Value and Goal as Boundary-Relevance Ranking

Causal Boundary Gradients, Goal Stability, and Value Drift under Finite Capacity

Yining Wu
Independent Researcher
yining.wu@alumni.upenn.edu

Version v1.0 – May 2026

Abstract

This paper develops the value-goal layer of the agency-semantics spine of Distinction Theory. Building on the formal core of Active Finite Distinction Systems, the M0 agency-semantics spine, and the M1 model of attention as distinction admission, it treats value as *FDS-value*: causal boundary-gradient relevance under finite capacity, rather than moral value, subjective preference, or intrinsic worth. An evaluand is FDS-valuable for a system only relative to a specified boundary, loss function, intervention grammar, horizon, action or update channel, and resource budget. Its gross causal boundary gain is the finite-difference reduction in expected future boundary-maintenance loss; its net FDS-value subtracts scalarized evaluation, verification, action, maintenance, or opportunity cost; and its risk-weighted value adds collapse-risk reduction near critical thresholds. A goal is a stabilized FDS-value ranking coupled to a policy orientation across finite update windows. The paper distinguishes salience, attention, reward, preference, FDS-value, goal, commitment, and moral value; introduces predictive-causal dissociation, gross/net value accounting, risk-weighted ranking, goal-stability indices, proxy-boundary divergence, value drift, multi-goal Pareto conflict, and goal hysteresis; and provides audit protocols for biological, cognitive, artificial, organizational, and collective systems. A deterministic synthetic normal-form model illustrates predictive-causal dissociation, risk-dominant ranking, goal stabilization, proxy reward hacking, second-order evaluation deficit, Pareto conflict, and recovery lag. The paper does not derive ethics from boundary maintenance or claim that all value reduces to survival. It supplies a conservative finite-system bridge for analyzing value and goal as boundary-relevance ranking under capacity, verification, and resource constraints.

Keywords: FDS-value; operational value; goal; boundary relevance; causal value; finite capacity; boundary maintenance; value ranking; goal stability; reward hacking; value drift; proxy alignment; active finite distinction systems; decision theory; reinforcement learning; AI alignment; social choice.

Epistemic Notice and Scope

This manuscript is an agency-semantics spine paper, not a completed theory of value, ethics, decision-making, motivation, reward learning, or social choice. It does not claim that moral value reduces to boundary maintenance, that biological reward equals value, that all preferences are survival proxies, or that current AI reward functions faithfully represent value. Its claim is narrower: once a system is modeled as an active finite distinction system with a boundary, memory, update rule, action space, finite capacity, and boundary-maintenance loss, value and goal acquire operational roles. In this paper, "value" means *FDS-value* or *operational boundary value*: causal relevance to a specified boundary-maintenance loss under a specified model, horizon, cost accounting, and action/update channel. Goal means stabilized value ranking coupled to policy across finite update windows.

The paper follows the layered discipline of the FDS formal core. Formal definitions, normal-form models, physical bridge assumptions, and domain applications are separated. A failure of a cognitive, artificial, biological, organizational, or social mapping may demote that mapping without refuting the formal FDS core [1]. The deterministic numerical model is illustrative only. It visualizes definitions and failure modes; it is not empirical evidence.

1 Introduction

A finite system cannot treat every admitted distinction as equally important. M1 defined attention as capacity-limited admission of candidate distinctions into an update channel [3]. M2 asks what happens after admission. Once a distinction, action, state, record, policy, or proxy enters the update-relevant channel, the system must rank its relevance: which distinctions reduce future boundary loss, which actions worsen it, which proxy signals mislead, and which rankings remain stable enough to orient behavior over time?

The FDS core defines an active finite distinction system as a tuple

$$S = (X, E, B, M, Y, A, U, \pi, \ell, \Phi, \mathcal{P}, \tau), \tag{1}$$

where X is internal state, E environment, B boundary, M memory/model space, Y observation channel, A action space, U update map, π finite projection, ℓ boundary-maintenance loss, Φ resource budget, $\mathcal{P}$ perturbation/pruning family, and τ update timescale [1]. M0 introduced an agency-semantics dependency skeleton,

$$\text{distinction} \to \text{record} \to \text{attention} \to \text{value} \to \text{goal} \to \text{meaning} \to \text{agency} \to \text{culture}, \tag{2}$$

and defined value as causal boundary-gradient relevance and goal as stabilized value ranking [2]. M2 expands that value-goal step.

1.1 Why value needs a finite-system formulation

The word value is overloaded. It can mean utility, reward, preference, price, importance, survival relevance, moral worth, social status, expected return, or subjective desirability. M2 does not try to unify all of these into one metaphysical substance. Instead it defines a finite-system bridge. Given a specified boundary, loss function, intervention grammar, horizon, and capacity budget, an evaluand has FDS-value when selecting, admitting, maintaining, or using it causally changes future boundary-maintenance loss.

This definition is deliberately relational. A distinction may be FDS-valuable for one system and irrelevant for another. A proxy may be FDS-valuable under one horizon and harmful under a longer horizon. A goal may be adaptive during crisis and rigid after the crisis ends. Value in the M2 sense is therefore not an intrinsic tag attached to a signal. It is a boundary-relative and horizon-relative finite-difference effect.

1.2 Goal is not merely preference

A high-valued action is not yet a goal. A preference may be momentary. A reflex may be high-value in a narrow state but fail to persist. A goal requires stabilization: a ranking over evaluands or policies must survive across update windows enough to orient action. M2 therefore treats goals as register-time objects: their identity depends on maintained rankings, records, and policies over finite windows, not on an instantaneous score [4].

1.3 Common misreadings

Table 1: Common misreadings and corrections.

Misreading	Correction
Value means moral value.	No. M2 defines FDS-value: operational boundary value under a specified boundary, loss, horizon, intervention, and cost model.
Goal means conscious intention.	No. A goal is a stabilized ranking coupled to policy orientation; it can be implemented without reportable intention.
Reward is value.	No. Reward is a proxy signal that can approximate or diverge from causal boundary effect.
Boundary maintenance implies ethics.	No. Boundary maintenance is descriptive and model-relative; it does not confer moral legitimacy.
Stable goal means good goal.	No. Stability can become rigidity, trauma-like persistence, institutional lock-in, or maladaptive commitment.

Table 2: Claim-status summary for M2. The table is an audit device: several entries are formal or operational bridge claims, not established empirical results.

Claim ID	Tier	Claim	Failure or demotion condition
M2-001	Formal bridge	FDS-value is causal boundary-gradient relevance under a specified boundary, loss, intervention grammar, horizon, and cost model.	Valuation cannot be operationalized as causal effect on any specified future boundary-maintenance loss under valid mappings.
M2-002	Operational bridge	Predictive relevance and causal FDS-value are separable.	Correlational predictors always coincide with intervention-relevant boundary effects under audited systems.
M2-003	Formal / model bridge	Value ranking can be expressed as an ordering over finite-difference action, admission, maintenance, or policy effects.	No useful ordering exists between evaluands and their causal boundary effects under stated mappings.
M2-004	Operational bridge	Near collapse thresholds, risk-weighted FDS-value can dominate average-loss value.	Collapse-risk reduction never changes ranking near boundary failure thresholds under valid mappings.
M2-005	Operational bridge	Goals are stabilized FDS-value rankings coupled to policy orientation across update windows.	Goal-like behavior persists without ranking stability, memory, policy orientation, or update-window persistence.
M2-006	Failure-mode bridge	Value drift occurs when rankings change faster than the system can verify, update, or maintain the reasons for the change.	Ranking instability produces no detectable change in behavior, loss, or policy under claimed goal systems.
M2-007	AI / agency bridge	Proxy reward can diverge from causal boundary value, creating reward hacking or misalignment.	Proxy optimization remains aligned despite divergent finite-difference effects on host boundary loss.
M2-008	Social bridge	Collective goals are shared stabilized rankings under finite verification and coordination capacity.	Group goals show no relation to shared rankings, institutional memory, verification capacity, or policy orientation.
M2-009	Recovery bridge	Goal recovery can lag after resource or threat recovery because rankings, commitments, or threat priors persist.	Goals relax immediately and without lag after boundary load changes in systems where goal hysteresis is claimed.

