

AI Companies as Cybernetic Systems

Author: Anonymous

Source: <https://ai-cybernetics.grok.me/>

Abstract

An AI company is modeled as an open, resource-dependent system whose observations, learning procedures, deployment actions, and governance rules close several feedback loops. A contextual ∞ -topos supplies a language for states, admissibility certificates, and coherent identification of implementations; categories and lenses retain the direction of irreversible operations. We give elementary results on certified update preservation, descent, the limits of observational regulation, and the insufficiency of semantic equivalence for identifying learning dynamics. At the differentiable level, an optimizer-dependent neural tangent kernel is derived exactly along gradient flow. Parameter jets, a neural tangent hierarchy, and an input-jet tangent kernel describe distinct extensions. Their order is independent of homotopy truncation. Every extensional optimizer on smooth objectives factors through their global holonomic jet representation, with runtime and oracle state retained; factorization through a jet at one point has a separate locality condition. A formal panoptic hypothesis specifies trace coverage and inequalities between observation, intervention, and governance capabilities. The construction combines identity types, univalence, and jets into a research program for the cybernetics of AI firms, without treating economic concentration, universal surveillance, or a fixed-kernel description of every neural system as mathematical consequences.

Keywords. Organizational cybernetics; homotopy type theory; ∞ -topos; lenses; neural tangent kernel; jets; constrained updates.

1 Scope and the cybernetic question

The company-level application is a proposed synthesis. Its starting hypotheses are that employee and customer traces can both become productive training data, that feedback changes the behavior being measured, and that data integration can create incentives for consolidation. These are conditional empirical and institutional claims. The mathematical results below concern explicitly specified models; they do not establish that every actual AI company collects all available traces or that a single company must emerge.

Wiener’s control-and-communication perspective [35], Ashby’s requisite variety [1], and Beer’s organizational cybernetics [2] motivate the separation of operation, coordination, adaptation, and policy. Modern categorical cybernetics supplies compositional interfaces and controllers [4; 27]; categorical learning supplies parametrized maps, learners, and reverse differentiation [9; 7]. Higher groupoids add a precise account of **witnessed equivalence and its coherence**. They are not substitutes for economic evidence or for differential equations.

Two further strands connect architecture and institutions: categorical deep learning studies algebraic architecture constraints [13; 12], while compositional game theory models interacting decision-makers [14]. Hedges explicitly proposes categorical-cybernetic analysis of AI-mediated markets and supply chains [16]. That proposal is a research agenda, not an established empirical law of AI companies.

Central question. Which observations and interventions can a firm perform, which specifications must those interventions preserve, and how do learning, governance, and the environment respond to one another? A model is useful when it makes these questions executable or falsifiable, rather than merely relabeling a company as a category.

2 Identity, higher groupoids, and contextual semantics

2.1 Foundational conventions and notation

A **category** has objects, arrows with specified source and target, identity arrows, and associative composition. A **functor** preserves these data. A natural transformation consists of arrows between two functors' values which commute with every source-category arrow. "Small" means that the objects and arrows belong to the chosen set-sized universe. C^{op} reverses arrows, and $\text{Fun}(C, D)$ denotes functors and their natural transformations. A **groupoid** is a category whose arrows are invertible. An **endomorphism** has equal source and target; an **automorphism** is an invertible endomorphism. A terminal object 1 has one arrow from each object, and an initial object 0 has one arrow to each object, with the corresponding mapping spaces contractible in higher categories. Products and limits are defined by their universal mapping properties.

For a precise higher model, a **simplicial set** is a functor $\Delta^{\text{op}} \rightarrow \text{Set}$, where Δ has finite ordered sets $[n] = \{0, \dots, n\}$ and order-preserving maps. The horn $\Lambda_i^n \subseteq \Delta^n$ consists of all faces except the i th. An ∞ -category can be modeled by a simplicial set filling every inner horn ($0 < i < n$); a Kan complex fills all horns and models an ∞ -groupoid. The ∞ -category $\mathcal{S}$ of spaces means homotopy types, modeled by Kan complexes with weak homotopy equivalences inverted. A weak homotopy equivalence induces a bijection of components and isomorphisms of all homotopy groups. The group $\pi_k(X, x)$ consists of based homotopy classes of maps from the k -sphere to (X, x) , for $k \geq 1$. All higher limits and pullbacks below are homotopy-coherent limits [25]. The **core** of a category retains its objects and equivalences.

In the internal type language, a universe U classifies the chosen small types. A dependent family $B: A \rightarrow U$ assigns a type to each $a: A$. Its dependent sum $\sum_{a:A} B(a)$ contains pairs (a, b) ; its dependent product $\prod_{a:A} B(a)$ contains sections assigning such a b to every a . A **witness** is an inhabitant of the stated type. A type X is **contractible** when there exist $c: X$ and paths $c = x$ for every $x: X$. For $f: A \rightarrow B$, its fiber at b is $\sum_{a:A} \text{Id}_B(f(a), b)$. An **equivalence** is a map with contractible fibers; $\text{Equiv}(A, B)$ is the type of these maps and witnesses. A homotopy from f to g is a section of $\prod_{a:A} \text{Id}_B(f(a), g(a))$. Intensional equality uses such identity types, distinct from the definitional equality used in evaluation rules. The type Prop consists of mere propositions, defined below; $\text{Bool} = 1 + 1$ has two alternatives. Decidability of a proposition P means a witness of $P + (P \rightarrow 0)$, not just the formation of its type [32].

A presheaf of spaces is a functor $C^{\text{op}} \rightarrow \mathcal{S}$. A **site** is a small category with a Grothendieck topology: specified covering sieves, where a sieve is a collection of arrows to an object closed under precomposition. The axioms require the maximal sieve to cover, stability under pullback, and transitivity of covering. A sheaf has compatible local sections which glue uniquely, or up to coherent homotopy in the space-valued case. Descent data include sections on a cover and agreements on all iterated overlaps. The Čech nerve lists these overlaps; its limit is the space of such coherent data. An ∞ -topos is an accessible left-exact localization of a presheaf ∞ -category. Here a localization has a fully faithful right adjoint; left exact means preserving finite limits, and accessible means preserving κ -filtered colimits for some regular cardinal κ . A diagram is κ -filtered when every subdiagram with fewer than κ arrows has a compatible cocone. A geometric morphism $f: E \rightarrow F$ is an adjoint pair $f^*: F \rightleftarrows E: f_*$ with f^* left exact [25; 22].

An **open system** exchanges inputs and outputs across a chosen boundary. A **controller** chooses inputs from observations and possibly memory; **feedback** feeds outputs into subsequent inputs. **Governance** specifies who may change controllers, objectives, interfaces, and admissibility rules. **Resources** are stocks or capacities consumed or replenished by operations. These terms acquire explicit maps, state variables, and authorization relations in the constructions below.

2.2 The identity tower

Work in intensional dependent type theory, with a chosen univalent universe when univalence is used. For $x, y: A$, an identity witness has type $p: \text{Id}_A(x, y)$; for $p, q: \text{Id}_A(x, y)$ there is a further type $\text{Id}_{\text{Id}_A(x, y)}(p, q)$, and the process iterates. Reflexivity and identity elimination induce inverses and composition together with higher coherence. Identity elimination says that a family $P(x, y, p)$ indexed by $p: \text{Id}_A(x, y)$ has a section for all (x, y, p) once its reflexive instances $P(x, x, \text{refl}_x)$ have sections, with the corresponding reflexive computation rule. It is the dependent substitution rule used for equality witnesses. This identity tower admits a weak ω -groupoid structure [24; 33]. It is stronger information than a binary relation of indistinguishability.

The groupoid model demonstrates that uniqueness of identity proofs is not a general consequence of intensional type theory [17]. Martin-Löf’s foundational presentation [26] and Voevodsky’s homotopy λ -calculus notes [34] provide foundational background. Grothendieck’s homotopical perspective [15] connects suitable models of higher groupoids to homotopy types.

Definition 2.1 Truncation levels

A type is a mere proposition, or (-1) -type, when any two inhabitants are equal. It is a set, or 0-type, when each of its identity types is a mere proposition. Inductively, it is an $(n+1)$ -type when its identity types are n -types. The reflection into n -types is written $\tau_{\leq n}$.

Its universal property is $\text{Map}(\tau_{\leq n} A, B) \simeq \text{Map}(A, B)$ for every n -type B ; Map denotes the mapping space. The (-2) -types are the contractible types. Calling a proposition “yes or no” does not make it a decidable Boolean: $P \vee \neg P$ is an additional assertion in constructive logic. Likewise, “discrete” in this note means homotopy 0-truncated; it need not mean discrete as a topological or smooth space. These distinctions matter when software must produce an actual certificate.

