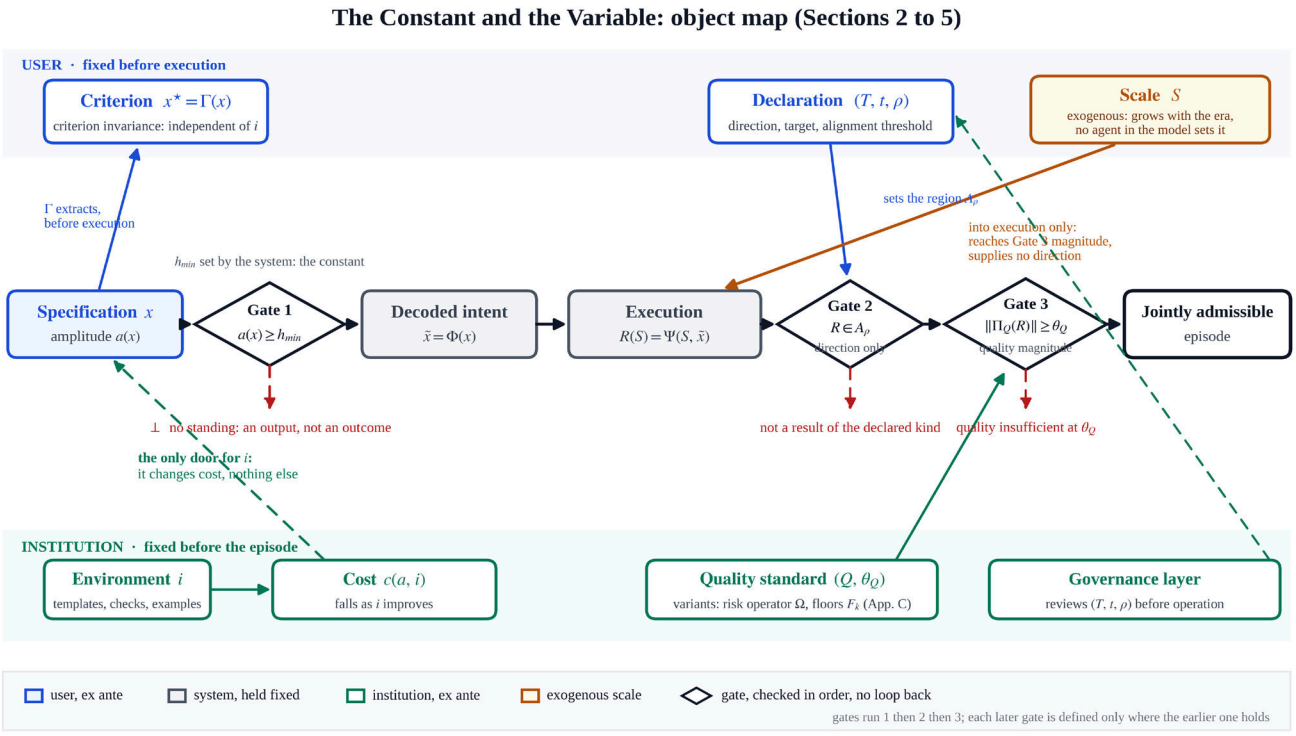

Fig. A1. The object map of sections 2 to 5. Ownership is marked by color; the environment i enters through cost alone; scale S enters execution alone; the gates are checked in order.

Table A1. Specification and cost

Symbol	Definition
$x \in X$	Specification provided by the user, including the goal, constraints, context, and criteria for success (§2).
$a(x) \in \mathbb{R}_+$	Amplitude. The completeness and well-formedness of the specification as an object. It is a property of the specification itself and is defined independently of the environment. Formally, $a: X \to \mathbb{R}_+$ (§2).
$i > 0$	Instruction-environment quality. A strictly positive scalar measure of how well the surrounding instruction environment, including templates, guided fields, examples, and clarification mechanisms, supports the user in writing a specification (§2).
$c(a, i)$	Specification cost. The efficient cost of producing a specification of amplitude a under instruction-environment quality i. It is continuously differentiable and satisfies $\partial c/\partial a > 0$ and $\partial c/\partial i < 0$ (§2).
$h_{\min}$	Decoding bar. The minimum amplitude required for reliable decoding. It is strictly positive and fixed by the decoding system. It is not altered by i (§3.1).
$c_{\min}(i) \coloneqq c\left(h_{\min}, i\right)$	Entry cost. The minimum cost required for admissibility. It satisfies $\partial c_{\min}/\partial i < 0$ (§3.2).
$a(\bar{c}, i)$	The amplitude attainable at any fixed cost $\bar{c}$ under instruction-environment quality i; the inverse of $c(a, i)$ in its first argument, distinct from the amplitude map $a(x)$. It satisfies $\partial a(\bar{c}, i)/\partial i > 0$ (§3.2).