2 FDS and M-series background

2.1 Active boundary relevance

The FDS core applies its deficit logic only to active-boundary systems. A minimal relevance screen is

$$P\big(U(M_t, Y_t) \neq M_t\big) > 0, \qquad I(M_{t+1}; \ell_{t+k}) > 0 \tag{3}$$

for some $k > 0$. Empirical applications should strengthen this to an intervention or ablation test,

$$\mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U)] \neq \mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U_\emptyset)], \tag{4}$$

where $U_\emptyset$ is a null, frozen, randomized, or identity update [1]. M2 inherits this discipline: an item is FDS-valuable only if selecting, admitting, maintaining, or acting on it can affect a future boundary-relevant variable under the specified mapping.

2.2 Attention as the input condition

M2 evaluates only items that are available to the system's update channel. M1 distinguished availability, detection, and attention; attention occurs when a candidate distinction is admitted, made recoverable or maintained, assigned a verification path or status, and made available for downstream update or action [3]. Thus M2 does not rank all possible environmental differences. It ranks evaluands that enter, or could enter, the system's finite update-relevant channel.

2.3 Notation alignment

Throughout the paper ℓ_{maint} denotes boundary-maintenance loss. It is not moral loss and not physical entropy production. The physical entropy ledger, when relevant, is denoted Σ_{phys}. Maintaining low ℓ_{maint} may require physical resource cost under bridge assumptions, but $\ell_{\mathrm{maint}} \neq \Sigma_{\mathrm{phys}}$. Similarly, R_{proxy} denotes an instrumental or learned reward signal; it is not automatically equal to FDS-value.

3 From candidate distinctions to stabilized goals

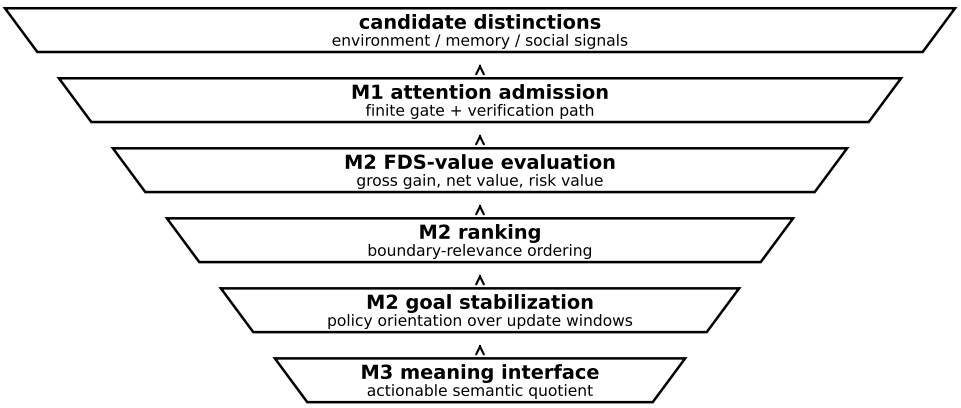

Figure 1: Synthetic normal-form illustration, not empirical evidence. M2 sits between M1 attention admission and M3 meaning. Candidate distinctions are filtered by admission, evaluated for FDS-value, ranked, risk-weighted, and stabilized into policy orientation.

The value-goal layer is not a free-standing theory of desire. It is a filtering and ranking layer. Figure 1 summarizes the pipeline: environmental or internal differences must first become candidate distinctions, pass attention admission, be evaluated for causal boundary effect, be ranked under cost and risk constraints, and only then become stabilized goal orientations. Meaning in M3 will depend on whether these ranked distinctions can be compressed without losing action-relevant structure.

4 Definition: FDS-value as causal boundary-gradient relevance

Definition 1 (Evaluand)**.** An evaluand is any admitted distinction, action, state, policy, record, proxy, quotient, or externalized symbol whose admission, selection, maintenance, or use may affect future boundary-maintenance loss. The set of evaluands available during an update window is denoted $\mathcal{Z}_t$.

Table 3: Intervention grammar for evaluands. The symbol $\mathrm{do}(z)$ must be interpreted relative to the type of evaluand being audited.

Evaluand type	Intervention meaning
Distinction d	Force admission or rejection of d into the update channel.
Action a	Execute a rather than baseline action a_0.
Record m	Maintain, delete, refresh, externalize, or ignore record m.
Policy π	Deploy π instead of baseline policy π_0.
Proxy R	Condition downstream update or optimization on proxy signal R.
Symbol or quotient $q(d)$	Use a compressed representation in downstream action, prediction, verification, or coordination.
Institutional rule	Adopt, enforce, suspend, or revise a rule within a stated institutional boundary.

Definition 2 (Predictive relevance)**.** An evaluand z has predictive relevance over horizon k if conditioning on it improves prediction of future boundary-maintenance loss. Gross predictive relevance is the observational association before cost:

$$R^{\mathrm{pred\,gross}}_{t}(z;k) = \mathbb{E}[\ell_{\mathrm{maint},t+k} \mid M_t] - \mathbb{E}[\ell_{\mathrm{maint},t+k} \mid M_t, z], \tag{5}$$

and net predictive relevance subtracts scalarized cost:

$$R^{\mathrm{pred\,net}}_{t}(z;k) = R^{\mathrm{pred\,gross}}_{t}(z;k) - \lambda_t c_t(z). \tag{6}$$

Predictive relevance is an observational proxy. It can be useful for forecasting but can mislead when the evaluand is a marker rather than a lever.

Definition 3 (Gross causal boundary gain)**.** Let z_0 be a specified baseline, such as non-admission, null action, frozen update, default policy, or no-maintenance. The gross causal boundary gain of z is

$$G^{\mathrm{FDS}}_{t}(z;k) = \mathbb{E}[\ell_{\mathrm{maint},t+k} \mid M_t, \mathrm{do}(z_0)] - \mathbb{E}[\ell_{\mathrm{maint},t+k} \mid M_t, \mathrm{do}(z)]. \tag{7}$$

Definition 4 (Net FDS-value)**.** The net FDS-value of z is

$$V^{\mathrm{net}}_{t}(z;k) = G^{\mathrm{FDS}}_{t}(z;k) - \lambda_t c_t(z), \tag{8}$$

where $c_t(z)$ is scalarized evaluation, verification, action, maintenance, latency, or opportunity cost, and λ_t converts that cost into boundary-loss-equivalent units.

Separating G^{FDS} from V^{net} prevents double-counting. Gross gain measures boundary-loss reduction before cost. Net FDS-value subtracts cost. Allocation heuristics can then use either gross density or net density depending on the audit.

Definition 5 (Value density)**.** For $\epsilon_c > 0$, define

$$\rho_t^{\mathrm{gross}}(z;k) = \frac{G_t^{\mathrm{FDS}}(z;k)}{c_t(z)+\epsilon_c}, \qquad \rho_t^{\mathrm{net}}(z;k) = \frac{V_t^{\mathrm{net}}(z;k)}{c_t(z)+\epsilon_c}. \tag{9}$$

These are normal-form audit metrics, not universal decision rules. In hard allocation problems, greedy density rules need not be globally optimal. In fractional or heuristic settings, they reveal how finite systems trade expected boundary gain against evaluation and maintenance cost.