2.3 A topos of organizational contexts

Definition 2.2 Contextual firm semantics

Choose a small site (C, J) . Objects of C describe contexts such as a team, project, deployment, jurisdiction, or accessible interface; arrows describe restriction of context. The topology J specifies which compatible local views jointly cover a context. Set

$$E = \text{Sh}_\infty(C, J) \subseteq \text{Fun}(C^{\text{op}}, S).$$

Firm states, interfaces, specifications, and equivalence witnesses are objects or structured objects internal to E .

This is an actual ∞ -topos, obtained by accessible left-exact sheafification [25]. The selected site is a presentation of the semantics. Neither the bare organization nor an arbitrary category of neural networks automatically forms a topos. A 0-truncated part supports ordinary sheaf logic and the Mitchell–Bénabou language [22]; suitable universes and dependent type-theoretic structure support the higher identity language [32]. An arbitrary such model does not automatically supply every desired computational rule for a proof assistant.

A state in context U is a generalized element $U \rightarrow S$. A global state is $1 \rightarrow S$. A point of a topos, a geometric morphism $\text{Set} \rightarrow E$ in the ordinary setting, is a different notion. Local inhabitation, a global section, and an executable construction must not be conflated: a semantic witness becomes an effective procedure only through an appropriate constructive interpretation or implementation.

For differentiation, choose smooth parameter spaces and derivative operations. A smooth manifold is a Hausdorff, second-countable space covered by Euclidean charts with smooth transition maps; smooth maps are smooth in those charts. A concrete option is to replace the context site by a product with a site of Cartesian spaces and smooth maps, with the corresponding product topology, and to use smooth finite-dimensional charts for parameter objects. Probability is extra structure too. A standard Borel space is measurably isomorphic to the Borel space of a complete separable metric space. A Markov kernel $K: X \rightsquigarrow Y$ assigns a probability measure $K(x, -)$ to each x , with $x \mapsto K(x, B)$ measurable for every measurable B . Composition is $(L \circ K)(x, B) = \int_Y L(y, B) K(x, dy)$. A Markov category packages such composition in a symmetric monoidal category, with compatible commutative copy maps $X \rightarrow X \otimes X$ and delete maps $X \rightarrow 1$, the latter natural and the unit terminal [11]. The tensor denotes parallel composition; copy need not be natural for stochastic maps. Neither calculus nor probability follows merely from the word “topos.”

Remark 2.3 Three independent resolutions

Homotopy order n , differential jet order r , and access/capability resolution κ answer different questions. Here κ labels an observation-and-action interface, not a numerical dimension; it is unrelated to the cardinal used in the accessibility convention. A representable smooth sheaf is the functor $U \mapsto \text{Hom}(U, M)$ for a smooth space M . A smooth representable parameter space can be a 0-truncated sheaf while possessing derivatives of all orders. A finite groupoid can have nontrivial identity structure without possessing any calculus. Increasing bandwidth or acquiring a vehicle changes accessible observations and actions; it does not by itself increase a homotopy truncation level.

3 A firm as a certified open system

3.1 States and admissible transitions

Let the ambient state object be

$$S_0 = H \times D \times \Theta \times O \times C \times R \times G,$$

where H describes human roles and capabilities, D data and provenance, Θ models and parameters, O optimizer/runtime memory, C controllers, R resources, and G governance. Products may be replaced by dependent sums when, for example, an optimizer’s state depends on the architecture. Specify a predicate $V: S_0 \rightarrow \text{Prop}$ and define

$$S = \sum_{s: S_0} V(s).$$

Thus an inhabitant of S contains both a proposed state and its validity witness. The predicate may include schema equations, authorization, resource bounds, deployment tests, and declared contractual conditions. Their concrete meaning must be supplied; a proof of the chosen predicate is not proof of every desirable social outcome.

Definition 3.1 Certified update

For an admissible input family $A: S_0 \rightarrow \text{Type}$, a certified update is a procedure which, given $s, p: V(s)$, and $a: A(s)$, returns $(s', p'): \sum_{s': S_0} V(s')$. An optional transition relation $T(s, a, s')$ can be included in the certificate. When a total implementation branches on acceptance, it must supply a decision procedure for its chosen checks; arbitrary semantic consistency is not thereby decidable.

Proposition 3.2 Preservation under iteration

Suppose an initial state has a witness $p_0: V(s_0)$ and every accepted update is certified. Then every state reached by a finite sequence of accepted updates has a witness of V . A rejected update that returns the original certified state also preserves V .

Proof. The base case is p_0 . At each successor step the update supplies the next witness. Rejection reuses the existing witness. Induction on the sequence length gives the claim. No uniqueness of certificates or excluded middle is required.

This formalizes constrained database updates and speculative changes. A valid rollback concerns the modeled state: deletion of already transmitted information, or reversal of an irreversible physical act, requires additional operations and cannot be assumed from this proposition.

For an ordinary categorical database schema B , an instance is a functor $D: B \rightarrow \text{Set}$ [31]. Functoriality enforces the schema's declared path equations. Authorization, statistical validity, and cross-table conditions beyond those equations must be specified separately. For a schema map $u: B \rightarrow B'$, restriction Δ_u is precomposition by u ; when they exist, its left and right adjoints Σ_u and Π_u are the left and right Kan extensions. An adjunction is a natural equivalence between the corresponding mapping spaces. These migration constructions are not automatic proofs of the additional conditions.

Proposition 3.3 Descent of certified operations

Let $\{U_i \rightarrow U\}$ be a cover. Suppose local states, transition maps, and certificates form a homotopy-coherent descent datum in their respective sheaf objects. Then they determine a certified operation over U , unique up to the contractible space prescribed by descent for that datum. **Proof.** Apply the sheaf equivalence between sections over U and the limit of sections over the Čech nerve to the structured object of certified operations. This object is assembled from dependent sums, products, and the specified certificate family. The fiber over the given descent datum is contractible.

Independent local approvals are not enough. Agreement on shared data, interfaces, and higher coherence is part of the premise. For example, two resource requests may each fit a budget locally and exceed it jointly; they do not form a valid global resource certificate.

3.2 Interfaces, lenses, and feedback

A deterministic open module consists of a state object S , an output map $o: S \rightarrow Y$, and an update $u: S \times X \rightarrow S$. Closing an environment/controller map $c: Y \rightarrow X$ gives

$$s_{t+1} = u(s_t, c(o(s_t))).$$

More generally, c has its own state and the environment contributes inputs; the full product state must then be retained. In a stochastic interpretation, updates are Markov kernels and composition uses their integral composition law.

For interfaces (A^+, A^-) and (B^+, B^-) , a simple lens is a pair

$$L^+: A^+ \rightarrow B^+, \quad L^-: A^+ \times B^- \rightarrow A^-.$$

The second component carries responses, requested changes, or cotangent signals back through the interface. For $L: A \rightarrow B$ and $M: B \rightarrow C$, composition is

$$(M \circ L)^+ = M^+ \circ L^+, \quad (M \circ L)^-(a, c) = L^-(a, M^-(L^+(a), c)).$$

This is the elementary bidirectional structure; dependent lenses permit response types to depend on the current output. Reverse differentiation sends a smooth map $f: A \rightarrow B$ to forward evaluation and cotangent pullback $(a, \bar{b}) \mapsto Df_a^T \bar{b}$ in Euclidean coordinates. The chain rule makes this construction compositional. An optimizer closes the parameter port, as in categorical learning [7]. The tangent space at a smooth point a consists of derivations $v: C_a^\infty \rightarrow R$ on smooth germs, linear over R and satisfying $v(fg) = f(a)v(g) + g(a)v(f)$; these are infinitesimal directions. Its dual, the cotangent space, contains linear functionals on those directions. A parameter port is the part of the interface carrying parameters and their requested changes. To close it is to supply an update law for those changes.

Parametric maps and architecture constraints. A layer with parameters P is a map $f: P \times A \rightarrow B$. Composing it with $g: Q \times B \rightarrow C$ gives a map with parameters $Q \times P$, $(q, p, a) \mapsto g(q, f(p, a))$. A reparametrization $\rho: P \rightarrow P'$ relates f to $f': P' \times A \rightarrow B$ when $f = f' \circ (\rho \times \text{id}_A)$. Such maps form the elementary structure behind the bicategorical *Para* construction; invertible reparametrizations give only its equivalence part. A 2-category has objects, 1-arrows, and 2-arrows between 1-arrows, with compatible vertical and horizontal composition; a bicategory weakens associative and unit laws to coherent invertible 2-arrows. An optic represents bidirectional behavior through a residual object M , with maps $A^+ \rightarrow M \times B^+$ and $M \times B^- \rightarrow A^-$. Representatives are identified by the relations induced by changing residual objects through compatible maps. Thus residual state is retained for the backward response [12].

