

The Theory of Life with 0 and 1

A Formal Model of Moral Direction
and Character Formation

Raphael Lugmayr

Working Paper, Version 1.0
7 October 2026

Abstract

This paper formalizes the proposal that morally significant choices have a direction and that repeated choices can reshape the dispositions from which future choices arise. An objective moral good is posited as a normative reference, rather than derived from mathematics. The symbols 0 and 1 encode opposed directions relative to that reference; they do not classify persons into two exhaustive kinds. Moral intensity, character, and choice probabilities remain continuous and potentially multidimensional. The framework distinguishes objective evaluation, fallible moral judgment, and the psychological process of character formation. A bounded recursive model couples choices to character updates and subsequent choice distributions. Conditional propositions establish invariance, memory weighting, robustness, contraction, feedback-induced multiplicity, and the possibility of trajectory reversal. Neutrality, conflicting virtues, intentions, context, and uncertainty are represented explicitly. A Markov formulation states the additional sufficiency assumptions needed for a stochastic interpretation. Illustrative examples and philosophical comparisons clarify the model without supplying empirical validation. The contribution is a disciplined formal vocabulary and a set of transparent consequences of stated assumptions, not a proven law of life, a moral measurement instrument, or a derivation of ethics from behavior.

Keywords: moral direction; character formation; virtue; habit; moral uncertainty; dynamical systems.

1 Introduction

Decisions affect both what happens in the world and how a person is disposed to act again. An honest disclosure may resolve an immediate problem while also making later honesty easier. Repeated avoidance may provide short-term relief while consolidating an avoidant response. This paper develops a formal interpretation of that recursive relationship:

$$\text{choice} \longrightarrow \text{character update} \longrightarrow \text{future choice distribution.} \quad (1)$$

The motivating intuition is directional: morally relevant action can orient a person toward the good or away from it. The title's 0/1 language expresses this polarity. It does not entail that all decisions are moral, that every dilemma admits a scalar answer, or that the worth of a person is a binary variable.

Three claims must be separated. First, a *normative claim* holds that there is an objective moral reference independent of an individual's approval. Second, a *psychological hypothesis* holds that practice can modify dispositions and thereby influence subsequent conduct. Third,

a *mathematical claim* states what follows from a specified transition rule. Only the third kind is proved here. Evidence for a learning process would not establish moral realism, and a proof about a recurrence would not establish that the recurrence describes human beings.

The paper’s contribution is a synthesis and clarification, rather than a claim to have discovered habituation or a new class of dynamical systems. Its basic ingredients—bounded states, weighted updates, probabilistic choice, and feedback—are familiar. Their joint use makes several otherwise hidden assumptions explicit: what counts as a moral direction, how neutral choices differ from unresolved judgments, when repetition supports improvement, and why strong habits need not imply irreversible character.

The mathematical framework occupies the central part of the paper. Sections 2–3 define the objects and the dynamics. Section 4 proves conditional results, and Section 5 provides fully specified examples. Philosophical sources and modern learning research provide interpretation in Section 6; they are not offered as evidence for the chosen equations. The final sections identify objections and an empirical research agenda. Throughout, *character* refers to selected dispositions, not a complete description of a self, an allocation of dignity, or a verdict on a life.

2 Definitions and normative assumptions

2.1 Events, agency, and the moral reference

Time $t \in \{0, 1, 2, \dots\}$ indexes modeled decision occasions, not days. Let $c_t \in \mathcal{C}$ describe the context, including available information, relationships, obligations, resources, and constraints. The nonempty finite set $\mathcal{A}(c_t)$ contains the actions available in that context; an omission can be an action when it is a meaningful available alternative.