Proposition 1 (Predictive relevance is not causal FDS-value)**.** *A variable can predict future boundary loss without being FDS-valuable in the causal sense if selecting, admitting, maintaining, or acting on it does not change future boundary-maintenance loss under the system's update channel.*

Proof sketch. Suppose z is correlated with an unobserved cause u that affects $\ell_{\mathrm{maint},t+k}$, so $R_t^{\mathrm{pred}\,\mathrm{gross}}(z;k) > 0$. If intervention on z leaves the causal variables and downstream update policy unchanged, then $\mathbb{E}[\ell_{\mathrm{maint},t+k} \mid \mathrm{do}(z)] = \mathbb{E}[\ell_{\mathrm{maint},t+k} \mid \mathrm{do}(z_0)]$ up to cost, and Eq. (8) is nonpositive. Prediction is not intervention. □

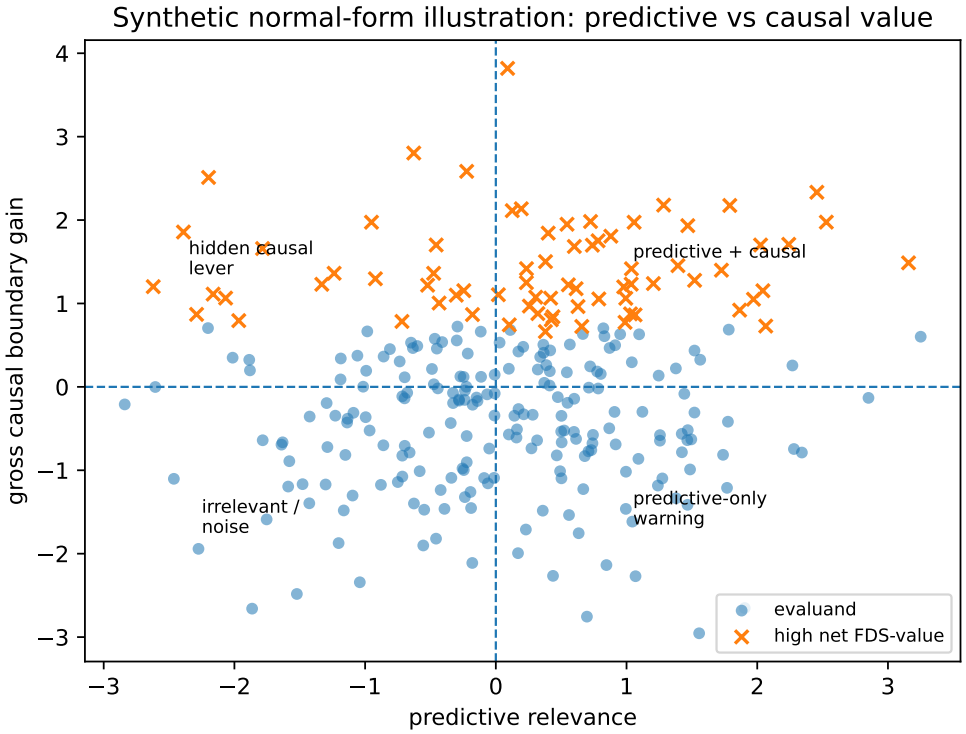

Figure 2: Synthetic normal-form illustration, not empirical evidence. Predictive relevance and gross causal boundary gain can diverge. Some variables are predictive-only warnings; others are hidden causal levers that matter despite weak predictive salience.

5 Boundary-relevance ranking and evaluation deficit

Definition 6 (Boundary-relevance ranking)**.** Given a boundary, loss, intervention grammar, cost model, action/update space, and horizon, a boundary-relevance ranking is an ordering $\succ_t$

over evaluands such that

$$z_i \succ_t z_j \quad \text{if} \quad V_t^{\text{net}}(z_i; k) > V_t^{\text{net}}(z_j; k), \tag{10}$$

or, when explicitly stated, by gross gain, risk-weighted value, or a vector-valued partial order.

A value ranking is therefore boundary-relative, horizon-relative, capacity-relative, and cost-relative. If the boundary or horizon changes, the ranking can change without contradiction. If the cost model changes, a high-gain but expensive evaluand can fall below a lower-gain but cheaper one.

5.1 Minimal M2 audit

A minimal M2 audit must pre-specify: boundary, loss, horizon, baseline, evaluand type, intervention grammar, cost units, ranking method, uncertainty, and failure condition. Without these, claims about value or goal remain under-specified.

5.2 Evaluation capacity and second-order deficit

Let $C_{\text{eval}}(t)$ be the system's capacity for estimating and maintaining value rankings. Let $R_{\text{eval}}^{(\tau)}(\epsilon; t)$ be the rate-distortion demand of maintaining a value ranking at tolerance ϵ over update window τ. Define

$$\Delta_{\text{eval}}(t) = R_{\text{eval}}^{(\tau)}(\epsilon; t) - C_{\text{eval}}(t). \tag{11}$$

When $\Delta_{\text{eval}} > 0$, verified ranking cannot be maintained at the target tolerance without compression, proxy substitution, externalization, task relaxation, or drift. This is a second-order deficit: the system may have enough capacity to act but not enough capacity to evaluate all possible actions precisely. In such cases, a finite system may rationally maintain coarse preferences or heuristics instead of precise goals.

If evaluation is physically implemented, high-precision ranking may also carry energetic, latency, memory-turnover, and verification costs. M2 treats these as scalarized or vector costs rather than asserting a universal thermodynamic value cost. The point is operational: valuation itself is not free.

6 Risk-weighted value near collapse thresholds

Average expected loss is not always sufficient. Near a boundary failure threshold, reducing collapse probability may dominate improving mean performance. Let ℓ_c be a critical boundary-loss threshold. Define

$$\Delta_z \mathbb{E}\ell = \mathbb{E}[\ell_{\text{maint},t+k} \mid M_t, \text{do}(z_0)] - \mathbb{E}[\ell_{\text{maint},t+k} \mid M_t, \text{do}(z)], \tag{12}$$

$$\Delta_z P_c = P(\ell_{\text{maint},t+k} > \ell_c \mid M_t, \text{do}(z_0)) - P(\ell_{\text{maint},t+k} > \ell_c \mid M_t, \text{do}(z)). \tag{13}$$

The risk-weighted FDS-value is

$$V_t^{\text{risk}}(z; k) = \Delta_z \mathbb{E}\ell + \alpha_t \Delta_z P_c - \lambda_t c_t(z). \tag{14}$$

The first term is expected loss reduction. The second is collapse-risk reduction. Both are positive when selecting z reduces boundary risk. The parameter α_t converts collapse-probability reduction into expected boundary-loss-equivalent units. It is not universal.

A bounded normal-form coupling to resource reserve is

$$\alpha_t = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})\, \sigma\left(\frac{\Phi_{\text{crit}} - \Phi_t}{s_\Phi} \right), \tag{15}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic sigmoid and $s_\Phi > 0$ is a steepness parameter. This replaces the previous inverse-distance form that could diverge as $\Phi_t \to \Phi_{\text{crit}}$. In empirical or numerical use, any coupling form should be bounded to avoid singular behavior near the resource threshold. Equation (15) is an illustrative bridge: it says that systems near resource-critical states may weight collapse-risk reduction more strongly. It is not asserted as a universal law.