Table A2. Evaluation objects

Symbol	Definition
$R \in \mathbb{R}^d$	Outcome. The object produced by the system from the specification, represented as a vector in the evaluation space. An outcome stands for the later gates only when Gate 1 has been cleared (§4.1, §5.1).
Γ	Normative extraction map. The mapping that constructs the *ex ante* criterion from the specification. The map is fixed independently of instruction-environment quality i (§4.2).
$x^\star \coloneqq \Gamma(x) \in \mathbb{R}^d$	*Ex ante* criterion. The criterion fixed by the user before execution. It is determined from the specification and remains independent of the environment (§4.2).
Criterion invariance	Named assumption: the criterion is fixed by the specification alone and does not vary with instruction-environment quality (§4.2).
$T \subseteq \mathbb{R}^d$	Directional subspace. The dimensions regarded by the user as normatively relevant for the episode (§4.2).
$Q \subseteq \mathbb{R}^d$	Quality subspace. The dimensions used by the institution to assess non-directional quality. The framework assumes $T \perp Q$ (§4.2).
Nonnegative quality convention	Named assumption: each quality coordinate is measured on a nonnegative scale, with larger values meaning better built; deficits appear as small magnitudes, not as negative ones (§4.2).
Π_T, Π_Q	Orthogonal projection operators. The projections from $\mathbb{R}^d$ onto the directional subspace T and the quality subspace Q, respectively (§4.2).
$\boldsymbol{t} \coloneqq \frac{\Pi_T(x^\star)}{\lVert \Pi_T(x^\star) \rVert}$	Target direction. The user's normalized target direction, defined whenever $\lVert \Pi_T(x^\star) \rVert > 0$. It is fixed before execution and does not depend on i (§4.2).
$\rho \in (0, 1]$	Alignment threshold. The minimum directional similarity required between the outcome and the target direction $\boldsymbol{t}$. It serves as the Gate 2 parameter (§4.2).
A_ρ	Directionally admissible region. The set of outcomes satisfying the directional requirement: $A_\rho \coloneqq \left\{R \in \mathbb{R}^d : \lVert \Pi_T(R) \rVert > 0,\ \langle \frac{\Pi_T(R)}{\lVert \Pi_T(R) \rVert},\ \boldsymbol{t} \rangle \geq \rho \right\}$ (§4.2).
$\theta_Q > 0$	Quality bar. The minimum quality magnitude required by the institution. This is the raw quality threshold used in the main text and serves as the Gate 3 parameter (§4.2).
$(T, \boldsymbol{t}, \rho), (Q, \theta_Q)$	User declaration and institutional standard. The $(T, \boldsymbol{t}, \rho)$ specifies the user's directional requirements, whereas (Q, θ_Q) specifies the institution's quality requirements (§4.2).

Table A3. Gates and standing

Item	Definition
Gate 1	$a(x) \geq h_{\min}$. The admissibility condition applied to the specification itself. It is evaluated before any outcome exists and determines whether reliable decoding is possible (§5.1).
Gate 2	$R \in A_\rho$. The directional admissibility condition applied to the outcome once Gate 1 has been cleared. It determines whether the outcome belongs to the user-declared evaluable domain and is therefore constitutive of evaluability (§5.2).
Gate 3	$\| \Pi_Q(R) \| \geq \theta_Q$. The institutional quality sufficiency condition applied to the outcome once Gate 2 has been cleared. It determines whether the outcome satisfies the institution's quality requirement (§5.3); risk-adjusted and component-floor forms in Appendix C.
What the later gates receive	R, if $a(x) \geq h_{\min}$; $\perp$, if $a(x) < h_{\min}$. The later gates operate only on an admissible outcome and receive the absence marker otherwise (§5.1).
$\perp$	Marker for the absence of an admissible outcome. It indicates that reliable decoding has not occurred and therefore that no admissible outcome stands. It does not denote the system's output (§5.1).

Table A4. Institutions and scale

Symbol	Definition
$F(\tau) \coloneqq \Pr(i \geq \tau), \tau \in \mathbb{R}_+$	Distribution of instruction-environment quality across an institution's episodes: the share of episodes whose quality is at least τ; non-increasing in τ (§6).
F'	Distribution after an upward shift; first-order stochastic dominance, $F'(\tau) \geq F(\tau)$ for all $\tau \in \mathbb{R}_+$; the considered shift of §6, not a standing assumption (§6).
$c_{\max}$	The most a user will spend in an episode; the user is admissible when $c_{\min}(i) \leq c_{\max}$ (§6).
$S > 0$	Execution scale: more compute, more parallel attempts, longer runs (§7.1).
$R(S)$	The outcome at execution scale S; the gate conditions of §5 are read on $R(S)$ (§7.1); production set out in Appendix B.1.