Definition 2.1 (Decision event). *A decision event is $e_t = (c_t, a_t, i_t, o_t)$, where $a_t \in \mathcal{A}(c_t)$, i_t describes intention and relevant knowledge, and o_t records the outcome. The event description must distinguish an outcome from its foreseeability and an action from the degree of voluntary control under which it is performed.*

An intention may be included in an action plan or treated as latent. The notation does not presume that an observer can recover it from behavior. Coercion, illness, ignorance, and restricted opportunities belong in the event description whenever they affect evaluation or learning. Moral quality and culpability are distinct questions; the model does not infer the latter from an action score.

Axiom 2.1 (Objective normative reference). *There is a moral reference $\mathcal{G}$ such that at least some judgments of better and worse are true or false independently of the evaluating agent’s endorsement. The reference can depend on morally relevant facts of an event without depending on approval merely as approval.*

This is a philosophical commitment of the paper, not a mathematical theorem. It leaves open the grounding of the good and does not establish a complete ordering of all actions. Context dependence is compatible with objectivity: a fixed principle may prescribe different conduct under different needs, promises, or risks. An observer’s uncertainty about those facts does not, by itself, change the reference.

Axiom 2.2 (Local numerical representation). *On a specified domain of events, selected moral considerations admit a bounded vector representation*

$$gg(e) = rg(e)vg(e) \in [-1, 1]^d, \quad rg(e) \in [0, 1], \quad vg(e) \in [-1, 1]^d. \quad (2)$$

Positive coordinates represent accordance with the relevant good and negative coordinates represent departure from it. The coordinate scales and their zero points are declared conventions.

The factor $r_{\mathcal{G}}$ records relevance or scope; $v_{\mathcal{G}}$ records the signed profile within that scope. Their product is the operative quantity. The factorization is not unique and is not claimed to be identifiable from observations. In an application, one could set $r_{\mathcal{G}} = 1$ on the entire relevant domain and use $g_{\mathcal{G}}$ directly.

The numerical representation is a stronger assumption than moral realism. Objective reasons could exist without possessing cardinal magnitudes or being mutually commensurable. Where a defensible representation is unavailable, the scalar constructions below must be withheld.

2.2 Two directions and continuous intensity

Definition 2.2 (Scalarization and directional code). *For a declared weight vector*

$$w \in \Delta^{d-1} := \{w \in \mathbb{R}^d : w_j \geq 0, \sum_j w_j = 1\}, \quad s_{\mathcal{G}}(e; w) = w^\top g_{\mathcal{G}}(e), \quad (3)$$

define, only when $s_{\mathcal{G}}(e; w) \neq 0$,

$$b_{\mathcal{G}}(e; w) = \begin{cases} 1, & s_{\mathcal{G}}(e; w) > 0, \\ 0, & s_{\mathcal{G}}(e; w) < 0, \end{cases} \quad m_{\mathcal{G}}(e; w) = |s_{\mathcal{G}}(e; w)|. \quad (4)$$

Then $s_{\mathcal{G}} = m_{\mathcal{G}}(2b_{\mathcal{G}} - 1)$.

The bit specifies direction and the real number specifies magnitude. A score of 0.02 and a score of 0.8 share a direction without sharing an intensity. The bit $b_{\mathcal{G}} = 0$ denotes a negative direction; it must not be confused with the numerical score $s_{\mathcal{G}} = 0$. Neither $b_{\mathcal{G}}$ nor $m_{\mathcal{G}}$ is a probability.

The weights are an explicit normative modeling choice. Unless justified as part of a defensible representation of $\mathcal{G}$, the sign of $s_{\mathcal{G}}$ is only a *conditional assessment under w* , not an established all-things-considered moral verdict. A binary code contains no argument for its own normative correctness.

Definition 2.3 (Neutral, balanced, and unresolved cases). *An event is neutral in the represented domain when $g_{\mathcal{G}}(e) = 0$, including $r_{\mathcal{G}}(e) = 0$. It is balanced under w when $w^\top g_{\mathcal{G}}(e) = 0$ but $g_{\mathcal{G}}(e) \neq 0$. An assessment is unresolved when available information or normative commitments do not warrant a sign.*

These cases require separate status labels in any implementation. Neutrality does not mean opposition to the good. A balanced score can conceal a serious conflict, and ignorance must not be encoded as moral neutrality. The 0/1 code is consequently partial: two fundamental directions do not imply two exhaustive classifications of every event.