Algebraic architecture constraints can also be expressed through monads. In an ordinary category, a monad consists of an endofunctor T and natural transformations $\eta: \text{Id} \rightarrow T$ and $\mu: T^2 \rightarrow T$ satisfying

$$\mu \circ T\eta = \text{id}_T = \mu \circ \eta T, \quad \mu \circ T\mu = \mu \circ \mu T.$$

An algebra $a: TA \rightarrow A$ satisfies $a \circ \eta_A = \text{id}_A$ and $a \circ \mu_A = a \circ T(a)$. These equations constrain how composed operations are implemented. Categorical deep learning lifts this algebraic viewpoint to appropriate 2-categories of parametric maps, relating model constraints to architecture implementations [13]. The lift and the chosen ambient category are substantive parts of the construction. Architecture algebra does not by itself specify loss, optimizer memory, or a firm's governance objective.

The temporal operations in (2) are generally irreversible. They belong to a category of transitions or systems. Only equivalences belong to its maximal higher groupoid. In an $(\infty, 1)$ -category higher morphisms are invertible; its ordinary 1-morphisms need not be. Reflexivity, symmetry, and transitivity of a relation alone do not provide the entire identity tower.

An executable trace $\gamma = (a_1, \dots, a_k)$ with specified state semantics denotes the composite $\llbracket \gamma \rrbracket = u_{a_k} \circ \dots \circ u_{a_1}$ when the types match. This makes a repeatable operation precise. An untyped log or a trace with missing causal context does not canonically denote an endomorphism. Recursive improvement can be modeled by an endomorphism of a type of certified agents; iteration additionally requires preserved admissibility and realizable operations, not merely a self-description.

Reading figure: loop. See the retained source snapshot.

4 Univalence, equivalence, and what equivalence forgets

Univalence asserts that the canonical map

$$\text{Id}_U(A, B) \longrightarrow \text{Equiv}(A, B)$$

is an equivalence [32]. It permits transport of type-dependent properties and structures along a supplied equivalence. The identity structure includes automorphisms; univalence does not reduce all equivalence witnesses to one proof. Nor is it derivable just from a verbal appeal to Leibniz's principle. The structuralist interpretation motivates a chosen foundation; it does not establish a uniqueness theorem for all possible foundations.

For firms or learners, equivalence must be defined on structured systems. An appropriate object includes $(S, o, u, V, \text{resources}, \text{optimizer})$; its equivalences preserve the declared interfaces and these chosen structures. Equivalence of underlying state types alone transports no unspecified costs or control policies. A chosen universe may forget them entirely.

Example 4.1 Equal predictions, different learning speeds

Let $F(\theta) = \theta$ and $\tilde{F}(\phi) = 2\phi$, both mapping R to R . The reparametrization $h(\phi) = 2\phi$ is a smooth bijection and $\tilde{F} = F \circ h$. For the same output loss $\ell(z)$ and Euclidean gradient flow in each coordinate, the scalar tangent kernels are $K = 1$ and $\tilde{K} = 4$. Hence

$$\dot{z} = -\ell'(z), \quad \dot{\tilde{z}} = -4\ell'(\tilde{z}).$$

The models have the same realizable predictions and equivalent parameter spaces, but different time-dependent predictions from matched initial outputs.

Proposition 4.2 Kernel invariance requires optimizer geometry

Suppose $\tilde{F} = F \circ h$ for a smooth diffeomorphism h , and let M and $\tilde{M}$ be positive semidefinite cotangent-to-tangent preconditioners. If $Dh \tilde{M} Dh^T = M \circ h$, then the output kernels agree at matched parameters. **Proof.** The chain rule gives $D\tilde{F} = DF Dh$. Consequently $D\tilde{F} \tilde{M} D\tilde{F}^T = DF M DF^T$. In Example 4.1, equality is restored by $\tilde{M} = 1/4$ when $M = 1$.

Reading figure: reparam. See the retained source snapshot.

This gives a concrete requirement for a higher groupoid of learner presentations: equivalences of optimizer-equipped systems must preserve the relevant geometry. Gauge symmetries such as hidden-unit permutations may be organized into an action groupoid, and, with descent, a quotient stack $[\Theta/G]$. Its stabilizers record presentation automorphisms. A quotient is not automatically a smooth manifold, and a kernel descends only when the required invariance holds. Here a group action is a map $G \times \Theta \rightarrow \Theta$ preserving group multiplication and identity; its action groupoid has an arrow $\theta \xrightarrow{g} g\theta$ for each $g \in G$. The stabilizer of θ is $\{g: g\theta = \theta\}$. A stack is a groupoid-valued sheaf satisfying coherent descent, and the quotient stack is the sheafified action groupoid. Gauge symmetry here means a reparametrization preserving the declared learner structure. A moduli object classifies such structures together with their equivalences; it retains automorphisms rather than only their orbit set.

Example 4.3 Truncation does not identify realizations

For every $n \geq 0$, the point and $K(Z, n+1)$ have equivalent n -truncations but are not equivalent spaces, since the latter has $\pi_{n+1} = Z$. Even full behavioral equivalence does not fix a machine: adjoining unused memory to a deterministic implementation leaves its observed behavior unchanged while changing its physical state count and possible cost.

The point $*$ is the one-point homotopy type. The Eilenberg–Mac Lane space $K(Z, k)$, for $k \geq 1$, is connected, has $\pi_k = Z$, and has all other positive-degree homotopy groups zero. Behavioral equivalence means equality of the declared observable input-output behavior; additional internal state is retained only when it is part of that declared structure. Thus a unique

specified operation class need not have a unique implementation. Also, $\tau_{\leq n}$ is not in general a left-exact localization of an ∞ -topos. In spaces,

$$\tau_{\leq 0}(* \times_{S^1} *) \simeq Z, \quad * \times_{\tau_{\leq 0}S^1} * \simeq *,$$

so 0-truncation fails to preserve this homotopy pullback. Sheafification and homotopy truncation must therefore play distinct roles.

5 Neural dynamics: the exact kernel identity

5.1 Finite models and continuous-time optimization

The notation C^r means derivatives through order r exist and are continuous; C^∞ means this for every finite r . Write DF for the Jacobian, $D^k F$ for the k -linear derivative, ∇L for the gradient in the declared inner product, and $\nabla^2 L$ for its Hessian. A dot denotes a time derivative. A real symmetric matrix M is positive semidefinite, written $M \geq 0$, when $v^T M v \geq 0$ for every v ; $M \geq \mu I$ means $M - \mu I \geq 0$. A Gram matrix has entries given by inner products of chosen vectors and is positive semidefinite. Neural tangent kernels below are matrix-valued Gram kernels; their meaning differs from a probabilistic Markov kernel. A loss is a scalar objective measuring the chosen discrepancy. A preconditioner maps its parameter covectors to update directions.

Fix a dataset and concatenate a differentiable network's training outputs into $F: \Theta \subseteq R^p \rightarrow R^q$. For vector-valued outputs q includes both sample and output indices. Let $\ell: R^q \rightarrow R$ be differentiable, set $J(\theta) = DF(\theta)$ and $g = \nabla \ell(F(\theta))$, and prescribe

$$\dot{\theta} = -M(\theta, t) J(\theta)^T g + v(\theta, t).$$

Here p is the number of parameter coordinates, q the number of stacked output coordinates, and $N \geq 1$ the sample count when a sample-normalized loss is used. A neural network means the chosen parametrized map f_θ ; F stacks its outputs on that sample. The transpose is denoted T , I is the identity matrix of the required size, and $\|\cdot\|$ is the Euclidean norm or its induced multilinear operator norm. The matrix $M \geq 0$ is a chosen preconditioner and v is an additional parameter drift. In geometric language M maps covectors to tangent vectors; the transpose uses the displayed Euclidean coordinates. Assume sufficient local regularity for the trajectory under consideration to exist. The following is a direct chain-rule calculation, not an infinite-width approximation.

Theorem 5.1 Optimizer-dependent tangent dynamics

Along (4),

$$K_M(\theta, t) = J(\theta) M(\theta, t) J(\theta)^T,$$

$$\dot{F} = -K_M g + J v,$$

$$\frac{d}{dt} \ell(F) = -g^T K_M g + g^T J v.$$

The matrix K_M is positive semidefinite. For $v = 0$ the fixed output loss is nonincreasing. If also $K_M \geq \mu I$ and $\ell(F) = \|F - y\|^2 / (2N)$, then $\|F(t) - y\| \leq e^{-\mu t / N} \|F(0) - y\|$ while this bound holds. **Proof.** Multiply (4) by DF to obtain (6). For any w , $w^T K_M w = (J^T w)^T M (J^T w) \geq 0$. The loss chain rule gives (7). In the squared-loss case, the residual $e = F - y$ satisfies $\dot{e} = -K_M e / N$, so $\frac{d}{dt} \|e\|^2 \leq -2\mu \|e\|^2 / N$; integrate this inequality.

Reading figure: residual. See the retained source snapshot.