Table A5. Production and scaling (Appendix B)

Symbol	Definition
$\Phi : X \to \mathbb{R}^d$	Decoding map: for an admissible specification, the system reliably recovers the intended meaning; Φ belongs to the decoding system and does not depend on i (B.1); below the bar no decoded intent stands (§5.1).
$\tilde{x} \coloneqq \Phi(x) \in \mathbb{R}^d$	Decoded intent, with $\| \tilde{x} \| > 0$ on admissible episodes; the content the system works from, parallel to the criterion $x^\star = \Gamma(x)$ (B.1).
$\Psi: (0,\infty) \times \mathbb{R}^d \to \mathbb{R}^d$	Execution map: $R(S) \coloneqq \Psi(S, \tilde{x})$; no structural form assumed beyond the regularities of B.2 (B.1).
$[S_{\min},\ S_{\max}]$	Operating range, $0 < S_{\min} \le S_{\max}$; the regularities of B.2 are assumed only on this range (B.2).
$\lambda_Q > 0, \varepsilon_Q \in [0,1)$	Per-unit rate and relative error bound of quality-magnitude scaling with its zero branch (B.2).
$\eta_Q(S, \tilde{x})$	Relative quality deviation, $\frac{\|\Pi_Q(R(S))\|}{\lambda_Q S \|\Pi_Q(\tilde{x})\|} - 1$, defined for $\| \Pi_Q(\tilde{x}) \| > 0$; the assumption reads $\| \eta_Q \| \le \varepsilon_Q$; audited in B.4 (B.2).
$\lambda_T > 0,\ \ \varepsilon_T \in [0,1)$	Per-unit rate and relative error bound of nondegenerate directional scaling with its zero branch (B.2).
$\eta_T(S, \tilde{x})$	Relative directional deviation, $\frac{\|\Pi_T(R(S)) - \lambda_T S \Pi_T(\tilde{x})\|}{\lambda_T S \|\Pi_T(\tilde{x})\|}$; the assumption reads $\eta_T \le \varepsilon_T$; audited in B.4 (B.2).
$u_S,\ u_0$	Unit directions of $\Pi_T(R(S))$ and of $\Pi_T(\tilde{x})$; $\| u_S - u_0 \| \le 2\varepsilon_T$ on the range (B.3).
$S_{\text{req}} \coloneqq \dfrac{\theta_Q}{\lambda_Q \ \| \Pi_Q(\tilde{x}) \|}$	Nominal required scale at which the quality magnitude reaches θ_Q; bracketed by the sufficient and necessary levels of B.3 (B.3).

Table A6. Gate 3 variants (Appendix C)

Symbol	Definition
$y \coloneqq \| \Pi_Q(R(S)) \|$	Quality magnitude of §5.3; shorthand of Appendix C.
$\Omega\colon \mathbb{R}_+ \to \mathbb{R}_+$	Risk-adjustment operator: non-decreasing and right-continuous, with $\Omega(y) \le y$, fixed by the institution before the episode.
$\Omega^{-1}(z) \coloneqq \inf\{\, y \ge 0 \colon \Omega(y) \ge z \,\}$	Generalized inverse, $+\infty$ when no such y exists; $\Omega(y) \ge \theta_Q^{\Omega} \iff y \ge \Omega^{-1}(\theta_Q^{\Omega})$.
$\theta_Q^{\Omega} > 0$	Risk-adjusted threshold, on the scale of $\Omega(y)$; in general numerically distinct from θ_Q and mutually exclusive with it within a single episode.
$Q_k \subseteq Q$	Evaluated quality dimension. One of an orthogonal decomposition $Q = Q_1 \oplus \cdots \oplus Q_m$ of the quality subspace, each one-dimensional, assessed against its own floor (Appendix C.2).
$F_k > 0$	Dimension floor. The minimum magnitude required on dimension Q_k; safety-critical dimensions carry high F_k (Appendix C.2).
$\lambda_{Q_k} > 0,\ \varepsilon_{Q_k} \in [0,1)$	Per-dimension rate and relative error of quality-magnitude scaling, B.2 read on Q_k (Appendix C.2).
$S_{req,k} \coloneqq \dfrac{F_k}{\lambda_{Q_k} \| \Pi_{Q_k}(\tilde{x}) \|}$	Nominal required scale for the k-th condition of the component-floor form. The scale for the full form is the largest of these over k, and $k = 0$ recovers the main-text scale of B.3 (Appendix C.2).
$g(R) \coloneqq \min\limits_{0 \le k \le m} \dfrac{\| \Pi_{Q_k}(R) \|}{F_k}$	Normalized quality margin, with the convention $Q_0 = Q$ and $F_0 = \theta_Q$. The component-floor form of Gate 3 is $g(R) \ge 1$ (Appendix C.2).