Neutrality is relative to the declared domain and time horizon. Choosing between equivalent meals may be irrelevant here, but become relevant when another person's needs, waste, a promise, or a health obligation is involved. Mere preference does not create moral significance.

2.3 Multidimensional character and virtue

Definition 2.4 (Character state). *The state $x_t = (x_{t,1}, \dots, x_{t,d}) \in \mathcal{K} := [-1, 1]^d$ represents selected, relatively persistent dispositions. Each coordinate is oriented so that higher values indicate a stronger disposition toward the corresponding appropriately specified virtue.*

Possible coordinates include honesty, justice, compassion, courage, and temperance. They are not assumed to be independent or exhaustive. *Courage* means appropriately responsive courage, not raw willingness to take risks; reckless and cowardly conduct can both receive negative evaluations. Thus the positive end of a coordinate need not mean maximizing a psychological quantity such as boldness.

The centered interval is convenient for analysis. The equivalent coordinate $z_{t,j} = (x_{t,j} + 1)/2 \in [0, 1]$ has no special metaphysical status. Neither $x_t = 0$ nor $z_t = 1/2$ describes half a good person. Numerical proximity to an endpoint is meaningful only within a fixed, justified calibration.

Scalar summaries discard information. For example, $x = (0.8, -0.8)$ and $x' = (0, 0)$ have the same equally weighted mean, but markedly different profiles. Componentwise dominance $x \succeq x'$, defined by $x_j \geq x'_j$ for all j , avoids some compensation problems while leaving many states incomparable. Neither dominance nor a weighted mean settles conflicts between duties.

2.4 Fallible access to the good

Let Θ index candidate moral interpretations and relevant factual models, and let q_t be a probability distribution over these candidates. The evaluator may use

$$\hat{s}_t(e) = \int_{\Theta} s_{\theta}(e; w) q_t(d\theta), \quad \hat{g}_t(e) = \int_{\Theta} g_{\theta}(e) q_t(d\theta), \quad (5)$$

provided their scales are meaningfully comparable. These are epistemic summaries, not definitions of objective goodness. The true reference need not be in the evaluator's candidate set. High confidence is therefore not a guarantee of correctness.

Before an action, uncertainty also concerns intentions, consequences, and contextual facts. Write $\hat{s}_t(a | c_t)$ for an assessment that averages over those uncertainties as well. This ex ante judgment must not be conflated with a realized event score or a judgment of blameworthiness after an unforeseen outcome.

When cross-theory averages are unjustified, one can instead retain the set of candidate assessments or an interval

$$I_t(a) = [\underline{s}_t(a), \bar{s}_t(a)]. \quad (6)$$

An interval lying strictly above zero warrants a positive sign relative to the covered candidates; one lying strictly below zero warrants a negative sign. An interval containing zero does not warrant either. Such intervals are assumption-dependent uncertainty sets, not automatically calibrated statistical confidence intervals. The difficulties of comparison across moral theories are substantive, as emphasized in the literature on moral uncertainty [6].

3 Model

3.1 Formation is a separate mechanism

The moral quality of an act, the actor's judgment of it, and its psychological effect are different objects. Let

$$h_t = H(e_t, x_t, q_t) \in \mathcal{K}, \quad D_t = \text{diag}(\alpha_{t,1}, \dots, \alpha_{t,d}), \quad 0 \leq \alpha_{t,j} \leq 1, \quad (7)$$

where h_t is a *formation target*. It describes which disposition profile the occasion tends to reinforce. The rates can be decomposed as $\alpha_{t,j} = \eta_{t,j} \rho_{t,j}$, with learning rate $\eta_{t,j}$ and coordinate-specific exposure or salience $\rho_{t,j}$, both in $[0, 1]$.