For $M = I$, $\nu = 0$, this is the empirical neural tangent kernel (NTK). In sample notation, including the output block structure,

$$K_{\theta}(x, x') = D_{\theta} f_{\theta}(x) D_{\theta} f_{\theta}(x')^T.$$

Jacot, Gabriel, and Hongler established the foundational NTK viewpoint and a constant-kernel infinite-width regime under their hypotheses [19]. Linearized wide-network results [23] and tensor-program calculations for broad architectural classes [36] extend its scope. They do not imply that an arbitrary finite transformer or other deployed network has a constant kernel. Feature learning can change J and K ; lazy training is a particular scaling regime [5] in which the parametrized map remains close to its linearization at initialization over the specified training interval.

5.2 Kernel evolution and the tangent hierarchy

For F , M sufficiently differentiable and no explicit time dependence of F ,

$$\dot{K}_M = \dot{J} M J^T + J \dot{M} J^T + J M \dot{J}^T, \quad \dot{J} = D_{\theta} J[\dot{\theta}].$$

The evolution already involves second parameter derivatives, and derivatives of the optimizer geometry. For the special scalar-output Euclidean flow with $L = (2N)^{-1} \sum_i (f_i - y_i)^2$, define recursively

$$H_{\theta}^{(2)}(i, j) = \langle \nabla f_i, \nabla f_j \rangle, \quad H_{\theta}^{(m+1)}(I, k) = \langle \nabla_{\theta} H_{\theta}^{(m)}(I), \nabla_{\theta} f_k \rangle.$$

The labels y_i are fixed scalar targets, $f_i = f_{\theta}(x_i)$, and I in $H^{(m)}(I)$ is an ordered m -tuple of sample indices; this local use is distinct from the identity matrix notation. Whenever the derivatives exist, the chain rule yields the exact identities

$$\dot{f}_i = -\frac{1}{N} \sum_k H_{\theta}^{(2)}(i, k) (f_k - y_k), \quad \dot{H}_{\theta}^{(m)}(I) = -\frac{1}{N} \sum_k H_{\theta}^{(m+1)}(I, k) (f_k - y_k).$$

This is the recursive structure of the neural tangent hierarchy (NTH) [18]. These higher tensors are not, in general, positive semidefinite two-input kernels or fully symmetric tensors. Freezing a finite level is a closure approximation; the identities themselves do not bound its error. Huang and Yau obtain controlled truncations under specified smoothness, width, initialization, and data assumptions. Their approximation theorem is not asserted here for every modern architecture or optimizer.

6 Jets, higher-order optimization, and a jet tangent kernel

6.1 Germs, the infinite jet tower, and holonomic sections

Fix an open parameter domain $\Theta \subseteq R^p$; the same definitions apply in manifold charts. For scalar objectives write $L = C^{\infty}(\Theta, R)$. The standard geometric constructions of jets and their coordinate transformations are developed in [30] and [21]. The factorization results below are elementary consequences of explicitly retaining zeroth-order values and runtime state.

Definition 6.1 Smooth germ and pointwise jet

The germ $[L]_\theta$ of a smooth function at θ is its equivalence class under agreement on some open neighborhood of θ . The germ algebra is $G_\theta = \varinjlim C^\infty(U, R)$, where the maps are restrictions and the colimit runs over neighborhoods $U \ni \theta$. Two germs have the same order- r jet when all their derivatives of orders at most r agree at θ . Their infinite jets agree when this holds at every finite order. The infinite jet fiber is the inverse limit

$$J_\theta^\infty = \varprojlim_r J_\theta^r,$$

with bonding maps forgetting the highest-order derivatives.

An inverse limit here is a compatible sequence under these forgetful maps; the germ direct limit identifies representatives agreeing after restriction. In Taylor coordinates the natural map $G_\theta \rightarrow J_\theta^\infty$ sends a germ to its full derivative sequence. Its kernel is the ideal of **flat germs**, whose derivatives of every order vanish at θ . It can identify distinct smooth germs. In one variable, $h(x) = e^{-1/x^2}$ for $x > 0$ and $h(x) = 0$ for $x \leq 0$ has a nonzero germ at zero but zero infinite jet there. Every derivative of h on $x > 0$ is an inverse-power polynomial times the exponential and tends to zero at the origin, which proves flatness. An analytic germ is represented by a locally convergent power series; for analytic germs the Taylor map is injective. Smooth germs, formal infinite jets, and analytic germs therefore require distinct types.

Definition 6.2 Global holonomic jet representation

The prolongation of L is its jet section over the entire domain, $P_\infty(L): \Theta \mapsto j_\theta^\infty L$. A section is **holonomic** when it is the prolongation of an actual smooth function. Let $\text{Hol}_\infty(\Theta)$ be precisely these sections. Zeroth-order projection defines $\varepsilon_0: \text{Hol}_\infty(\Theta) \rightarrow L$ by $[\varepsilon_0(s)](\theta) = s_0(\theta)$. Thus $\varepsilon_0 P_\infty = \text{id}_L$ and $P_\infty \varepsilon_0 = \text{id}_{\text{Hol}_\infty}$. The global section retains $L(\theta)$ at every point. This differs from the derivative sequence at a single point. With unnormalized coordinates $s_\alpha(\theta)$, holonomicity implies $\partial_j s_\alpha = s_{\alpha+e_j}$, where e_j is the j th unit multiindex. Conversely, smooth component functions satisfying these equations are the derivatives of s_0 , by induction. Arbitrary pointwise formal coefficient assignments need not satisfy this compatibility.

The assignment $U \mapsto C^\infty(U, R)$ is a sheaf: local smooth functions agreeing on overlaps glue to one smooth function. Germs are its stalks, namely the direct limits over neighborhoods. Prolongation and zeroth-order projection commute with restrictions on the corresponding holonomic jet sheaf. This gives the local-to-global organization of objective data inside smooth contextual semantics. An algorithm querying outside an open set U is not an operation determined solely by the objective restricted to U .

6.2 Jet algebra and its limitations

In a chosen parameter chart, the order- r parameter jet of a C^r map F is its Taylor polynomial modulo terms of degree $r + 1$:

$$j_\theta^r F(\delta) = \sum_{|\alpha| \leq r} \frac{\partial^\alpha F(\theta)}{\alpha!} \delta^\alpha.$$

At a fixed point it records derivatives through order r , not the entire map. Under coordinate changes these records transform by the chain rule. They can be organized into jet bundles, whose fibers collect the jets at each base point, and compositional differential structures [3; 30; 21]. Here $\alpha = (\alpha_1, \dots, \alpha_p) \in N^p$ is a multiindex, $|\alpha| = \sum_j \alpha_j$, $\alpha! = \prod_j \alpha_j!$, and $\delta^\alpha = \prod_j \delta_j^{\alpha_j}$. For a one-variable jet over a characteristic-zero algebra A , normalized Taylor coefficients obey

$$(a + b)_k = a_k + b_k, \quad (ab)_k = \sum_{j=0}^k a_j b_{k-j}, \quad (Da)_k = (k+1)a_{k+1}.$$

Equivalently, with unnormalized derivatives, $(ab)^{(k)} = \sum_{j=0}^k \binom{k}{j} a^{(j)} b^{(k-j)}$. Thus multiplication is a convolution, not pointwise multiplication of derivative lists. The product order shown also works for noncommutative coefficients when the variable is central and D is a derivation; arbitrary reordering is not permitted. An algebra here is an associative algebra over a field; characteristic zero ensures the displayed integer coefficients have their usual nonzero values.

Lemma 6.3 Differentiation lowers finite jet order

For $r \geq 1$, differentiation of germs induces $D: J^r \rightarrow J^{r-1}$, but ordinary polynomial differentiation does not induce an endomorphism of $A[z]/(z^{r+1})$ for a characteristic-zero field A . **Proof.** If two representatives differ by $z^{r+1}h$, their derivatives differ by a multiple of z^r , so they have the same $(r-1)$ -jet. But z^{r+1} is zero in the quotient and its derivative $(r+1)z^r$ is nonzero there. An endomorphism defined by differentiating representatives is therefore not well defined.

The infinite jet, or formal power series, admits the coefficient shift without this finite-order loss. Truncated polynomial representatives can be differentiated after choosing a representative, but that is a chosen approximation to the germ's derivative. Regular jets do not automatically include Laurent series with a pole at the base point. Shared differential formulas are not an equivalence of all polynomial, Laurent, smooth, and formal theories.

6.3 A universal jet representation of optimization algorithms

Definition 6.4 Stateful objective-based optimizer

Let Ξ be a runtime state space containing current parameters and any memory, query history, population of candidates, or random seed used by the algorithm. Let C contain declared auxiliary data, constraints, precision, schedules, objective presentations, and oracle access rules. A deterministic optimizer step is a possibly partial map $A: L \times \Xi \times C \rightarrow Y$, where Y contains subsequent states or stopping results. For a single-state iteration one can take $Y = \Xi + \text{Stop}$, where Stop is the declared set of terminal answers and statuses. It is **extensional in the objective** when equal functions with the same runtime state and auxiliary data yield the same definedness and output. A stochastic step is instead a Markov kernel to Y , with its measurable structures specified. An optimization algorithm is an iteration of such steps, including its query, acceptance, and stopping rules; the name alone does not assert successful minimization.