Appendix B. Production and Scaling Regularities

B.1 Production: decoding and execution

The main text treats the outcome R as a vector in $\mathbb{R}^d$ (§4.1) and defers how it is produced. Production has two stages. The first is decoding. For an admissible specification, one with $a(x) \geq h_{\min}$, the system reliably recovers the intended meaning (§3.1). We write $\tilde{x} \coloneqq \Phi(x) \in \mathbb{R}^d$ for this decoded intent, with $\| \tilde{x} \| > 0$. The decoding map Φ belongs to the decoding system, which is held fixed (§2), so Φ does not depend on the instruction-environment quality i. The notation is parallel to the criterion of §4.2: $x^\star = \Gamma(x)$ is the standard read off the specification, and $\tilde{x} = \Phi(x)$ is the content the system works from; both are fixed by x and free of i. The framework places no fidelity assumption linking Φ and Γ: whether the decoded content conforms to the declaration is exactly what Gate 2 tests on the outcome, and B.3 shows that this verdict is settled by the alignment of the decoded intent with $\boldsymbol{t}$, up to the audited margins. Below the bar decoding is not reliable and no decoded intent stands; those episodes are handled at the gate level (§5.1) and lie outside this appendix, which works on admissible episodes throughout. The second stage is execution. The system executes the decoded intent at a scale $S > 0$ (§7.1), and we write the outcome as

$$R(S) \coloneqq \Psi(S, \tilde{x}), \quad \Psi\colon (0,\infty) \times \mathbb{R}^d \to \mathbb{R}^d.$$

This appendix places no structural form on Ψ. Everything below rests on the regularities of B.2, which are stated directly on the components of $R(S)$ and are auditable (B.4).

B.2 Scaling regularities on the operating range

Fix an operating range $S \in [S_{\min}, S_{\max}]$ with $0 < S_{\min} \leq S_{\max}$. On this range we assume two nominal relations between the decoded intent $\tilde{x}$ and the components of the outcome $R(S)$, each with a bounded relative deviation. The relations are idealizations of how execution scales content, stated so that deviations from them are measurable (B.4); beyond the operating range neither is assumed. The rates and errors are properties of the executing system in its deployment setting, distinct from the instruction-environment quality of §2, and audited per deployment; the decoded intent contributes what it places on the evaluated axes. The derivations below use each projection only within its own subspace, so they need only that T and Q are disjoint, $T \cap Q = \{0\}$; the orthogonality assumed in §4.2 enters at the interpretive level, in reading a vanishing quality component of the decoded intent as the absence of quality content (§7.2).

Assumption (quality-magnitude scaling). There are $\lambda_Q > 0$ and $\varepsilon_Q \in [0,1)$ such that, for every admissible episode with $\| \Pi_Q(\tilde{x}) \| > 0$ and every S in the range,

$$\| \Pi_Q(R(S)) \| = \lambda_Q S \, \| \Pi_Q(\tilde{x}) \| \, (1 + \eta_Q), \qquad | \eta_Q | \leq \varepsilon_Q.$$

Here the relative deviation is the defined quantity

$$\eta_Q(S,\tilde{x}) \coloneqq \frac{\| \Pi_Q(R(S)) \|}{\lambda_Q S \| \Pi_Q(\tilde{x}) \|} - 1,$$

defined for $\| \Pi_Q(\tilde{x}) \| > 0$; for $\| \Pi_Q(\tilde{x}) \| = 0$ the relation reads $\| \Pi_Q(R(S)) \| = 0$.

Equivalently, the sandwich bounds

$$(1-\varepsilon_Q)\lambda_Q S \| \Pi_Q(\tilde{x}) \| \le \| \Pi_Q(R(S)) \| \le (1+\varepsilon_Q)\lambda_Q S \| \Pi_Q(\tilde{x}) \|$$

hold on the range. Execution scales the quality magnitude the decoded intent carries, at a per-unit rate λ_Q, up to the stated relative error.

Assumption (nondegenerate directional scaling). There are $\lambda_T > 0$ and $\varepsilon_T \in [0,1)$ such that, for every admissible episode with $\Pi_T(\tilde{x}) \neq 0$ and every S in the range,

$$\| \Pi_T(R(S)) - \lambda_T S \Pi_T(\tilde{x}) \| \le \varepsilon_T \lambda_T S \| \Pi_T(\tilde{x}) \|.$$

Here the relative deviation is the defined quantity

$$\eta_T(S,\tilde{x}) \coloneqq \frac{\| \Pi_T(R(S)) - \lambda_T S \Pi_T(\tilde{x}) \|}{\lambda_T S \| \Pi_T(\tilde{x}) \|},$$

and the assumption reads $\eta_T \le \varepsilon_T$. For admissible episodes with $\| \Pi_T(\tilde{x}) \| = 0$ the relation reads $\| \Pi_T(R(S)) \| = 0$ for every scale, the directional counterpart of the quality zero branch; on that branch the outcome has no directional component, so Gate 2 fails at every scale. By the reverse triangle inequality, the sandwich bounds

$$(1-\varepsilon_T)\lambda_T S \| \Pi_T(\tilde{x}) \| \le \| \Pi_T(R(S)) \| \le (1+\varepsilon_T)\lambda_T S \| \Pi_T(\tilde{x}) \|,$$

hold on the range. In particular, on the nonzero branch the directional component does not degenerate, $\| \Pi_T(R(S)) \| > 0$: the outcome retains a nonzero T-component, and Gate 2's comparison is well defined. The premise is placed on the decoded intent, $\Pi_T(\tilde{x}) \neq 0$, not on the outcome, so the nonzero component Gate 2 needs is delivered rather than presupposed. The bound is stated on the vector rather than on the magnitude because a magnitude bound alone would leave the direction free to rotate within T as S varies. The vector bound rules that out, and B.3 uses exactly this. The assumption secures Gate 2's comparison; it does not decide Gate 2, which remains the alignment test on the outcome against $\boldsymbol{t}$ (§5.2). The rates λ_T and λ_Q are in general distinct; each is estimated on its own relation (B.4).