Axiom 3.1 (Bounded recursive formation). *On the modeled time scale, character evolves according to*

$$\boxed{x_{t+1} = (I - D_t)x_t + D_th_t.} \quad (8)$$

This is a tractable hypothesis about selected dispositions, not an empirically established universal transition law. It allows an isolated occasion to affect one coordinate strongly and another not at all. It also permits interdependence through H , without assuming that all virtues improve together.

For example, a bounded intermediate response $v_t^{\text{learn}} \in \mathcal{K}$ can be mapped into $h_t = Bv_t^{\text{learn}}$, where B is nonnegative and every row sums to at most one. This preserves boundedness and represents specified positive spillovers. Negative couplings require a different map, such as one whose rows have absolute sums at most one. Neither form of coupling is implied by the existence of virtues.

The special case $h_t = g_{\mathcal{G}}(e_t)$ is called the *aligned benchmark*. It makes mathematical consequences transparent but assumes a correspondence that must not be smuggled into the general theory. More generally, a justified mapping L may be needed to relate event scores to disposition targets: $h_t = L(g_{\mathcal{G}}(e_t), e_t)$. Behavior can reinforce a harmful disposition even when it feels rewarding, and a sincerely good intention can coexist with mistaken judgment. Accordingly, neither $h_t = \hat{g}_t$ nor $h_t = g_{\mathcal{G}}$ is imposed in the general model.

An occasion irrelevant to coordinate j has $\alpha_{t,j} = 0$ and leaves that coordinate unchanged. Moral neutrality alone does not force this: a neutral routine could indirectly shape attention or later responses. Conversely, a morally important isolated act could have a small persistent effect. Normative relevance $r_{\mathcal{G}}$, formation salience ρ , and responsibility are not interchangeable.

3.2 Character-dependent choice

Let $u_t(a, c)$ be anticipated nonmoral utility, including convenience or social reward, and let $\phi(a, c) \in \mathbb{R}^d$ describe the relation of an available action to the modeled dispositions. A simple choice rule is

$$\ell_t(a) = \beta x_t^\top \phi(a, c_t) + \kappa \hat{s}_t(a | c_t) + \gamma u_t(a, c_t), \quad (9)$$

$$\pi_t(a | x_t, q_t, c_t) = \frac{\exp(\ell_t(a))}{\sum_{a' \in \mathcal{A}(c_t)} \exp(\ell_t(a'))}, \quad \beta, \kappa, \gamma \geq 0. \quad (10)$$

All logits are assumed finite. The coefficients determine the relative influence of learned disposition, current moral judgment, and other incentives. The sign of an entry of ϕ must be specified; the mere use of softmax does not ensure that a larger x_j favors good conduct. Here β measures sensitivity to dispositions. Learning rates in D_t determine the speed of formation, a separate notion.

With a fixed context, estimates, and incentives, the log odds obey

$$\log \frac{\pi_t(a)}{\pi_t(a')} = \beta x_t^\top (\phi(a, c_t) - \phi(a', c_t)) + \kappa (\hat{s}_t(a) - \hat{s}_t(a')) + \gamma (u_t(a) - u_t(a')). \quad (11)$$

Thus a change in character can change the distribution of future actions without making any one action inevitable. This is a statistical description of choice; randomness is not a solution to the philosophical problem of free will.

The set $\mathcal{A}(c)$ represents practical availability, not moral permission. A separate admissibility set $\mathcal{A}_{\mathcal{G}}(c)$ can encode noncompensable prohibitions. Descriptive choice may put positive

probability on actions outside that normative set. Restricting the model to morally permissible actions by construction would conceal the phenomenon it aims to study.