Code inspection and representation-dependent costs are explicit auxiliary data when they affect the algorithm. A randomized program may equivalently use a deterministic step with its seed in Ξ . The oracle is the interface answering objective or derivative queries; a black-box algorithm knows the objective through those answers rather than through a closed formula.

Theorem 6.5 Universal global-jet factorization

Every optimizer of Definition 6.4 on smooth objectives has a unique optimizer on global holonomic jets,

$$\tilde{A}(s, \xi, c) = A(\varepsilon_0(s), \xi, c), \quad A(L, \xi, c) = \tilde{A}(P_\infty(L), \xi, c).$$

For partial maps its domain is transported by ε_0 . The same statement holds for stochastic steps, with the measurable structure on holonomic sections transported from L . Stopping behavior and the entire iterated trajectory are preserved from matched initial runtime states and corresponding objective and auxiliary-input streams. **Proof.** The inverse identities in Definition 6.2 give the displayed factorization. Every s in the holonomic space equals $P_\infty(\varepsilon_0(s))$, so any factorization must have the

displayed value and domain; this proves uniqueness. For stochastic steps use the same substitution in every measurable output event. The transported measurable structures make this substitution measurable. Induction over steps preserves deterministic trajectories; integral composition preserves the stochastic trajectory law and stopping events.

This is the precise sense in which every specified objective-based optimization algorithm is an instance of the global jet construction. It includes derivative-free methods because the objective's values are its zeroth-order component. The theorem is a representation result: infinite derivatives can be available without each algorithm actually computing or using them. It also holds with global order- r holonomic sections for any $r \geq 0$; increasing the order enriches the accessible differential operations, rather than being necessary to recover a function already retained at order zero. Extending an optimizer to nonholonomic formal sections is a separate modeling choice and is not determined by this theorem.

Proposition 6.6 Compositional equivalence of optimizer interfaces

Fix auxiliary data. Let Opt_L have runtime state spaces as objects and total steps $A: L \times \Xi \rightarrow \Xi'$ as arrows, with composition $(B \star A)(L, \xi) = B(L, A(L, \xi))$. Replacing L by $\text{Hol}_\infty(\Theta)$ gives a category isomorphic to Opt_L through $A \mapsto \bar{A}$. Partial steps and stochastic steps have the corresponding partial-map and kernel compositions. **Proof.** Identity steps return ξ ; associativity follows from substitution. Since $\varepsilon_0(s) = L$ is shared by both composed steps, $B \approx A = B \star A$. Precomposition with P_∞ is inverse to precomposition with ε_0 . The same argument transports domains, and kernel composition uses the integral law already specified.

Composition permits search, differentiation, filtering, acceptance, resource checks, and release gates to use one typed objective representation while retaining their separate runtime states.

Proposition 6.7 Criterion for a pointwise-jet optimizer

Fix θ and a total germ-local rule $A_\theta: G_\theta \times \Xi \times C \rightarrow Y$, with Y a set. It factors through realizable order- r pointwise jets, including $r = \infty$, exactly when its value is constant on germs having the same r -jet, for each fixed (ξ, c) . For a partial rule its domain must also be a union of these equivalence classes. **Proof.** A factorization is constant on the fibers of the jet map. Conversely, assign to each realizable jet the value of any representative germ. The premise makes this assignment independent of representative, giving the unique descended rule. The same reasoning applies to a partial rule precisely when definedness also descends.

Here **germ-local** means dependence only on agreement on a neighborhood; **pointwise-jet determined** means dependence only on the derivative data at the base point. These are distinct conditions. For example, choose a smooth function β supported in $(1/2, 3/2)$ with $\beta(1) = 1$. The objectives $L_0(x) = x^2$ and $L_1(x) = x^2 - 2\beta(x)$ have the same germ at zero. Searching the candidates $\{0, 1\}$ returns zero for L_0 and one for L_1 . This legitimate value-based search cannot factor through the current germ or current pointwise jet. It does factor through the global holonomic jet section. Likewise the flat-germ example in Definition 6.1 explains why the infinite pointwise jet can forget information even in a germ-local problem.

Orders, queries, and optimizer state. Gradient methods read first-order jets of L at their query points. Newton methods read second-order jets. A line search selects a step length along a declared direction by querying candidate values; it can be expressed through zeroth-order evaluations of the global jet section. Finite differences use several such evaluations to approximate derivatives. Quasi-Newton methods update an approximate Hessian or inverse Hessian from successive gradient and displacement pairs, so their memory belongs to Ξ . Momentum and Adam combine first-order queries with retained memory. Population and randomized search methods retain several candidate points and query values at each. A trust-region method optimizes a local Taylor model under a bound $\|\Delta\theta\| \leq \rho$ with $\rho > 0$; higher-order tensor methods use higher multilinear derivatives. Constraints and acceptance tests remain part of the update interface.

For a composite loss $L = \ell \circ F$, algorithms that use the network's representation, such as Gauss–Newton, retain F and ℓ as auxiliary data, or use their jet sections jointly. Equality of the scalar loss alone need not identify this representation. Minibatch methods likewise specify the sampled objective family and the batch-selection state. Chart changes act on jets by the chain rule, and coordinate-invariant training still requires transport of the optimizer geometry as in Proposition 4.2.

For nonsmooth or discrete objectives, classical infinite derivative jets may not exist. The universal value interface remains: a function $L: \Theta \rightarrow R$ is its global zeroth-order section. Subgradients, generalized derivatives, or surrogate derivatives can be additional declared oracle data. Consequently every objective-based algorithm has a global value/jet interface in its appropriate regularity setting. This does not identify arbitrary black-box search with a rule using only one smooth germ's derivative sequence. The distinction is also computational: a runtime can expose finite jet queries on demand, without storing an infinite array or supplying an oracle for uncomputable objectives. A subgradient of a convex function L at θ is a vector v with $L(y) \geq L(\theta) + v \cdot (y - \theta)$ for every y in its domain. Convexity means $L(tx + (1-t)y) \leq tL(x) + (1-t)L(y)$ for $0 \leq t \leq 1$ on a domain containing each such line segment. Other generalized derivatives mean explicitly chosen set-valued replacements for classical derivatives; a surrogate derivative is an explicitly supplied replacement used by the algorithm, for example from a smooth approximation.

6.4 Second and higher-order optimizers

For $L = \ell \circ F$, the parameter Hessian is

$$\nabla_{\theta}^2 L = J^T (\nabla_F^2 \ell) J + \sum_{a=1}^q g_a \nabla_{\theta}^2 F_a.$$

Newton-type updates use this second jet. A damped positive-definite approximation can supply $M = (\hat{H} + \lambda I)^{-1}$ in Theorem 5.1, where $\hat{H}$ is the declared symmetric Hessian approximation and damping λ is chosen so that $\hat{H} + \lambda I$ is positive definite. Gauss–Newton retains the first term of (11); a raw indefinite inverse Hessian need not produce descent. Higher Taylor models, trust-region methods, and tensor methods can use larger jets, but a jet by itself is not an optimization algorithm: objective, step selection, constraints, and acceptance must also be chosen.

For an actual finite parameter step $\Delta\theta$, Taylor's theorem gives

$$F(\theta + \Delta\theta) - F(\theta) = \sum_{k=1}^r \frac{1}{k!} D^k F_{\theta} [\Delta\theta^{\otimes k}] + R_{r+1},$$

with $\|R_{r+1}\| \leq B \|\Delta\theta\|^{r+1} / (r+1)!$ when the $(r+1)$ st derivative is bounded by B along the connecting segment. This quantifies the step-size limitation of a finite jet description.

6.5 An explicitly defined input-jet tangent kernel

The phrase “jet tangent kernel” does not specify a single universal object. For a precise version, take continuous inputs $x \in R^d$, vector outputs, and multiindices $|\alpha| \leq r$. Define jet observables $Z_{\alpha, \theta}(x) = \partial_x^{\alpha} f_{\theta}(x)$. Assume the mixed derivatives needed below exist and commute. Stack the chosen sampled observables into $Z(\theta)$ and set

Definition 6.8 Order- r input-jet tangent kernel

For a parameter preconditioner M , define the block kernel

$$K_{M;\alpha\beta}^{\text{jet},r}(x, x') = D_\theta \partial_x^\alpha f_\theta(x) M [D_\theta \partial_{x'}^\beta f_\theta(x')]^T.$$

The observable coordinates in this definition are unnormalized derivatives; using Taylor-normalized coordinates rescales the corresponding blocks.

For M independent of input, differentiating the ordinary output block kernel gives $K_{M;\alpha\beta}^{\text{jet},r} = \partial_x^\alpha \partial_{x'}^\beta K_M(x, x')$.