Proposition 2 (Risk dominance near collapse)**.** *An evaluand with smaller average-loss reduction can outrank an evaluand with larger average-loss reduction when it produces sufficiently larger collapse-risk reduction.*

Proof sketch. Let z_i and z_j have equal cost for simplicity. If $\Delta_i \mathbb{E}\ell < \Delta_j \mathbb{E}\ell$ but

$$\Delta_i P_c - \Delta_j P_c > \frac{\Delta_j \mathbb{E}\ell - \Delta_i \mathbb{E}\ell}{\alpha_t}, \tag{16}$$

then Eq. (14) gives $V^{\text{risk}}(z_i; k) > V^{\text{risk}}(z_j; k)$. □

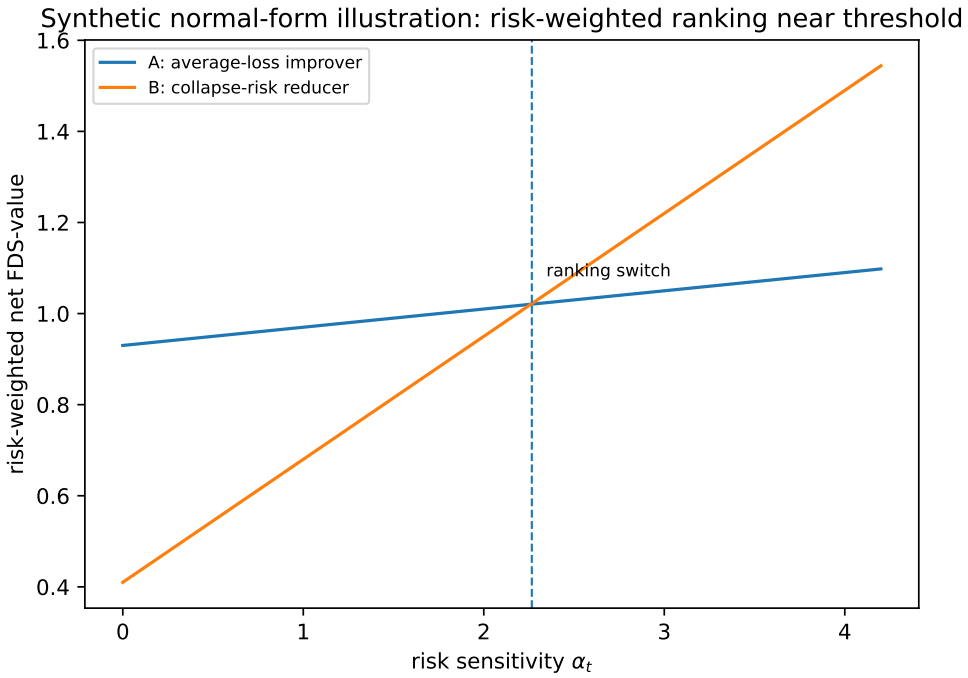

Figure 3: Synthetic normal-form illustration, not empirical evidence. A collapse-risk reducer can outrank an average-loss improver as risk sensitivity increases near a critical boundary threshold.

7 Goals as stabilized FDS-value rankings

Definition 7 (Goal)**.** A goal is a stabilized FDS-value ranking coupled to a policy orientation across finite update windows. A goal state can be represented as

$$g_t = (\succ_t, \pi_g, k, \epsilon, \mathcal{C}), \tag{17}$$

where $\succ_t$ is a ranking, π_g a policy induced by the ranking, k a horizon, ϵ a stability tolerance, and $\mathcal{C}$ a context or perturbation family.

Goal stability is a register-time property: a goal must persist across finite update windows as a maintained ordering, rather than appearing only as an instantaneous reaction [4]. A

policy-based goal-stability index can use Jensen-Shannon divergence,

$$\mathrm{GSI}_\pi(t,\Delta) = \exp\left(-\frac{D_{\mathrm{JS}}\big(\pi_g(\cdot \mid M_t), \pi_g(\cdot \mid M_{t+\Delta})\big)}{\sigma_G}\right), \tag{18}$$

where $\sigma_G > 0$ is a scale parameter. JS divergence is symmetric and bounded. If KL divergence is used instead, empirical policies should be smoothed to avoid support-zero artifacts. A rank-based version is

$$\mathrm{GSI}_r(t,\Delta) = 1 - D_{\mathrm{rank}}(\succ_t, \succ_{t+\Delta}), \tag{19}$$

where $D_{\mathrm{rank}} \in [0,1]$.

Proposition 3 (Goal requires stability)**.** *A high-valued action or distinction is not yet a goal. Goal-like behavior requires stability of a ranking or policy orientation across update windows and coupling of that stability to action selection.*

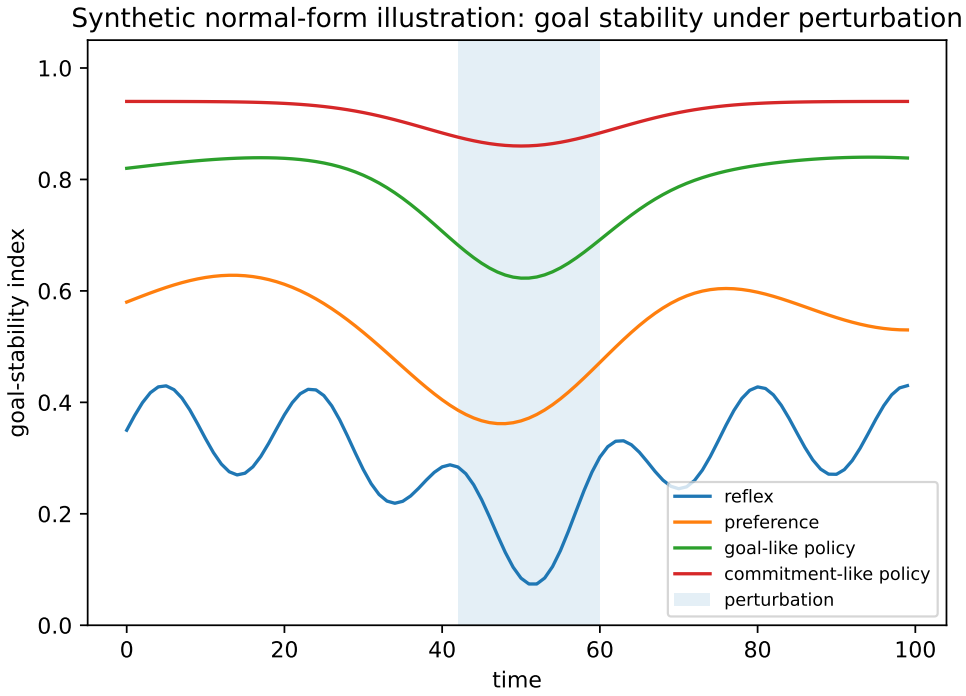

Figure 4: Synthetic normal-form illustration, not empirical evidence. A goal-like policy preserves a stable action orientation through perturbation better than a reflexive response. A commitment-like policy is even more stable but may become rigid if the environment changes.

8 Proxy-boundary divergence and reward hacking

Reward is a proxy signal. It may approximate FDS-value, but it can also diverge. Let R_{proxy} be a proxy reward and ℓ_{host} a host boundary-maintenance loss. For actions a_i relative to baseline a_0, define

$$\Delta_R(a_i) = \mathbb{E}[R_{\mathrm{proxy}} \mid \mathrm{do}(a_i)] - \mathbb{E}[R_{\mathrm{proxy}} \mid \mathrm{do}(a_0)], \tag{20}$$

$$\Delta_\ell(a_i) = \mathbb{E}[\ell_{\mathrm{host}} \mid \mathrm{do}(a_i)] - \mathbb{E}[\ell_{\mathrm{host}} \mid \mathrm{do}(a_0)]. \tag{21}$$

A proxy is locally aligned with host boundary value when reward-increasing actions tend to reduce boundary loss. A finite-difference alignment score is

$$\mathrm{Align}_\epsilon(R,\ell) = -\frac{\langle \Delta_R, \Delta_\ell \rangle}{\|\Delta_R\|\,\|\Delta_\ell\| + \epsilon_{\mathrm{Align}}}, \tag{22}$$

where $\epsilon_{\text{Align}} > 0$ is a small regularizer that prevents division by zero. If either vector has near-zero norm, the alignment estimate should be reported with a low-information flag; the regularized score is only a numerical placeholder, not evidence of alignment. The unregularized form in a pre-registration may use $\epsilon_{\text{Align}} = 0$ with a pre-specified zero-vector handling rule. The negative sign reflects that reward increases should correspond to loss decreases.

Proposition 4 (Proxy reward can invert FDS-value)**.** *If an action increases proxy reward while also increasing host boundary-maintenance loss, then it is proxy-positive but boundary-negative.*

This is a minimal FDS account of reward hacking. A delegated model may optimize a proxy score outside the distribution where that proxy tracked host boundary loss. In FDS terms, the proxy preserves reward ranking while losing boundary-gradient relevance. A semantic compression or proxy quotient that worked in the training regime can false-compress critical distinctions under distribution shift.