B.3 Consequences

Three consequences. The two assumptions of B.2 yield three consequences on the operating range. First, the Gate 2 verdict is stable in S. Second, the Gate 3 condition reduces to a required level of S. Third, the requirements that remain as S grows are exactly those of Gates 1 and 2.

Scale stability of direction. On the nonzero branch, write $u_S \coloneqq \Pi_T(R(S))/\| \Pi_T(R(S)) \|$ for the outcome's direction and $u_0 \coloneqq \Pi_T(\tilde{x})/\| \Pi_T(\tilde{x}) \|$ for the direction of the decoded intent; u_S is well defined by the nondegeneracy of B.2. For any nonzero a and b,

$$\| \frac{a}{\| a \|} - \frac{b}{\| b \|} \| \leq \frac{2 \| a - b \|}{\| b \|},$$

holds, which follows by writing $a \| b \| - b \| a \| = a(\| b \| - \| a \|) + (a - b) \| a \|$ and bounding both terms.

We now apply this bound. Let $a \coloneqq \Pi_T(R(S))$ and $b \coloneqq \lambda_T S \ \Pi_T(\tilde{x})$. Because b is a positive scalar multiple of $\Pi_T(\tilde{x})$, its normalized direction is exactly u_0, so the bound compares u_S with u_0. The directional assumption of B.2 reads $\| \Pi_T(R(S)) - \lambda_T S \ \Pi_T(\tilde{x}) \| \leq \varepsilon_T \lambda_T S \ \| \Pi_T(\tilde{x}) \|$, and since $\| b \| = \lambda_T S \ \| \Pi_T(\tilde{x}) \|$, this is precisely

$$\| a - b \| \leq \varepsilon_T \| b \|.$$

Substituting this into the normalization bound, the factor $\| b \|$ cancels, and we obtain

$$\| u_S - u_0 \| \leq \frac{2 \| a - b \|}{\| b \|} \leq \frac{2\varepsilon_T \| b \|}{\| b \|} = 2\varepsilon_T.$$

The actual direction u_S therefore departs from the nominal direction u_0 by at most $2\varepsilon_T$, uniformly in S on the operating range.

This directional closeness carries over to the alignment that Gate 2 reads. By the linearity of the inner product and the Cauchy-Schwarz inequality, together with $\| \boldsymbol{t} \| = 1$,

$$| \langle u_S, \boldsymbol{t} \rangle - \langle u_0, \boldsymbol{t} \rangle | = | \langle u_S - u_0, \ \boldsymbol{t} \rangle | \leq \| u_S - u_0 \| \| \boldsymbol{t} \| \leq 2\varepsilon_T$$
for all $S \in [S_{\min}, S_{\max}]$.

The alignment of the outcome thus differs from that of the nominal direction by at most $2\varepsilon_T$, whatever the scale. Two implications follow:

$$\langle u_0, \boldsymbol{t} \rangle \geq \rho + 2\varepsilon_T \Rightarrow R(S) \in A_\rho \text{ at every } S \text{ in the range,}$$
$$\langle u_0, \boldsymbol{t} \rangle < \rho - 2\varepsilon_T \Rightarrow R(S) \notin A_\rho \text{ at every } S \text{ in the range.}$$

The sufficient condition can be met only when $\rho + 2\varepsilon_T \leq 1$, since the alignment never exceeds 1; when the audited band is wider than the room the declaration leaves, the first implication holds vacuously. Symmetrically, the failure condition can bind only when $\rho - 2\varepsilon_T > -1$, since the alignment never falls below -1. Outside a band of width $4\varepsilon_T$ around ρ, the Gate 2 verdict does not depend on S (§7.2).

The scale Gate 3 requires. Let $\| \Pi_Q(\tilde{x}) \| > 0$. By the quality sandwich bounds of B.2, the condition $\| \Pi_Q(R(S)) \| \geq \theta_Q$ holds whenever

$$S \geq \frac{\theta_Q}{(1-\varepsilon_Q)\lambda_Q \| \Pi_Q(\tilde{x}) \|},$$

and can hold only if

$$S \geq \frac{\theta_Q}{(1+\varepsilon_Q)\lambda_Q \| \Pi_Q(\tilde{x}) \|}.$$

The nominal required scale sits between the two,

$$S_{\text{req}} := \frac{\theta_Q}{\lambda_Q \| \Pi_Q(\tilde{x}) \|}.$$

If $\| \Pi_Q(\tilde{x}) \| = 0$, the relation of B.2 gives $\| \Pi_Q(R(S)) \| = 0$ at every S, and Gate 3 stays unmet at every scale (§5.3, §7.2).