3.3 A generative cycle and its Markov interpretation

For a closed specification let $S_t = (x_t, q_t, c_t)$. Any changing incentive parameters or relevant memory must also be included in S_t or in its context component. A decision cycle consists of:

1. drawing a_t from the choice distribution $\pi(\cdot \mid S_t)$;
2. drawing or determining (i_t, o_t) conditional on S_t, a_t ;
3. computing h_t, D_t and updating x_t by (8);
4. revising q_t through observation and reflection and advancing the context, including consequences for other people.

The objective score $g_{\mathcal{G}}(e_t)$ can be evaluated conceptually alongside the cycle. It does not need to be available to the actor or observer. Belief revision may use evidence, testimony, or argument; a Bayesian notation does not establish that moral truth has an agreed likelihood function.

Axiom 3.2 (State sufficiency for the Markov version). *Conditional on S_t and the chosen action, the laws of intentions, outcomes, formation, belief revision, and next context do not depend further on the earlier history. Their mechanisms are time invariant unless an intervention or time variable is explicitly included.*

Under this additional assumption, a transition kernel is

$$K(S, B) = \sum_{a \in \mathcal{A}(c)} \pi(a \mid S) \int \mathbf{1}_B((I - D)x + Dh, q', c') P(di, do \mid S, a) R(dq', dc' \mid S, a, i, o), \quad (12)$$

where h, D are the declared measurable functions of the event and state. Standard measurable spaces and kernels are assumed. Equation (12) gives a Markov model on the augmented state; x_t alone need not be Markov. A history-dependent response omitted from the state would invalidate the sufficiency assumption.

An environment can reinforce conduct through changed relationships, opportunities, or rewards. Such feedback enters through R , not only through D_t . Intervention-dependent kernels describe changes in practice, accountability, or opportunity. No Markov assumption implies stationarity, ergodicity, or convergence by itself.

3.4 A one-dimensional analytical reduction

For exact results, fix a context with two actions $a \in \{-1, +1\}$. Assume that their directional assessments and formation targets agree: $g_{\mathcal{G}}(a) = h(a) = a$. With a constant rate $0 < \alpha \leq 1$, set

$$\mathbb{P}(a_t = +1 \mid x_t) = \sigma(\beta x_t + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}}, \quad x_{t+1} = (1 - \alpha)x_t + \alpha a_t. \quad (13)$$

This is a special case of the policy above, obtained with scalar action features $\phi(a) = a/2$ and an appropriate fixed logit difference b . The bias b collects fixed deliberative and contextual

influences. The exact conditional mean is

$$\mathbb{E}[x_{t+1} \mid x_t = x] = F(x) := (1 - \alpha)x + \alpha \tanh\left(\frac{\beta x + b}{2}\right). \quad (14)$$

The deterministic recurrence $y_{t+1} = F(y_t)$ is a *mean-response model*. It is not generally the recursion for $\mathbb{E}[x_t]$, since $\mathbb{E}[F(x_t)] \neq F(\mathbb{E}[x_t])$ for nonlinear F . Its equilibria and basins describe a deterministic approximation; sample paths of (13) remain stochastic.

4 Propositions

The results below are deductions within the stated model. Their proofs do not validate the normative reference, numerical calibration, or psychological mechanism. Superscripts denoting coordinates are avoided; inequalities between vectors are componentwise unless stated otherwise.

4.1 Boundedness, memory, and local improvement

Proposition 4.1 (Invariance and selective irrelevance). *If $x_0 \in \mathcal{K}$, $h_t \in \mathcal{K}$, and $0 \leq \alpha_{t,j} \leq 1$, then $x_t \in \mathcal{K}$ for every t . Moreover, $\alpha_{t,j} = 0$ implies $x_{t+1,j} = x_{t,j}$.*

Proof. Each coordinate of (8) is

$$x_{t+1,j} = (1 - \alpha_{t,j})x_{t,j} + \alpha_{t,j}h_{t,j},$$

a convex combination of two elements of $[-1, 1]$. Induction gives invariance. Setting $\alpha_{t,j} = 0$ gives the second assertion. $\square$

In particular, the model avoids an unbounded cumulative sum of moral scores. A person's represented disposition cannot diverge merely because many decisions have occurred.