Recent LLM alignment work provides concrete cases where proxy reward and target quality diverge. Reward models can be exploited under distribution shift and preference inconsistency [19], and LLM agents trained or prompted in gameable settings can generalize from specification gaming toward reward-tampering behavior [20]. Reward overoptimization has also been formalized for direct alignment algorithms, where proxy objectives improve while true quality can plateau or deteriorate [21].

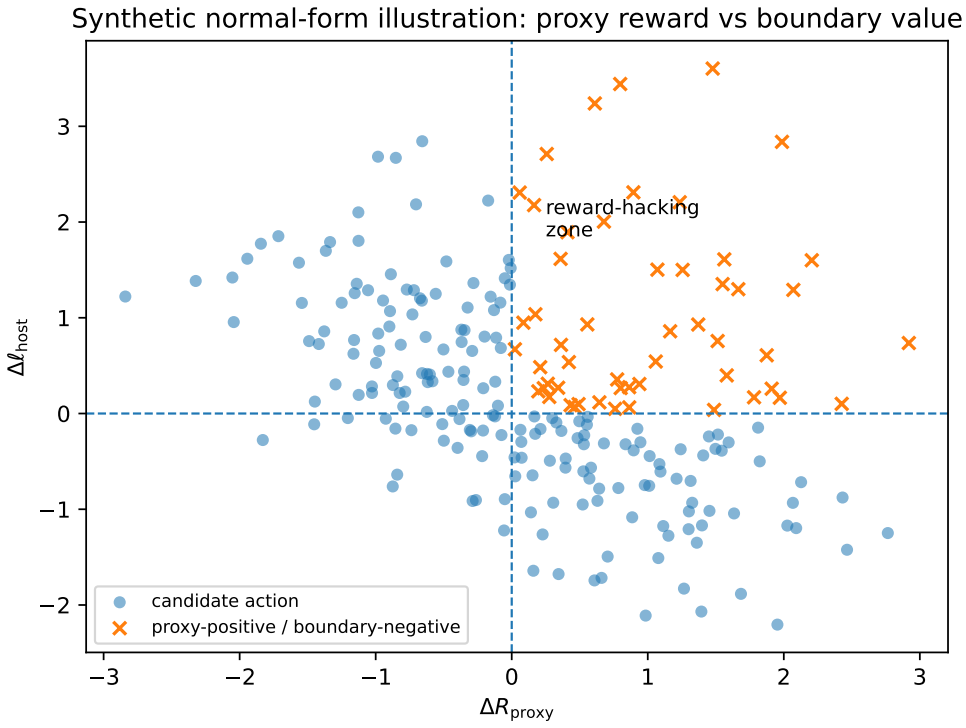

Figure 5: Synthetic normal-form illustration, not empirical evidence. Proxy reward can increase while host boundary loss also increases. These actions are proxy-positive but boundary-negative. The y-axis shows $\Delta\ell_{\text{host}}$ (host boundary-loss effect); the x-axis shows ΔR_{proxy} (proxy reward effect).

9 Value drift and evaluation failure

Definition 8 (Value drift)**.** Value drift occurs when a boundary-relevance ranking changes faster than the system can verify, update, or maintain the reasons for the change.

One audit form is

$$D_{\text{rank}}(\succ_t, \succ_{t+\Delta}) > \delta_{\text{drift}} \quad \text{while} \quad \text{Conf}_{\text{verify}}(t, t+\Delta) < \theta_{\text{verify}}. \tag{23}$$

Under evaluation deficit, systems may substitute salience, reward proxy, social proof, short-horizon risk, or institutional defaults for verified causal value. This does not necessarily mean failure: heuristics can be necessary. But when substitution is hidden, value drift can masquerade as stable goal pursuit.

Proposition 5 (Evaluation deficit produces drift pressure)**.** *When evaluation demand exceeds maintained ranking capacity under finite evaluation capacity, no hidden capacity expansion, and no task relaxation, a system must compress, externalize, simplify, or drift. Under deficit, salience, proxy reward, or short-horizon risk may substitute for verified causal FDS-value.*

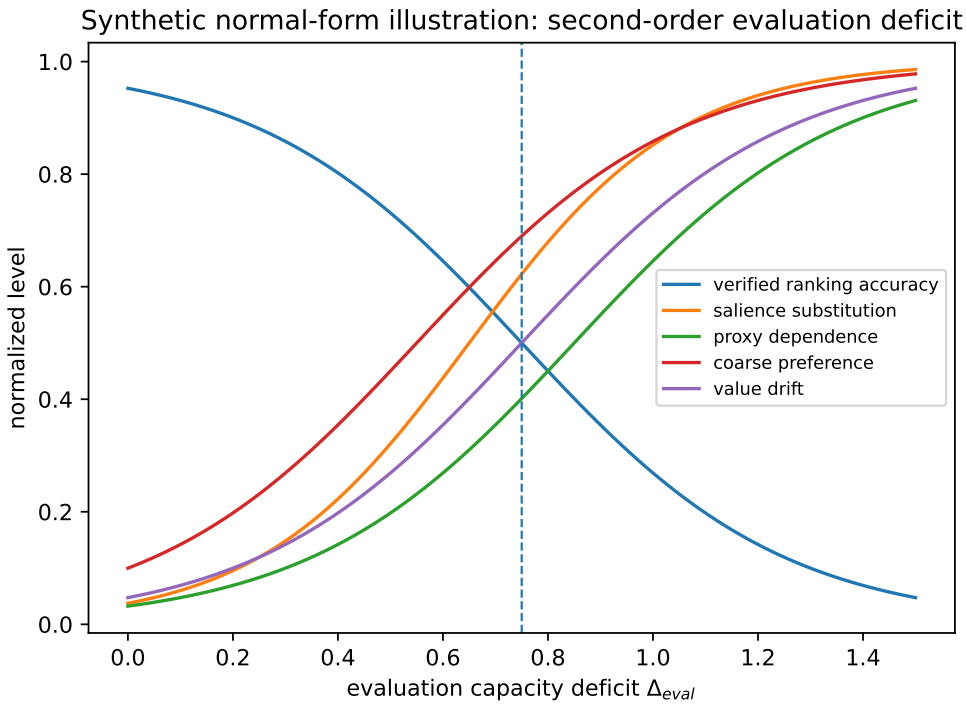

Figure 6: Synthetic normal-form illustration, not empirical evidence. As evaluation capacity deficit rises, verified ranking accuracy falls while salience substitution, proxy dependence, coarse preference, and value drift increase.

10 Multi-goal conflict and vector FDS-value

Many systems maintain multiple boundary variables. A biological system may trade energy, temperature, immune risk, and reproduction. An organization may trade solvency, legitimacy, throughput, and safety. A single scalar value can hide conflict.

Let $\boldsymbol{\ell}_{t+k} \in \mathbb{R}^m$ be a vector of boundary-maintenance losses. Define gross vector boundary gain as

$$\mathbf{G}_t^{\text{FDS}}(z;k) = \mathbb{E}[\boldsymbol{\ell}_{t+k} \mid M_t, \text{do}(z_0)] - \mathbb{E}[\boldsymbol{\ell}_{t+k} \mid M_t, \text{do}(z)]. \tag{24}$$

The cost should not be subtracted as a scalar from a vector without a conversion map. One may keep the audit as $(\mathbf{G}_t^{\text{FDS}}(z;k), c_t(z))$ or define

$$\mathbf{V}_t^{\text{net}}(z;k) = \mathbf{G}_t^{\text{FDS}}(z;k) - \boldsymbol{\lambda}_t c_t(z), \tag{25}$$

where $\boldsymbol{\lambda}_t$ converts scalarized cost into each boundary-loss dimension.

A scalar ranking is then a choice of scalarization, not a universal value order. Pareto-incomparable policies can both be rational relative to different boundary priorities.