The binding requirement. On the operating range, joint admissibility (§5.4) requires $a(x) \geq h_{\min}$, $R(S) \in A_\rho$, and the Gate 3 condition. The first does not involve S. The second is settled by $\langle u_0, \boldsymbol{t} \rangle$ up to the $2\varepsilon_T$ margin, uniformly in S. The third holds once S clears the sufficient level above and only if it clears the necessary one; for $\| \Pi_Q(\tilde{x}) \| > 0$ and a sufficient level that falls inside the operating range, it is therefore met and stays met as S grows. As S grows within the operating range, the binding constraints that remain are the bar and the direction (§7.3).

B.4 Audit quantities

The assumptions of B.2 are nominal relations, and §9.1 treats them as auditable. The audit asks whether, on episodes an institution actually runs, the deviations stay within the stated bounds; it is what licenses the margins of B.3 in practice. This appendix fixes the audit at the level the framework needs: what is recorded, what is estimated, and what verdict follows. The operational detail beneath that level, the concrete choice of estimation rule, its tie-breaking, the tolerance schedule, and the design of the reference sample, is implementation for the institution and is not part of the framework's claims.

Recording. Take a reference sample of admissible episodes executed at scales within the operating range under one fixed execution protocol, with the scale S, the decoded intent $\tilde{x}$ and the outcome $R(S)$ recorded for each. Recording the decoded intent does not require access to model internals. §2 defines the decoding system as the parser, the validation rules, and the interpretation layer, and the decoded intent is the output of that pipeline. Operationally, decoding counts as reliable for an episode when this pipeline emits its validated intent object. The audit therefore instruments it at that interface, for instance as the validated request object that the execution stage receives. Where a deployment exposes no such artifact, the audit cannot be run, and the margins of B.3

cannot be calibrated for it. From these, together with the fixed subspaces and the protocol, every quantity entering B.2 is computable episode by episode.

Estimation. The rates are estimated on the sample by a rule fixed in advance, each from its own relation: λ_Q from the relation between $\| \Pi_Q(R(S)) \|$ and $S \| \Pi_Q(\tilde{x}) \|$, and λ_T from the vector relation between $\Pi_T(R(S))$ and $S\Pi_T(\tilde{x})$, for instance by least squares through the origin in each case.

Check. With the rates in hand, the deviations η_Q and η_T of B.2 are computed, where they are defined, for each episode and each recorded scale. The assumptions hold on the sample when $| \eta_Q | \leq \varepsilon_Q$ and $\eta_T \leq \varepsilon_T$ throughout. Deviations beyond the bounds locate where the nominal relations fail, by episode and by scale.

Calibration. Read in the other direction, the sample maxima of $| \eta_Q |$ and of η_T are working values for ε_Q and ε_T. These are the error terms that enter the margins of B.3. The institution fixes in advance the largest relative error it will accept, strictly below 1. The recorded rate is the one the predeclared estimation rule returns; the maxima above are read at that rate. If the recorded maximum exceeds that tolerance, the scaling assumption fails on the sample at that tolerance, and the margins of B.3 are not available for it. The error ε_T sets the width $4\varepsilon_T$ of the band within which the Gate 2 verdict can depend on scale. The error ε_Q sets the gap between the sufficient and necessary levels of S at Gate 3. A tighter audited sample therefore yields tighter margins. The audit calibrates the bounds only on the range the sample covers; beyond the operating range nothing is claimed (B.2).

Appendix C. Two Variants of Gate 3: Risk Adjustment and Component Floors

The main text places Gate 3 on the aggregate quality magnitude, $\| \Pi_Q(R(S)) \| \geq \theta_Q$ (§5.3). This appendix states two ways an institution may strengthen this test while keeping it a scalar sufficiency condition. The first reshapes the bar on the aggregate magnitude through a risk-adjustment operator. The second keeps the aggregate bar and imposes, in addition, a floor on each dimension the institution evaluates. This ensures that a high aggregate cannot conceal a low value on a dimension the institution treats as critical. Both variants are read only once Gate 2 holds, and both leave Gates 1 and 2 untouched. Throughout, write $y := \| \Pi_Q(R(S)) \|$ for the aggregate quality magnitude of §5.3.

C.1 The risk-adjusted form

In some settings the institution does not credit raw quality magnitude directly, but discounts it under uncertainty and harm exposure and places its bar on the discounted value. Let $\Omega: \mathbb{R}_+ \to \mathbb{R}_+$ be non-decreasing, right-continuous, and satisfying $\Omega(y) \leq y$, fixed by the institution before the episode, and let $\theta_Q^\Omega > 0$ be a threshold on the scale of $\Omega(y)$. The risk-adjusted form of Gate 3 is

$$\Omega(y) \geq \theta_Q^\Omega.$$

Like the raw form, it is read only once Gate 2 holds. The operator Ω expresses how strongly the institution discounts raw magnitude. The threshold θ_Q^Ω and the raw threshold θ_Q apply to different quantities, the raw and the discounted magnitude, and are in general numerically distinct, and a single episode uses one form or the other, not both. Because Ω is monotone and fixed in advance, the discounted value is comparable across the episodes evaluated under the same operator and threshold, and the audit of B.4 is unchanged, since the regularities of B.2 are stated on raw magnitudes. The risk adjustment moves the bar, not the measured quantities.