Proposition 6.9 Jet-observable dynamics

For training on $\ell_{\text{jet}}(Z(\theta))$ with $\dot{\theta} = -M(D_\theta Z)^T \nabla_Z \ell_{\text{jet}}$, the stacked input jets satisfy $\dot{Z} = -K_M^{\text{jet},r} \nabla_Z \ell_{\text{jet}}$ and $K_M^{\text{jet},r} = (D_\theta Z) M (D_\theta Z)^T \geq 0$. **Proof.** Apply the chain rule and the Gram-matrix argument of Theorem 5.1 to Z in place of F .

This is appropriate for derivative matching and Sobolev objectives; Sobolev training is an established example [8]. A sampled Sobolev objective, for example, is $\frac{1}{2} \sum_{i,\alpha} w_\alpha \|Z_{\alpha,\theta}(x_i) - t_{i,\alpha}\|^2$, with nonnegative weights w_α and prescribed derivative targets $t_{i,\alpha}$. The term means that function values and chosen derivatives both enter the loss. It is distinct from the parameter-jet tower in (10) and the NTH in (9). One may also construct weighted Gram matrices from higher parameter derivatives, but without an optimizer law such matrices do not determine training.

For discrete tokens there are no ordinary input derivatives in the token index; input jets can instead be taken in continuous embedding variables when that is the intended model. Parameter derivatives of logits remain meaningful wherever the parametrization is differentiable.

6.6 What no fixed kernel can determine

For momentum, optimizer memory is essential: matched parameters can have different velocities. For Adam [20], with gradient $g_t = \nabla_\theta L_t$,

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t, \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^{\odot 2}, \\ \Delta \theta_t &= -\eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}. \end{aligned}$$

Here $0 \leq \beta_1, \beta_2 < 1$, the learning rate is $\eta > 0$, and $\varepsilon > 0$ regularizes division; operations are componentwise, $g_t^{\odot 2}$ squares each component, and with zero initial moments $\hat{m}_t = m_t / (1 - \beta_1^t)$, $\hat{v}_t = v_t / (1 - \beta_2^t)$ are bias corrections. The update depends on history through (m_t, v_t) , not solely on a kernel at θ_t . Minibatches, schedules, weight decay, and architecture changes add further state or drift. If a diffusion model for training is specified, $d\theta = b dt + B dW$, Itô's formula gives

$$dF_a = (DF_a b + \frac{1}{2} \text{tr}(BB^T \nabla^2 F_a)) dt + DF_a B dW.$$

Here W is standard Brownian motion, b the specified drift vector, B the noise-amplitude matrix, and tr the matrix trace. The Hessian correction is absent from deterministic gradient flow. ReLU, the rectified linear unit $u \mapsto \max(0, u)$, and other nonsmooth mechanisms require piecewise-smooth analysis, generalized derivatives, or a chosen surrogate; higher classical

jets need not exist on switching boundaries. Quantized updates and sampling require discrete semantics. Derivative-free optimization fits the global zeroth-order interface; it need not depend on pointwise derivatives. Quantization means a declared map to a finite-precision value set, and sampling draws outcomes from a declared law.

As Definition 6.1 shows, the infinite jet at one point need not determine a smooth germ. Theorem 6.5 instead uses the global holonomic section and the full optimizer state. Analyticity restores local determination by a convergent Taylor series; it does not remove optimizer or environmental dependence. The justified differential statement is the chain-rule identity for differentiable parametrized models under specified update laws. A single fixed “jet kernel” does not determine all training, inference, deployment, and organizational dynamics.

7 Observation, trace capture, and organizational control

7.1 Capability-relative observation

For an actor a , choose an observation map $O_a: S \rightarrow Y_a$ and an admissible action family $A_a(s)$. A capability change may refine O_a , enlarge A_a , or alter resource access. In a set-level deterministic model, states s, s' are observationally equivalent to a when $O_a(s) = O_a(s')$. In higher semantics, use the homotopy fibers of O_a and retain witnesses when appropriate. This describes capability through accessible observations and actions, with no assumption that capability is literally a homotopy level.

Theorem 7.1 Observation-fiber obstruction to regulation

Let $P: S \times A \rightarrow S$ be a deterministic transition, restrict Y to the reachable observation image $O(S)$, and let $B \subseteq S$ be the desired safe region. Define $Q(s) = \{a \in A: P(s, a) \in B\}$. A controller $c: Y \rightarrow A$ using only $O: S \rightarrow Y$ regulates all states only if

$$\bigcap_{s \in O^{-1}(y)} Q(s) \neq \emptyset \quad \text{for every } y \in O(S).$$

Conversely, a supplied selection from these intersections defines a regulator. In constructive type theory, the converse requires an actual dependent selection, not merely an assertion that each intersection is nonempty. **Proof.** If c regulates, then $c(y) \in Q(s)$ for every s with $O(s) = y$, proving (14). A given choice a_y in each intersection, with its membership witnesses, defines $c(y) = a_y$ and proves safety for each state.

In particular, if indistinguishable states need incompatible actions, no extra computation applied to the same instantaneous observation can solve the problem. Memory or new sensors can change the observation object. If a finite disturbance set requires a different unique action for each disturbance, any regulating observation must distinguish every disturbance: $|O(S)| \geq |S|$, or at least $\log_2 |S|$ bits of distinguishable observation. This restricted theorem illustrates requisite variety; it is not a blanket entropy inequality for every control system. Conant and Ashby’s good-regulator result has its own optimality and modeling assumptions [6].

7.2 Employee traces, customer traces, and the panoptic hypothesis

An **observational trace** is a finite sequence of typed interaction events with declared times, participants, and outputs. An employee trace comes from work performed in an employee role; a customer trace comes from use of a product or service in a customer role. One person can occupy both roles. Data provenance specifies origin, transformations, and the permissions attached to the data. These traces become inputs to a learner only through a specified collection and data-processing map. This observational use of “trace” differs from an executable operation sequence.

Let D be existing data, T a newly captured work trace, and Y a future prediction target, all modeled as finite random variables on a declared probability space. The Shannon conditional entropy, in bits, is

$$H(Y \mid D) = - \sum_{d,y} \Pr(d, y) \log_2 \Pr(y \mid d), \quad 0 \log_2 0 = 0.$$

Log loss for a predicted distribution q is $-\log_2 q(Y)$; the Bayes-optimal predictor is the true conditional distribution, minimizing expected loss. Conditional mutual information is $I(Y; T \mid D) = H(Y \mid D) - H(Y \mid D, T)$. Conditional relative entropy compares distributions through $\sum_x p(x) \log_2 [p(x)/q(x)]$, averaged over the conditioning data. For finite random variables under log loss, the Bayes-optimal expected loss is conditional entropy. Hence the gain from adding the trace is

$$H(Y \mid D) - H(Y \mid D, T) = I(Y; T \mid D) \geq 0.$$

This follows directly from the definition of conditional mutual information and nonnegativity of conditional relative entropy. It gives a precise condition under which internal traces have predictive value. It does not establish equal value to customer data, a monetary valuation, or an incentive to capture every trace: acquisition cost, provenance, representativeness, permissions, and objective choice remain independent variables.

For worker and customer trace families T_W, T_C , equal predictive value would be the additional equation $I(Y; T_W \mid D) = I(Y; T_C \mid D)$ for the declared target, baseline, and data budget. It does not follow from both being traces. For example, with D constant, Y a fair bit, T_W an independent fair bit, and $T_C = Y$, these two quantities are zero and one bit, respectively. Information value, collection coverage, and authority must therefore be specified independently.

Definition 7.2 Observation refinement and authority

For a common state domain Ω , observation maps $O_a: \Omega \rightarrow Z_a$ and $O_b: \Omega \rightarrow Z_b$ satisfy $O_a \leq O_b$ when $O_a = k \circ O_b$ for some map $k: Z_b \rightarrow Z_a$. Thus a 's view can be reconstructed from b 's. This is a preorder: reflexivity uses the identity and transitivity uses composition. Mutual refinement defines equivalence of information; the preorder is not assumed antisymmetric. For a common intervention interface $A(\omega)$, let $A_a(\omega) \subseteq A(\omega)$ be actor a 's authorized actions. For a common policy space Λ , let $\Gamma_a(\omega) \subseteq \Lambda \times \Lambda$ be its authorized policy changes. Policies specify controllers, objectives, observation access, and update criteria. Authority domination means inclusion of the corresponding action and policy-change sets in each state. Strict domination means at least one inclusion is proper.

The channels and rights concern a declared organizational interface, rather than every fact an actor knows or every action an actor can perform. Controller inspection and challenges to updates are explicit observation or governance commands when included in this interface.