Proposition 4.2 (Exact memory weights). *For one coordinate, suppress its index and let $x_{t+1} = (1 - \alpha_t)x_t + \alpha_th_t$. Then*

$$x_t = \left[\prod_{k=0}^{t-1} (1 - \alpha_k) \right] x_0 + \sum_{s=0}^{t-1} \alpha_s \left[\prod_{k=s+1}^{t-1} (1 - \alpha_k) \right] h_s, \quad (15)$$

where empty products equal one. The coefficients are nonnegative and sum to one. If $\alpha_t = \alpha \in (0, 1)$, this becomes

$$x_t = (1 - \alpha)^t x_0 + \alpha \sum_{s=0}^{t-1} (1 - \alpha)^{t-1-s} h_s. \quad (16)$$

Proof. Substitute the recurrence repeatedly, or apply induction in t . If the existing coefficients sum to one, multiplication by $1 - \alpha_t$ and addition of the new coefficient α_t preserve their sum. The constant-rate expression follows directly. $\square$

The formula holds pathwise for the realized sequence of targets and rates, even when that sequence is endogenous. It does not imply that the targets can be chosen independently of prior character. Under a constant rate, the initial state's weight decays exponentially. The half-life of that coefficient, measured in update occasions, is $\log(1/2)/\log(1 - \alpha)$. This is a feature of the assumed filter, not a universal psychological constant.

Proposition 4.3 (A conditional improvement criterion). *If $h_{t,j} \geq x_{t,j}$ for every coordinate, then $x_{t+1} \succeq x_t$. A coordinate increases strictly if both $\alpha_{t,j} > 0$ and $h_{t,j} > x_{t,j}$. In the common-rate case $D_t = \alpha_t I$, a fixed weighted summary $W_t = w^\top x_t$ obeys*

$$W_{t+1} - W_t = \alpha_t (w^\top h_t - W_t). \quad (17)$$

Proof. Subtract $x_{t,j}$ from the coordinate update to obtain $\alpha_{t,j}(h_{t,j} - x_{t,j})$. Multiplication by $w^\top$ gives (17) when rates are common. $\square$

This criterion exposes a limitation of a tempting slogan. A positively evaluated action does not necessarily increase every disposition, nor even its weighted summary. In the aligned scalar benchmark, $x_t = 0.8$, $h_t = g_{\mathcal{G}}(e_t) = 0.2$, and $\alpha_t = 0.1$ give $x_{t+1} = 0.74$. The state tracks a smoothed disposition target rather than accumulated merit. A theory requiring every good action to raise the state would need a different, explicitly defended update law.

4.2 Uncertainty and perturbation

Proposition 4.4 (Robust directional assessment). *Fix $w \in \Delta^{d-1}$. If $\|\hat{g} - g_{\mathcal{G}}\|_\infty \leq \varepsilon$, then*

$$|\hat{s} - s_{\mathcal{G}}| \leq \varepsilon. \quad (18)$$

Consequently, $\hat{s} > \varepsilon$ implies $b_{\mathcal{G}} = 1$, and $\hat{s} < -\varepsilon$ implies $b_{\mathcal{G}} = 0$.

Proof. The triangle inequality yields

$$|w^\top (\hat{g} - g_{\mathcal{G}})| \leq \sum_j w_j |\hat{g}_j - g_{\mathcal{G},j}| \leq \varepsilon \sum_j w_j = \varepsilon.$$

The two strict inequalities therefore preserve the true sign. $\square$

The proposition is useful only if an error bound can be justified. It does not supply one. If the weights are also estimated, with $\hat{w}, w \in \Delta^{d-1}$, the same argument gives

$$|\hat{w}^\top \hat{g} - w^\top g_{\mathcal{G}}| \leq \varepsilon + \|\hat{w} - w\|_1, \quad (19)$$

using $\|g_{\mathcal{G}}\|_\infty \leq 1$. Certainty about coordinates cannot compensate for an undefended aggregation rule.