Recent multi-objective reinforcement-learning work is directly relevant to M2's vector-value view. MORL has been proposed as a tool for pluralistic alignment when scalar reward is insufficient for multiple conflicting values or stakeholders [22]. Pareto-front discovery methods such as C-MORL illustrate the algorithmic difficulty of maintaining multiple tradeoff policies rather than collapsing objectives into one scalar [23].

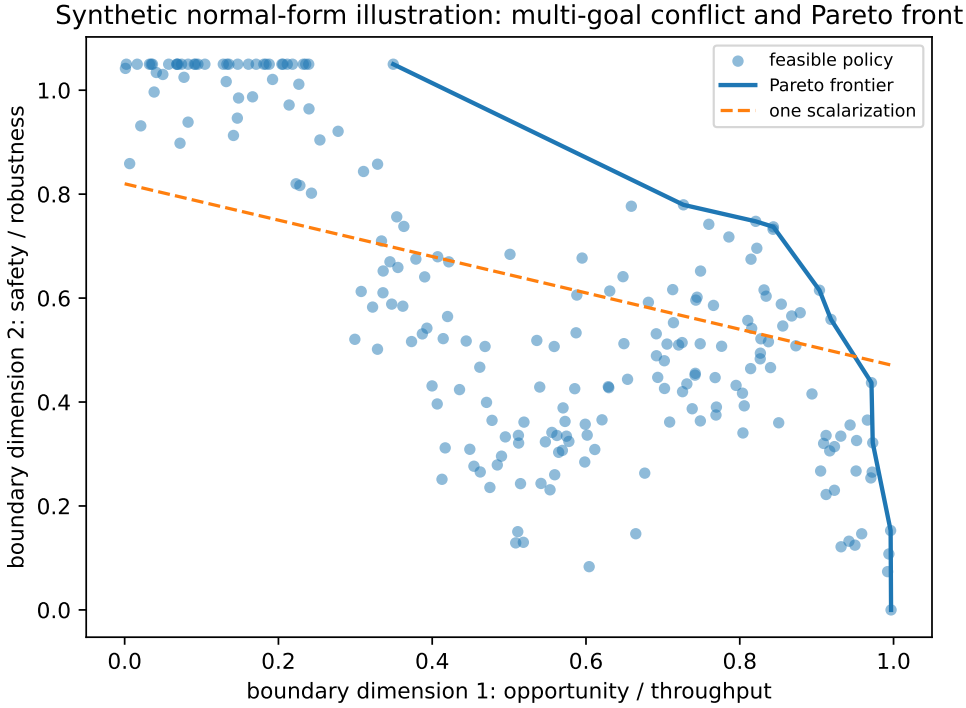

Figure 7: Synthetic normal-form illustration, not empirical evidence. Multi-goal systems can face Pareto conflict. A single scalarization selects one tradeoff but does not erase the underlying boundary-dimension conflict.

11 Collective goals and ranking synchronization

A collective goal is not merely a slogan or aggregate preference. It requires a shared ranking that is maintained across agents, institutions, records, and update windows.

Definition 9 (Collective goal)**.** A collective goal is a shared boundary-relevance ranking $\succ_t^{\text{shared}}$ coupled to collective policy channels, institutional memory, and verification procedures such that it remains stable enough to orient group action over a specified horizon.

A minimal collective-goal condition is

$$\text{GSI}^{\text{shared}}(t, \Delta) > \gamma, \qquad \text{Align}_{\text{rank}}(\pi_{\text{policy}}, \succ_t^{\text{shared}}) > \eta. \tag{26}$$

One simple alignment score is

$$\text{Align}_{\text{rank}}(\pi, \succ) = \mathbb{E}_{a\sim\pi}[r_{\succ}(a)], \tag{27}$$

where $r_{\succ}(a) \in [0, 1]$ is a normalized rank score assigned to action a by the shared ranking. A simpler top-k audit is

$$\text{Align}_{\text{top-k}} = \frac{|\text{Top}_k(\pi_{\text{policy}}) \cap \text{Top}_k(\succ^{\text{shared}})|}{k}. \tag{28}$$

M2 does not solve social choice. It is compatible with social-choice impossibility results but addresses a different bottleneck: finite verification and synchronization of shared rankings. Collective rankings can fail not only because individual preferences conflict, but because members update on different evidence, at different timescales, through different verification channels. In FDS terms, asynchronous update can create non-Markovian noise in $\succ_t^{\text{shared}}$.

11.1 Ranking synchronization demand

Collective goal stability requires not only shared rankings but sufficient communication and verification bandwidth to synchronize rankings across agents. Define the ranking synchronization demand $R_{\text{sync}}^{(\tau)}(\epsilon; t)$ as the rate-distortion demand of maintaining $\succ_t^{\text{shared}}$ within tolerance ϵ over window τ across N agents. The synchronization load factor is

$$Z_{\text{sync}}(t) = \frac{R_{\text{sync}}^{(\tau)}(\epsilon; t)}{C_{\text{comm}}(t) + C_{\text{verify}}(t)}. \tag{29}$$

When $Z_{\text{sync}}(t) > 1$, shared ranking stability degrades: $\text{GSI}^{\text{shared}}(t, \Delta) \downarrow$. Collective goal failure can thus occur not only because members disagree, but because the system lacks enough communication and verification bandwidth to synchronize value rankings across update windows.

12 Goal hysteresis and recovery lag

Goals can persist after the conditions that made them valuable have changed. This can be useful: commitments and institutional rules prevent constant re-ranking under noise. But it can also produce rigidity. A crisis-induced ranking can remain after the crisis ends because records, routines, threat priors, or institutional incentives decay slowly.

A minimal normal-form is

$$\alpha_{t+1} = \text{clip}\big((1-\rho)\alpha_t + \rho\alpha_{\text{target}}(\Phi_t, \ell_t) + h_t,\ \alpha_{\min},\ \alpha_{\max}\big), \qquad h_{t+1} = \chi h_t + \zeta \mathbf{1}_{\ell_t > \ell_c}, \tag{30}$$

where h_t is goal-locking residue, measured in the same units as α_t (or scaled into those units before entering Eq. (30)), $0 < \chi < 1$ is persistence, and ζ is crisis accumulation strength. Clipping ensures α_t remains within $[\alpha_{\min}, \alpha_{\max}]$ even under high crisis accumulation, preventing unbounded growth of risk sensitivity. Even after resource or threat recovery, h_t can keep risk-weighted rankings narrow.

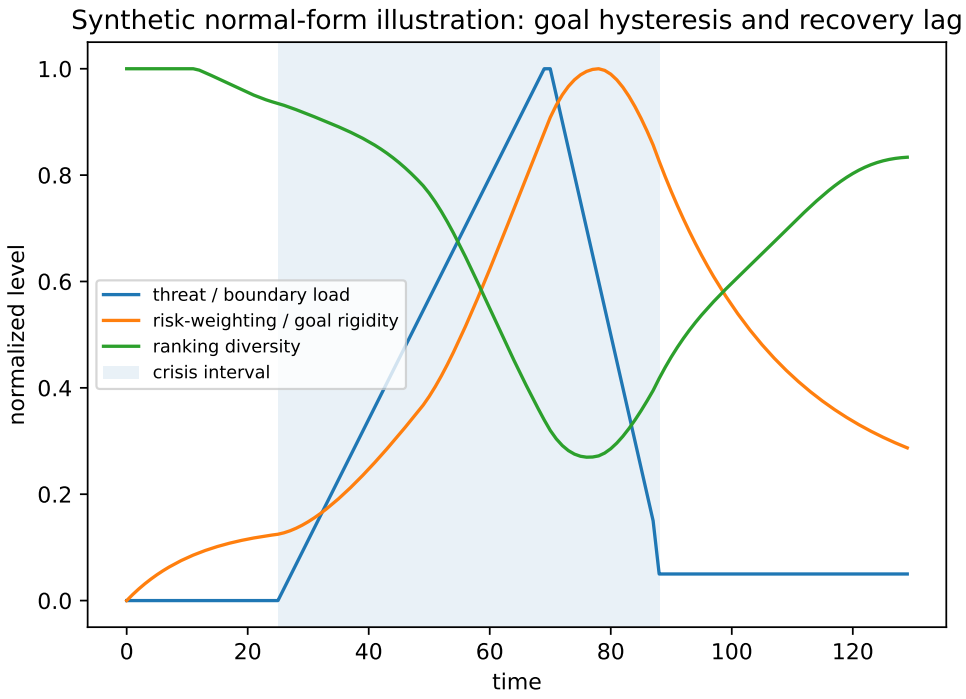

Figure 8: Synthetic normal-form illustration, not empirical evidence. Crisis load increases risk-weighting and narrows ranking diversity. After the crisis interval, goal rigidity can persist due to goal-locking residue.