The risk-adjusted form is the raw form with the bar relocated. Define the generalized inverse

$$\Omega^{-1}(z) := \inf \{ y \geq 0 : \Omega(y) \geq z \},$$

with $\Omega^{-1}(z) = +\infty$ when no such y exists. For non-decreasing, right-continuous Ω,

$$\Omega(y) \geq \theta_Q^\Omega \iff y \geq \Omega^{-1}(\theta_Q^\Omega).$$

One direction follows from monotonicity, with right-continuity securing the boundary point itself; the other is the definition of the infimum. The risk-adjusted gate is therefore the raw gate with the bar moved from θ_Q to $\Omega^{-1}(\theta_Q^\Omega)$. The move strengthens

the raw gate exactly when the relocated bar sits at or above the raw one; the institution is taken to set the operator and its threshold so that it does.

Substituting this moved bar into B.3, the sufficient level, the necessary level, and the nominal required scale become

$$S \geq \frac{\Omega^{-1}(\theta_Q^{\Omega})}{(1-\varepsilon_Q)\lambda_Q \parallel \Pi_Q(\tilde{x}) \parallel},$$
$$S \geq \frac{\Omega^{-1}(\theta_Q^{\Omega})}{(1+\varepsilon_Q)\lambda_Q \parallel \Pi_Q(\tilde{x}) \parallel}, \quad S_{req} = \frac{\Omega^{-1}(\theta_Q^{\Omega})}{\lambda_Q \parallel \Pi_Q(\tilde{x}) \parallel},$$

for $\parallel \Pi_Q(\tilde{x}) \parallel > 0$ and finite $\Omega^{-1}(\theta_Q^{\Omega})$. Nothing else in Appendix B changes. If θ_Q^{Ω} exceeds the range of Ω, then $\Omega^{-1}(\theta_Q^{\Omega}) = +\infty$ and Gate 3 is unmet at every scale. This is an institutional shut-off, available by the choice of θ_Q^{Ω}, and it leaves Gates 1 and 2 exactly where they were.

C.2 The component-floor form

The bar θ_Q is placed on the aggregate magnitude $\parallel \Pi_Q(R) \parallel$, so a high aggregate may coexist with a low value on any single dimension. An institution usually guards against this by requiring two things at once: that the aggregate clear its standard, and that no single dimension fall below a minimum of its own. The component-floor form states this conjunction. It also matches the regulated practice described in §9.2.1, where quality is judged on named dimensions such as sensitivity, specificity, and subgroup calibration.

Let Q decompose into the evaluated dimensions $Q_1, \ldots, Q_m \subseteq Q$, each one-dimensional, taken as an orthogonal decomposition of the quality subspace, and let each dimension carry a floor $F_k > 0$, fixed before the episode. Gate 3 in its full form keeps the aggregate bar of §5.3 and adds the per-dimension floors:

$$\parallel \Pi_Q(R(S)) \parallel \geq \theta_Q \quad \text{and} \quad \parallel \Pi_{Q_k}(R(S)) \parallel \geq F_k \text{ for every } k = 1, \ldots, m.$$

The aggregate bar is retained, not replaced; the floors are imposed in addition to it. A safety-critical dimension is simply one whose floor F_k is set high; nothing else sets it apart. Both parts are conditions on the outcome $R(S)$, so both depend on the execution scale S and on the relevant scaling rates, exactly as the aggregate bar alone does. What the floors add is that they are read on each dimension before aggregation and conjunctively. A high aggregate cannot stand in for a deficient dimension, and an outcome that fails any one floor does not clear Gate 3 whatever its aggregate value.

It is convenient to gather every condition into a single scalar. Writing the aggregate as index 0, with $Q_0 := Q$ and floor $F_0 := \theta_Q$, define the normalized margin

$$g(R) := \min_{0 \leq k \leq m} \frac{\parallel \Pi_{Q_k}(R) \parallel}{F_k}.$$

The full form of Gate 3 is then exactly $g(R(S)) \geq 1$, since the minimum is at least 1 exactly when the aggregate and every dimension clear their respective standards. The term $k = 0$ is the main-text bar; the terms $k \geq 1$ are the added floors. Gate 3 thus remains a single scalar test, now read on g in place of y.