Definition 7.3 Trace coverage

Choose a finite operational state abstraction Ω , including the relevant firm, environment, and interaction histories. Let $I = I_W \sqcup I_C$ be a finite set of employee and customer role instances, with weights $w_i > 0$ and $\sum_i w_i = 1$. For each i choose a nonconstant trace observable $\tau_i: \Omega \rightarrow T_i$, with T_i finite, and an observation map O_i . Write f for the firm authority role and $O_f: \Omega \rightarrow Z_f$ for its finite observation channel. Take each observation codomain to be its reachable image. Define

$$I_f = \{i: \text{there is } r_i: Z_f \rightarrow T_i \text{ with } \tau_i = r_i \circ O_f\}, \quad C_f = \sum_{i \in I_f} w_i.$$

The number $C_f \in [0, 1]$ is informational trace coverage for these observables and weights; r_i is a recovery witness. High coverage means $C_f \geq q$ for a declared threshold $0 < q \leq 1$.

Recoverability is equivalent to τ_i being constant on each fiber of O_f : this condition defines r_i on the reachable image, and a recovery map immediately implies it. Nonconstant trace observables exclude uninformative constant functions from the coverage count. Recoverability does not by itself say that a particular storage or collection mechanism was used. For noisy channels one can instead declare a probability law and errors ε_i , requiring $\Pr[r_i(O_f) \neq \tau_i] \leq \varepsilon_i$. Choosing ε_i below the best constant-predictor error $1 - \max_t \Pr(\tau_i = t)$ ensures that this approximate recovery is informative.

Definition 7.4 Panoptic architecture and panoptic hypothesis

Let P contain the state domain, roles, observation maps, trace observables, weights, actions, and governance relations just defined. It is a **panoptic architecture** at threshold q , written $\text{Pan}_q(P)$, when: - $C_f \geq q$; - for every $i \in I_f$, $O_i \leq O_f$ and, in every state, $A_i(\omega) \subseteq A_f(\omega)$ and $\Gamma_i(\omega) \subseteq \Gamma_f(\omega)$; - for some $i \in I_f$ and state ω , at least one of these two authority inclusions is proper. For a declared collection F of firms and a fixed assessment period, the panoptic hypothesis is the empirical assertion $H_{\text{pan}}(q, F): \forall f \in F, \text{Pan}_q(P_f)$. The literal claim about every AI company uses that entire declared class for F ; a case study can instead test a restricted collection.

This is an operational definition inspired by the asymmetry of observation and authority in the historical account of panopticism [10]. It specifies the mathematical meaning of the term in this note. The threshold, role weights, trace scope, recovery errors, and authorization evidence are part of the hypothesis and must be reported. Positive predictive value alone proves none of its three conditions. A model can have full trace coverage and fail the definition when all action and policy-change rights are equal. Conversely, authority concentration does not establish high trace coverage. Controllability means the ability to reach declared target states through allowed inputs; observational predictability is a separate property.

Proposition 7.5 Panoptic status is invariant under structured equivalence

In the finite model, bijective changes of state, observation, trace, action, and policy coordinates preserving roles, weights, maps, and authorization relations preserve Pan_q . **Proof.** A recovery map transports by composing with the observation bijection's inverse and the trace bijection. Hence exactly the same role instances contribute to C_f . Observation factorizations transport in the same way. Bijections preserving the authorization relations preserve inclusion and proper inclusion of the corresponding sets. All three conditions are therefore preserved, in both directions.

In higher contextual semantics a recovery witness has the internal type

$$r: Z_f \rightarrow T_i \quad \text{Id}_{\text{Map}(\Omega, T_i)}(\tau_i, r \circ O_f).$$

Mere recoverability is its (-1) -truncation, while the full type retains maps and coherence witnesses. Numerical coverage above uses a finite, decidable abstraction; a general higher recovery proposition is not automatically a decidable Boolean. An observation-only equivalence which forgets permissions need not preserve panoptic status. The structured equivalence must retain the authority data as well.

Deployment also changes data: write $e_{t+1} = P(e_t, a_t)$, $d_{t+1} = O(e_{t+1})$, and $\theta_{t+1} = \text{Opt}(\theta_t, d_{t+1}, o_t)$. This feedback is related to performative prediction [28]. It can amplify, attenuate, or redirect effects; its sign and stability must be estimated, not inferred from the existence of a loop.

7.3 Composed decision-makers and shared providers

A company contains and connects decision-makers with distinct objectives. The set-based open-game formalism [14] represents a component $G: (X, S) \rightarrow (Y, R)$ by a strategy set Σ , play $P: \Sigma \times X \rightarrow Y$, coplay $C: \Sigma \times X \times R \rightarrow S$, and a best-response relation

$$B: X \times (Y \rightarrow R) \rightarrow \text{Rel}(\Sigma).$$

An equilibrium in context (x, k) is a strategy σ with $(\sigma, \sigma) \in B(x, k)$. Sequential and parallel composition retain the environmental continuation k , rather than asserting that each isolated optimum remains optimal after composition. The type R specifies the outcomes relevant to this component; S returns outcomes to its environment. A deployment gate can thus expose actions forward and evaluation consequences backward. The formalism represents best responses, not a guarantee that a neural decision-maker implements them or converges to an equilibrium.

Hedges's research proposal [16] highlights the dependence created when apparently separate economic decision-makers use a shared upstream AI provider. A minimal probabilistic example makes the distinction precise. Suppose a finite latent provider regime W affects two otherwise conditionally independent action channels. Then

$$\Pr(a_1, a_2 \mid y_1, y_2) = \sum_w \Pr(w \mid y_1, y_2) \Pr(a_1 \mid y_1, w) \Pr(a_2 \mid y_2, w)$$

under the stated conditional-independence assumptions. This mixture generally does not factor into the two marginal channels. For $W \sim \text{Bernoulli}(1/2)$ and $a_1 = a_2 = W$, it gives $\Pr(a_1 = a_2 = 1) = 1/2$, whereas the product of marginals is $1/4$. This is a shared-cause example, not evidence of actual collusion. Common providers, jointly updated policies, and correlated failures should therefore be represented as shared state in an industry model; corporate separation alone does not prove independence of decision channels.

8 Coupled stability, resources, and consolidation

8.1 Learning stability does not imply firm stability

An equilibrium is a state with zero time derivative. It is asymptotically stable when all sufficiently close initial states remain close and converge to it. For the linear systems below this is equivalent to every eigenvalue having negative real part. A kernel mode is an eigenvector $v \neq 0$ with $K_M v = \lambda v$; its eigenvalue is λ . Consider a local two-mode approximation with learning residual e and an environmental or organizational response z :

$$\begin{pmatrix} \dot{e} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} -a & b \\ c & -d \end{pmatrix} \begin{pmatrix} e \\ z \end{pmatrix}, \quad a, d > 0.$$

One possible frozen-kernel interpretation is $a = \lambda(K_M)/N$ for a learning mode, while b, c, d encode the surrounding response and timescales.

Proposition 8.1 Two-mode closed-loop criterion

The equilibrium of (16) is asymptotically stable exactly when $ad - bc > 0$. If $bc > ad$, it has a positive real eigenvalue.

Proof. The characteristic polynomial is $\lambda^2 + (a + d)\lambda + (ad - bc)$. Its roots have negative real parts exactly when both coefficients after the leading term are positive. For negative determinant the roots are real and have opposite signs. At

$ad = bc$ a zero eigenvalue prevents asymptotic stability.

Reading figure: twomode. See the retained source snapshot.

Increasing an isolated learning rate, or making its loss decrease, does not by itself control b and c . Delays, switching releases, and nonlinear saturation require a richer model. A categorical wiring diagram specifies composition; it does not provide the missing stability estimates.

Resources can be made explicit through, for example,

$$\dot{E} = -P_{\text{compute}} - P_{\text{deployment}} + P_{\text{supply}}, \quad \dot{B} = \text{revenue} - \text{operating cost},$$

with E an available energy stock and B a financial balance, together with material stocks and capacity constraints. Power is energy per unit time. A viability region includes resource floors as well as technical and institutional conditions. Maintaining that region is a control problem with disturbances; parameter-loss minimization proves only the narrower statement in Theorem 5.1.

Beer's viable-system roles suggest a decomposition into operational model teams, coordination of shared infrastructure, allocation and audit, adaptation to the environment, and policy/identity [2]. This is a modeling correspondence, not evidence that an AI company is a biological organism or already a viable system in Beer's technical sense.

8.2 A conditional consolidation comparison

Suppose compatible data objects admit a pushout $D_1 \sqcup_{D_{12}} D_2$ in the chosen semantic category. This expresses a universal integration construction: for every test object X ,

$$\text{Map}(D_1 \sqcup_{D_{12}} D_2, X) \simeq \text{Map}(D_1, X) \times_{\text{Map}(D_{12}, X)} \text{Map}(D_2, X).$$

This defines the pushout as the gluing of the span through D_{12} . It entails no improvement in statistical quality or profit. Permissions, shared identifiers, and interface compatibility are prerequisites for the construction to represent a real integration.