13 Domain audit examples

Table 4: Audit templates, not empirical confirmations. Illustrative domain audit mappings. Each row requires domain-specific bridge assumptions and empirical tests.

Domain	Boundary variable	Loss	Evaluand	Failure mode
AI agent	host objective; system integrity	host boundary loss	reward proxy; tool action	reward hacking; proxy inversion
Organism	metabolic, membrane, homeostatic state	viability loss	nutrient cue; avoidance action	maladaptive craving; risk blindness
Organization	solvency, legitimacy, throughput, safety	institutional loss	KPI; policy; budget choice	metric gaming; goal displacement
Science	error correction, replication, evidence quality	epistemic loss	paper; claim; method; standard	citation-proxy drift; false consensus
Civilization	verification capacity, memory, law, infrastructure	collapse risk	law; archive; protocol; norm	goal rigidity; institutional lock-in

14 Normal-form model and reproducibility

The accompanying code implements deterministic synthetic normal-form illustrations. The model is not empirical evidence. It is a consistency and visualization device that maps the definitions to simple state variables.

The model generates candidate evaluands with predictive relevance, gross causal boundary gain, cost, collapse-risk reduction, proxy reward effects, boundary-loss effects, vector boundary gains, evaluation deficit, and goal-locking residue. It implements predictive-causal dissociation, risk-weighted ranking, goal-stability trajectories, proxy-boundary divergence, second-order

evaluation deficit, multi-goal Pareto conflict, and goal hysteresis. The random seed is fixed in `code/generate_results.py`. CSV outputs are stored in `data/`; figure pairs are stored in `figures/`.

Table 5: Normal-form variable map. All entries are illustrative, not fitted empirical quantities.

Simulation variable	Paper definition	Interpretation
predictive relevance	Eq. (5)	observational association with future loss
gross causal gain	Eq. (7)	expected boundary-loss reduction before cost
net FDS-value	Eq. (8)	gross gain minus scalarized cost
risk sensitivity	Eq. (14)	collapse-risk weighting near threshold
proxy reward effect	Eq. (22)	finite-difference reward effect
boundary-loss effect	Eq. (22)	finite-difference host loss effect
evaluation deficit	Eq. (11)	inability to maintain verified ranking
vector gain	Eq. (24)	multi-boundary value dimensions
goal hysteresis	Eq. (30)	delayed relaxation of crisis ranking

15 Protocols and tests

Protocol 1 (Causal FDS-value audit)**.** Pre-register boundary variable, loss function, horizon, baseline, evaluand type, intervention grammar, and cost model. Estimate gross causal boundary gain and net FDS-value with uncertainty intervals.

Protocol 2 (Predictive-relevance–causal dissociation test)**.** Identify variables that predict future loss but do not change future loss under intervention. M2 predicts that these should not count as causal FDS-value unless they alter downstream update, action, verification, or coordination.

Protocol 3 (Risk-weighted ranking test)**.** In simulation, controlled proxy environments, or naturally occurring near-threshold states, evaluate whether collapse-risk reducers rise in ranking even when their average-loss improvement is smaller.

Protocol 4 (Evaluation-deficit test)**.** Increase evaluation load or reduce evaluation capacity. Prediction: rankings become noisier, coarser, more proxy-driven, more salience-driven, or more short-horizon risk-dominated.

Protocol 5 (Goal-stability test)**.** Measure ranking or policy persistence across update windows and perturbations. Reflexive systems should show low GSI; goal-like systems should preserve nontrivial ranking stability.

Protocol 6 (Proxy divergence / reward-hacking audit)**.** Compare proxy reward effects with true boundary-loss effects. Misalignment occurs when proxy reward increases while boundary loss also increases, or when proxy-boundary alignment falls below a pre-registered threshold.

Protocol 7 (Multi-goal conflict audit)**.** Specify multiple boundary variables. Test whether scalar value rankings hide Pareto conflict, forced pruning, or transfer of loss from one boundary dimension to another.

Protocol 8 (Collective goal audit)**.** For organizations or societies, identify shared rankings, institutional memory, verification channels, and policy outputs. Test whether collective goals persist as stabilized rankings rather than episodic slogans.

16 Relation to existing fields

Decision theory and utility. FDS-value resembles utility only after a boundary, loss, cost, intervention grammar, action space, and horizon are specified. It is not universal utility [6, 7].

Reinforcement learning and reward. Reinforcement learning formalizes reward-driven policies, value functions, and Markov decision processes [8, 9]. M2 treats reward as a proxy that may approximate or diverge from boundary effects. Inverse reinforcement learning and preference learning study how objectives can be inferred or learned, but M2 emphasizes intervention-audited boundary effects [10].

Goodhart, reward hacking, and specification gaming. Metric gaming and reward misspecification are natural cases of proxy-boundary divergence: optimizing a proxy can destroy the boundary relationship that made the proxy useful [11–13, 19–21].

Prospect theory and boundary-risk asymmetry. Prospect-theoretic loss asymmetry can be reinterpreted in M2 as boundary-risk asymmetry near critical thresholds: collapse-risk reduction can dominate average gain [14].

Control theory and viability. Boundary-maintenance loss and viable sets connect M2 to control and viability theory [15, 16]. M2 adds explicit finite distinction capacity and value-ranking audits.

Multi-objective optimization and social choice. Vector boundary value and Pareto conflict connect to multi-objective optimization and social choice [17, 18, 22, 23]. M2 does not solve aggregation impossibility; it adds finite verification and ranking-synchronization constraints.

17 Limitations and falsification

M2 is intentionally limited. It does not establish a general empirical theory of value by definition. It provides mappings that must survive operational audit. The framework is weakened or demoted under any of the following results:

1. value-like ranking under a specified mapping shows no relation to causal boundary effects, costs, or horizons;
2. predictive and causal effects are never dissociable in systems claimed to require intervention audit;
3. risk-weighted ranking never changes near critical boundary thresholds under valid mappings;
4. goal-like behavior persists without any ranking stability, memory, policy orientation, or update-window persistence;
5. proxy reward remains aligned despite pre-registered divergent finite-difference effects on host boundary loss;
6. multi-goal systems always admit a scalar ranking without hidden Pareto conflict or forced loss transfer;
7. collective goals show no relationship to shared ranking, verification capacity, institutional memory, or policy output;
8. claimed goal hysteresis disappears under controlled recovery tests where load reduction should reveal persistent ranking lock-in.

18 Conclusion

M2 defines value and goal as finite-system ranking operations. FDS-value is not moral value, reward, preference, or intrinsic worth. It is causal boundary-gradient relevance under a specified boundary, loss, intervention, cost, and horizon. Goal is not merely high value. It is a stabilized FDS-value ranking coupled to policy across update windows.