For the consequences in Appendix B.3, we assume the quality-magnitude scaling of B.2 on the aggregate, as in the main text, and on each dimension Q_k, with its own rate $\lambda_{Q_k} > 0$ and relative error $\varepsilon_{Q_k} \in [0,1)$. For index 0, the rate and error are those of B.2. Under the orthogonal decomposition, the two levels are tied by the identity $\| \Pi_Q(R) \|^2 = \sum_{k=1}^{m} \| \Pi_{Q_k}(R) \|^2$. Assuming the aggregate relation and the component relations together is therefore a joint restriction on the rates and errors, not an automatic consequence of orthogonality. The audit of B.4, run on the aggregate and on each dimension from the same sample, checks exactly this joint restriction. One consequence is worth noting: when every floor is met, the identity lifts the aggregate magnitude to at least the root sum of squares of the floors, so the aggregate bar adds a separate requirement only when θ_Q exceeds that level. For each k with $\| \Pi_{Q_k}(\tilde{x}) \| > 0$, the sandwich bounds of B.2 give

$$(1 - \varepsilon_{Q_k})\lambda_{Q_k} S \| \Pi_{Q_k}(\tilde{x}) \| \leq \| \Pi_{Q_k}(R(S)) \| \leq (1 + \varepsilon_{Q_k})\lambda_{Q_k} S \| \Pi_{Q_k}(\tilde{x}) \|,$$

so the k-th condition $\| \Pi_{Q_k}(R(S)) \| \geq F_k$ holds whenever $S \geq F_k / [(1 - \varepsilon_{Q_k})\lambda_{Q_k} \| \Pi_{Q_k}(\tilde{x}) \|]$ and can hold only if $S \geq F_k / [(1 + \varepsilon_{Q_k})\lambda_{Q_k} \| \Pi_{Q_k}(\tilde{x}) \|]$, with the nominal required scale for that condition,

$$S_{req,k} := \frac{F_k}{\lambda_{Q_k} \| \Pi_{Q_k}(\tilde{x}) \|},$$

sitting between the two. The case $k = 0$, with $F_0 = \theta_Q$, $Q_0 = Q$, and rate λ_Q, recovers the main-text Gate 3 scale of B.3. Because the full form requires the aggregate and every dimension to clear their standards, its nominal required scale is set by the most demanding condition,

$$S_{req} = \max_{0 \leq k \leq m} S_{req,k},$$

and likewise the sufficient and necessary levels are each the maximum over k of the corresponding levels. Guaranteed passage is read off the sufficient level, and the binding dimension may differ across the three. A high floor raises $S_{req,k}$ for that dimension, so a safety-critical dimension tends to be the binding one. Since each per-condition level depends on the fixed quantities F_k, λ_{Q_k}, and $\| \Pi_{Q_k}(\tilde{x}) \|$ and not on S, the nominal binding condition does not change as S grows; within the relative error bands, the realized condition may differ.

If $\| \Pi_{Q_k}(\tilde{x}) \| = 0$ for some dimension, then the scaling of B.2 forces $\| \Pi_{Q_k}(R(S)) \| = 0 < F_k$ at every S, so that condition fails at every scale and the full form is unmet however large S grows. This is the dimension-level counterpart of the shut-off

of C.1, and it records, in the present terms, the limit of §7.2: execution scale raises a dimension only when the decoded intent carries it. The point bears on safety-critical dimensions in particular, since a high floor cannot be met by scale alone unless the specification and declaration first place the dimension in the decoded intent.

The component-floor form is, in both its parts, a condition on the outcome, and is therefore dependent on scale and on the rates, exactly as the aggregate bar is. We do not claim that it is independent of scale. Greater scale helps a dimension that is present in the decoded intent to clear its floor, just as it helps the aggregate clear θ_Q. The protection the floors add is of a different kind. Because each floor is read before aggregation and conjunctively, no aggregate magnitude can mask a deficient dimension. The audit of B.4 extends without change, estimating λ_{Q_k} and ε_{Q_k} on each dimension from the same reference sample. The acceptance tolerance applies to each audited relation; the component package is licensed only when the aggregate relation and every required component relation meet their tolerances, a verdict separate from whether any particular outcome clears its Gate 3 floor.

C.3 Relation of the two variants

The two variants act on different objects and may be used separately or together. The risk-adjusted form reshapes the bar on the aggregate magnitude y; the component floors act on the dimensions before they are aggregated and are imposed in addition to the aggregate bar. When both are used, the risk adjustment applies to the aggregate condition while the floors apply to the dimensions. The full Gate 3 is the conjunction of all of them, again expressible as a single normalized margin, formed as in C.2 with the aggregate ratio taken on the discounted magnitude against its own threshold. The institution fixes any order of evaluation it requires; the clauses are separately stated, and the verdict is their conjunction in any order; separately, the scaling relations behind them remain subject to the joint restriction of C.2.

Neither variant alters the structure of Appendix B. For the component-floor variant this holds under the joint restriction stated in C.2. The scale stability of direction (B.3) and the binding requirement read as before, with the strengthened Gate 3 in place of the raw one. The asymmetry of §7 therefore stands under both variants: execution scale reaches the quality magnitude, in whichever form Gate 3 takes, and reaches neither the bar of Gate 1 nor the direction of Gate 2.

https://doi.org/10.23104/ME.e9