Let Π_{sep} and Π_{joint} be explicitly specified net objectives. Decompose their difference as

$$\Pi_{\text{joint}} - \Pi_{\text{sep}} = S_{\text{duplicate}} + S_{\text{complement}} - C_{\text{integration}} - C_{\text{coordination}} - C_{\text{policy}} - C_{\text{failure}}.$$

Here the S terms are avoided duplication and complementary-data gains, while the C terms are the respective additional costs. Under this accounting, integration is preferred for this objective precisely when the right side is positive. This is a conditional comparison, not a prediction: independent objectives, diseconomies, limited resources, failure concentration, and governance can reverse the sign. No date or inevitability of industry-wide consolidation follows from categorical universal properties or from (15).

9 Further analysis and a minimal research protocol

The framework supplies a starting point with several measurable extensions. For a first case study, choose one workflow from employee or customer interaction to training, evaluation, release, and subsequent observations. Record its state variables, time delays, optimizer memory, permissions, resource costs, and rejection paths. Specify V before analyzing updates. Identify

which equivalences are intended to preserve predictions only, and which must also preserve optimizer geometry, costs, or control rights.

Estimate empirical tangent-kernel modes or Jacobian-vector products on a declared sample; materializing a full kernel is often unnecessary. Track kernel change, jet approximation error over actual step sizes, and sensitivity of behavior to release and data-selection decisions. Estimate the coupled response terms in (16); test whether broader observations resolve the obstruction in Theorem 7.1. These measurements separate information capture, effective intervention, and governance.

Four theoretical extensions follow naturally:

- **Compositional certificates.** Give module contracts and prove that their shared-resource and interface witnesses assemble into the descent datum of Proposition 3.3. Arbitrary local tests do not suffice.
- **A moduli object of learners.** Define equivalences preserving specified forward behavior, losses, optimizer geometry, and permissions. Compute stabilizers and ask which kernels and costs descend to its quotient stack.
- **Multiobjective and stochastic governance.** Replace a single scalar company objective by stakeholder-indexed objectives and a specified mechanism for selecting actions. Add Markov kernels, delays, and resource-aware policies.
- **Jet closure with error control.** Derive architecture- and optimizer-specific truncation estimates, rather than importing a smooth fully-connected NTH bound into an unrelated runtime.

Relating cocycles to homotopies requires a specified complex or derived moduli problem. A cochain complex has vector spaces C^k and maps $d^k: C^k \rightarrow C^{k+1}$ with $d^{k+1}d^k = 0$. Cocycles are $\ker d^k$, coboundaries are $\text{im } d^{k-1}$, and cohomology is $H^k = \ker d^k / \text{im } d^{k-1}$. A cochain homotopy between maps f, g is a degree- (-1) family h with $f - g = dh + hd$. In an ordinary cochain complex, cocycles are not automatically null-homotopic: two cocycles represent the same cohomology class when their difference is a coboundary. For a smooth quotient by a Lie group, the infinitesimal action gives a two-term complex

$$g \xrightarrow{\rho_\theta} T_\theta \ominus$$

in degrees $-1, 0$. Here a Lie group is a smooth manifold with smooth group operations, g is its tangent space at the identity, and ρ_θ is the derivative of the action there. The cokernel is the target vector space modulo the image. Its kernel records infinitesimal stabilizers and its cokernel records tangent directions modulo the action. A derived enhancement can introduce genuine obstruction and higher-homotopy information, but it must be constructed [29]; it is not supplied by ordinary Taylor coefficients.

Finally, a one-point model exists for many purely equational algebraic theories, including the theory of groups, but not for every consistent theory. The theory of nontrivial fields, with $0 \neq 1$, is a counterexample. An external model gives relative consistency through a sound metatheory; it does not automatically give an executable or physically realizable model. The practical aim is a hierarchy of explicitly presented models and checkable claims connecting formal semantics to observed organizational behavior. A model interprets the operations and predicates of a theory so that its axioms hold. A purely equational theory specifies operations and equations between their terms; logical consistency means that a contradiction is not derivable. Soundness of a metatheory means its derivations preserve truth in the models under consideration. A derived deformation problem, proposed here as a further extension, retains homotopy-valued families of perturbations and their higher relations; its actual complexes and obstruction maps must be specified before any additional claim is proved.

References

- [1]

W. Ross Ashby. *An Introduction to Cybernetics*. Chapman and Hall, London, 1956. Requisite variety: Chapter 11.

- [2]

Stafford Beer. *Brain of the Firm*. John Wiley and Sons, second edition, 1981.

- [3]

R. F. Blute, J. R. B. Cockett, and R. A. G. Seely. Cartesian differential categories. *Theory and Applications of Categories*, 22(23):622–672, 2009.

- [4]

Matteo Capucci, Bruno Gavranović, Jules Hedges, and Eigil Fjeldgren Rischel. Towards foundations of categorical cybernetics. *Electronic Proceedings in Theoretical Computer Science*, 372:235–248, 2022.

- [5]

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, volume 32, 2019.

- [6]

Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970.

- [7]

G. S. H. Cruttwell, Bruno Gavranović, Neil Ghani, Paul Wilson, and Fabio Zanasi. Categorical foundations of gradient-based learning. *Programming Languages and Systems (ESOP 2022)*, LNCS 13240, pages 1–28. Springer, 2022.

- [8]

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. *Advances in Neural Information Processing Systems*, volume 30, 2017.

- [9]

Brendan Fong, David I. Spivak, and Rémy Tuyéras. Backprop as functor: A compositional perspective on supervised learning. *34th Annual ACM/IEEE Symposium on Logic in Computer Science*, pages 1–13, 2019.

- [10]

Michel Foucault. *Discipline and Punish: The Birth of the Prison*. Vintage Books, 1995. Translated by Alan Sheridan. “Panopticism,” pp. 195–228.

- [11]

Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020.

- [12]

Bruno Gavranović. Fundamental Components of Deep Learning: A Category-Theoretic Approach. PhD thesis, University of Strathclyde, 2024.

- [13]

Bruno Gavranović, Paul Lessard, Andrew Joseph Dudzik, Tamara von Glehn, João Guilherme Madeira Araújo, and Petar Veličković. Position: Categorical deep learning is an algebraic theory of all architectures. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, pages 15209–15241, 2024.

- [14]

Neil Ghani, Jules Hedges, Viktor Winschel, and Philipp Zahn. Compositional game theory. Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science, pages 472–481, 2018.

- [15]

Alexander Grothendieck. Pursuing stacks. Circulated manuscript, 1983. Revised transcription edited by Mateo Carmona with Ulrik Buchholtz, 2021.

- [16]

Jules Hedges. AI safety meets value chain integrity. Cybercat Institute, December 11, 2023. Research proposal.

- [17]

Martin Hofmann and Thomas Streicher. The groupoid interpretation of type theory. In Twenty-Five Years of Constructive Type Theory, Oxford Logic Guides 36, pages 83–111. Oxford University Press, 1998.

- [18]

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. Proceedings of the 37th International Conference on Machine Learning, PMLR 119, pages 4542–4551, 2020.

- [19]

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. Advances in Neural Information Processing Systems, volume 31, pages 8571–8580, 2018.

- [20]

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. International Conference on Learning Representations, 2015.

- [21]

Ivan Kolář, Jan Slovák, and Peter W. Michor. Natural Operations in Differential Geometry. Springer, 1993.

- [22]

Saunders Mac Lane and Ieke Moerdijk. Sheaves in Geometry and Logic: A First Introduction to Topos Theory. Universitext. Springer, 1992. Corrected reprint 1994. Internal language: Chapter VI.

- [23]

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, volume 32, 2019.

- [24]

Peter LeFanu Lumsdaine. Weak ω -categories from intensional type theory. *Logical Methods in Computer Science*, 6(3:24), 2010.

- [25]

Jacob Lurie. *Higher Topos Theory*. *Annals of Mathematics Studies* 170. Princeton University Press, 2009.

- [26]

Per Martin-Löf. *Intuitionistic Type Theory*. Bibliopolis, Naples, 1984. Notes by Giovanni Sambin.

- [27]

David Jaz Myers. *Categorical systems theory*. Book draft, last updated September 3, 2023.

- [28]

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, pages 7599–7609, 2020.

- [29]

J. P. Pridham. Unifying derived deformation theories. *Advances in Mathematics*, 224(3):772–826, 2010.

- [30]

D. J. Saunders. *The Geometry of Jet Bundles*. Cambridge University Press, 1989.

- [31]

David I. Spivak. Functorial data migration. *Information and Computation*, 217:31–51, 2012.

- [32]

The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study, 2013. See especially Chapters 2, 3, and 9.

- [33]

Benno van den Berg and Richard Garner. Types are weak ω -groupoids. *Proceedings of the London Mathematical Society*, 102(2):370–394, 2011.

- [34]

Vladimir Voevodsky. A very short note on homotopy λ -calculus. Unpublished note, 2006. Revised version dated 2009 in the author’s archive.

- [35]

Norbert Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press, second edition, 1961. First edition 1948.

- [36]

Greg Yang. *Tensor programs II: Neural tangent kernel for any architecture*. Preprint, 2020.