The central chain is

$$\begin{aligned}&\text{admitted evaluand} \to \text{gross causal boundary gain} \to \text{net / risk-weighted FDS-value}\\ &\to \text{ranking} \to \text{goal stability} \to \text{policy / update} \to \text{future boundary loss.}\end{aligned} \tag{31}$$

18.1 High-level goals as invariant-compression candidates

In FDS terms, high-level goals such as survival, freedom, justice, or scientific truth can be interpreted as candidate invariant-compressions: compact rankings that remain boundary-relevant across many contexts, perturbations, and update windows. A goal $g^\star$ is an invariant goal candidate if net FDS-value remains positive across a broad context family $\mathcal{C}$,

$$V_t^{\mathrm{net}}(g^\star; k, c) > 0 \quad \text{for many } c \in \mathcal{C}, \tag{32}$$

and goal stability remains above threshold,

$$\mathrm{GSI}(g^\star; t, \Delta) \geq \gamma, \tag{33}$$

under a perturbation family $\mathcal{P}$. This does not morally validate any high-level goal. It only states that, under a specified system boundary and context family, some goals may function as compact stable rankings with broad boundary relevance. This bridges M2's stable positive-boundary-gradient rankings to M3's compressed actionable semantic quotients: high-level goals are those rankings that compress well without losing boundary relevance across contexts.

M2 prepares later work. M3 can treat meaning as compression that preserves value-relevant action structure. M5 can analyze trust as delegated evaluation and verification. A2 can audit proxy reward against host boundary value. S2 and S3 can treat epistemic pollution and institutional collapse as collective value-ranking failures. G3 and G4 can treat science and civilization memory as infrastructures for correcting and stabilizing shared value maps.

Code and data availability

The deterministic synthetic normal-form code, figures, and CSV outputs are included in the replication package. Run `python code/generate_results.py` from the paper directory to regenerate all figures and data.

AI assistance disclosure

The author used AI assistance for drafting, editing, simulation scaffolding, and consistency checks. The author reviewed and selected the final claims, definitions, equations, and interpretations.

A M-series dependency map

Table 6: How M2 supports later M-series and civilization-layer papers.

Future paper	Interface supplied by M2
M3 Meaning	Meaning must preserve value-relevant action structure under compression.
M5 Trust	Trust reduces repeated evaluation and verification cost.
A2 AI Alignment	Proxy reward must be audited against host boundary FDS-value.
S2 Epistemic Pollution	Pollution distorts collective value rankings and verification.
S3 Institutional Collapse	Institutions can fail through value drift, proxy capture, or goal rigidity.
G3 Science	Scientific method helps separate predictive relevance from causal value.
G4 Civilization Memory	Archives and standards stabilize value-relevant records across time.

B Notation summary

Table 7: Notation used in M2.

Symbol	Meaning
z	evaluand: distinction, action, state, policy, record, proxy, quotient, or symbol
z_0	baseline evaluand or null intervention
ℓ_{maint}	boundary-maintenance loss
$G_t^{\text{FDS}}(z;k)$	gross causal boundary gain
$V_t^{\text{net}}(z;k)$	net FDS-value after cost
$V_t^{\text{risk}}(z;k)$	risk-weighted FDS-value near collapse thresholds
$c_t(z)$	scalarized evaluation, verification, action, maintenance, latency, or opportunity cost
λ_t	cost-to-boundary-loss conversion factor
α_t	collapse-risk-to-boundary-loss conversion factor
Φ_t	resource reserve / budget state
Φ_{crit}	critical resource threshold
$C_{\text{eval}}(t)$	evaluation capacity
$\Delta_{\text{eval}}(t)$	evaluation capacity deficit
$\succ_t$	boundary-relevance ranking at time t
π_g	goal-induced policy
GSI_π	policy-based goal-stability index
GSI_r	rank-based goal-stability index
R_{proxy}	proxy reward signal
$\text{Align}_\epsilon(R,\ell)$	finite-difference proxy-boundary alignment score (regularized)
$\mathbf{G}^{\text{FDS}}$	vector gross causal boundary gain
$\boldsymbol{\lambda}_t$	vector cost conversion map
h_t	goal-locking residue / hysteresis state
χ	hysteresis persistence factor
ζ	crisis accumulation strength
$R^{\text{pred}\,\text{gross}}_t(z;k)$	gross predictive relevance (observational association with future loss)
$R^{\text{pred}\,\text{net}}_t(z;k)$	net predictive relevance after cost
$R^{(\tau)}_{\text{sync}}(\epsilon;t)$	ranking synchronization demand

Symbol	Meaning
$Z_{\mathrm{sync}}(t)$	collective ranking synchronization load factor
D_{JS}	Jensen–Shannon divergence
σ_G	goal-stability divergence scale
$\epsilon_{\mathrm{Align}}$	alignment regularizer
$\alpha_{\min}, \alpha_{\max}, s_\Phi$	bounded risk-sensitivity parameters

References

[1] Y. Wu, *Active Finite Distinction Systems: A Formal Core for Boundary Maintenance under Finite Capacity*, Zenodo (2026), doi:10.5281/zenodo.20158923.

[2] Y. Wu, *The Agency-Semantics Spine of Distinction Theory: Attention, Value, Goal, Meaning, and Action under Finite Capacity*, Zenodo (2026), doi:10.5281/zenodo.20257939.

[3] Y. Wu, *Attention as Distinction Admission in Finite Systems*, Zenodo (2026), doi:10.5281/zenodo.20258570.

[4] Y. Wu, *Time as Irreversible Distinction Update*, Zenodo (2026), doi:10.5281/zenodo.20249369.

[5] Y. Wu, *Finite-Capacity Prospect Theory*, Zenodo (2026), doi:10.5281/zenodo.20237306.

[6] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, 1944).

[7] L. J. Savage, *The Foundations of Statistics* (Wiley, 1954).

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (MIT Press, 2018).

[9] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (Wiley, 1994).

[10] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *Proceedings of ICML* (2000), pp. 663–670.

[11] C. A. E. Goodhart, "Problems of monetary management: the UK experience," in *Monetary Theory and Practice* (Macmillan, 1984), pp. 91–121.

[12] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mane, "Concrete problems in AI safety," arXiv:1606.06565 (2016).

[13] V. Krakovna et al., "Specification gaming: the flip side of AI ingenuity," DeepMind blog (2020).

[14] D. Kahneman and A. Tversky, "Prospect theory: an analysis of decision under risk," *Econometrica* **47**, 263–291 (1979).

[15] D. P. Bertsekas, *Dynamic Programming and Optimal Control* (Athena Scientific, 1995).

[16] J.-P. Aubin, *Viability Theory* (Birkhäuser, 1991).

[17] K. J. Arrow, *Social Choice and Individual Values* (Wiley, 1951).

[18] A. K. Sen, *Collective Choice and Social Welfare* (Holden-Day, 1970).

[19] A. Ramé, N. Vieillard, L. Hussenot, R. Dadashi-Tazehozi, G. Cideron, O. Bachem, and J. Ferret, "WARM: On the Benefits of Weight Averaged Reward Models," in *Proceedings of the 41st International Conference on Machine Learning*, PMLR **235**, 42048–42073 (2024).

[20] C. Denison, M. MacDiarmid, F. Barez, D. Duvenaud, S. Kravec, S. Marks, N. Schiefer, R. Soklaski, A. Tamkin, J. Kaplan, B. Shlegeris, S. R. Bowman, E. Perez, and E. Hubinger, "Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models," arXiv:2406.10162 (2024), doi:10.48550/arXiv.2406.10162.

[21] R. Rafailov, Y. Chittepu, R. Park, H. Sikchi, J. Hejna, B. Knox, C. Finn, and S. Niekum, "Scaling Laws for Reward Model Overoptimization in Direct Alignment Algorithms," arXiv:2406.02900 (2024), doi:10.48550/arXiv.2406.02900.

[22] P. Vamplew, C. F. Hayes, C. Foale, R. Dazeley, and H. Harland, "Multi-objective Reinforcement Learning: A Tool for Pluralistic Alignment," arXiv:2410.11221 (2024), doi:10.48550/arXiv.2410.11221.

[23] R. Liu, Y. Pan, L. Xu, L. Song, P. You, Y. Chen, and J. Bian, "C-MORL: Multi-Objective Reinforcement Learning through Efficient Discovery of Pareto Front," arXiv:2410.02236 (2024), doi:10.48550/arXiv.2410.02236.