Proposition 4.5 (Stability under bounded formation error). *Consider two recurrences with the same constant common rate $\alpha \in (0, 1]$, and target sequences satisfying $\|h_t - \tilde{h}_t\|_\infty \leq \delta$. Writing $d_0 = \|x_0 - \tilde{x}_0\|_\infty$, one has*

$$\|x_t - \tilde{x}_t\|_\infty \leq (1 - \alpha)^t d_0 + \delta [1 - (1 - \alpha)^t]. \quad (20)$$

Proof. Subtract the recurrences and use the triangle inequality: $d_{t+1} \leq (1 - \alpha)d_t + \alpha\delta$. Iteration and the finite geometric sum give the bound. $\square$

This result concerns matched target sequences with a known error bound. If a small state error causes the two systems to choose different actions, that bound need not remain small. It is not a general robustness theorem for arbitrary endogenous feedback.

4.3 Stable habits and reinforcing feedback

Proposition 4.6 (Contraction under weak feedback). *For the deterministic map F in (14), let $0 < \alpha \leq 1$, $0 \leq \beta < 2$, and $b \in \mathbb{R}$. Then F has a unique fixed point $y^* \in [-1, 1]$, and every deterministic trajectory converges to it. Specifically,*

$$|y_t - y^*| \leq L^t |y_0 - y^*|, \quad L = 1 - \alpha + \alpha\beta/2 < 1. \quad (21)$$

Proof. The map preserves $[-1, 1]$, and

$$F'(y) = 1 - \alpha + \frac{\alpha\beta}{2} \operatorname{sech}^2\left(\frac{\beta y + b}{2}\right).$$

Thus $0 \leq F'(y) \leq L < 1$. The mean value theorem gives a contraction of the complete interval into itself. The contraction mapping theorem gives existence, uniqueness, and the displayed rate. $\square$

Stability is not a moral endorsement. With $b < 0$, the unique fixed point is negative: the fixed-point equation has a negative root, since $\tanh(b/2) < 0$ at zero and the continuous difference $\tanh((\beta y + b)/2) - y$ is positive at -1 . Convergence can stabilize a harmful pattern.

Proposition 4.7 (Multiple deterministic equilibria). *Let $b = 0$, $0 < \alpha \leq 1$, and $\beta > 2$. Then F has exactly three fixed points, $-y^*, 0, y^*$, with $y^* \in (0, 1)$. Zero is unstable, and the two nonzero fixed points are locally attracting. Every initial $y_0 > 0$ converges to y^* ; every $y_0 < 0$ converges to $-y^*$.*

Proof. Fixed points satisfy $y = \tanh(ky)$, where $k = \beta/2 > 1$. For $f(y) = \tanh(ky) - y$, one has $f(0) = 0$, $f'(0) = k - 1 > 0$, $f(1) < 0$, and $f''(y) < 0$ for $y > 0$. Strict concavity gives exactly one positive root. Odd symmetry gives exactly one negative root. At zero, $F'(0) = 1 - \alpha + \alpha k > 1$. At y^* , strict concavity of $\tanh(ky)$ implies $k \operatorname{sech}^2(ky^*) < \tanh(ky^*)/y^* = 1$. Hence $0 \leq F'(y^*) < 1$, and symmetry handles $-y^*$. For $0 < y < y^*$, $y < F(y) < y^*$; for $y^* < y \leq 1$, $y^* < F(y) < y$. Monotone bounded iteration proves the basin claims. $\square$

The critical value 2 depends on the chosen scaling of the logit. It is not a biological threshold. At $\beta = 2, b = 0$, the derivative at zero equals one, so the strict-contraction proof does not apply. The strong-feedback case illustrates path dependence: small differences in initial state can select different long-run deterministic profiles. It neither establishes moral polarization in a population nor makes these profiles absorbing states of the stochastic process.

Indeed, with finite β, b , (13) assigns positive probability to both actions at every $x \in [-1, 1]$. The stochastic model has no permanent all-positive or all-negative choice forced by these parameters. Statements about its invariant distributions require additional analysis and are not inferred from deterministic equilibria.

4.4 Reversal and the limits of behavioral inference

Proposition 4.8 (Trajectory reversal under sustained practice). *Suppose the scalar state at occasion T is $x_T < 0$, and for the next n updates the target is $h_t = 1$, with constant $0 < \alpha < 1$. Then*

$$x_{T+n} = 1 - (1 - x_T)(1 - \alpha)^n. \quad (22)$$

In particular $x_{T+n} > 0$ whenever

$$n > \frac{\log(1/(1 - x_T))}{\log(1 - \alpha)}. \quad (23)$$

Proof. Subtract the update from one to obtain $1 - x_{t+1} = (1 - \alpha)(1 - x_t)$, and iterate. Solving $(1 - \alpha)^n < 1/(1 - x_T)$ gives the condition, with the inequality reversed when dividing by the negative logarithm. $\square$

For $\alpha = 1$, one such update sets the state to one immediately. The proposition proves a conditional possibility of reversal, not the ability of a person to secure a prescribed sequence. Reflection can change the policy before character changes; altered practices can then change the state. In the deterministic mean-response model, the sufficient condition $b > \beta$ makes the target $\tanh((\beta y + b)/2)$ strictly positive throughout $[-1, 1]$. While $y \leq 0$, each increment is at least $\alpha \tanh((b - \beta)/2) > 0$, so a trajectory crosses zero in finitely many steps. This is a second, policy-mediated illustration, not an estimated intervention threshold.

Proposition 4.9 (Normative non-identification from behavior alone). *Consider two specifications with the same initial state law, policy, formation rule, and environmental kernels, but different objective evaluations g_G and $g_{G'}$. If those evaluations do not enter the shared behavioral mechanisms, the two specifications induce the same distribution of observed action histories.*

Proof. The conditional distribution of the next observed action is the same under either specification, as is the law of the next latent state. Starting with the shared initial law, induction yields equality of every finite joint action distribution. $\square$

The condition holds, for example, when the actor follows the same subjective judgments in both specifications and formation depends on the same experienced events. Better behavioral prediction cannot by itself select the true moral reference. This is a limited identification result, not an argument that moral knowledge is impossible or that normative reasons cannot use empirical facts.

5 Illustrative examples

All values in this section are constructed. They are neither ratings of actual people nor estimates from a dataset. Their role is to expose the implications and limits of the equations.

5.1 An ordinary decision changes with context

Consider getting out of bed. If the actor is rested and has a reasonable commitment to someone relying on them, rising can express responsibility. If the actor is ill and needs recovery, resting may be the responsible choice. The observable alternative alone does not determine its score. Intention matters too: resting to recover and invoking recovery as a pretext for knowingly abandoning an obligation need not be the same event. An outside observer may be unable to distinguish them.

For a toy two-coordinate system, let $x = (0.2, 0.4)$ represent reliability and care, let $D = \text{diag}(0.1, 0.05)$, and suppose a specified event has formation target $h = (0.8, 0.6)$. Then

$$x' = (0.9 \cdot 0.2 + 0.1 \cdot 0.8, 0.95 \cdot 0.4 + 0.05 \cdot 0.6) = (0.26, 0.41).$$

If an equivalent meal choice is irrelevant to both modeled dispositions, setting $D = 0$ gives $x' = x$. The latter conclusion concerns this occasion and model; it is not a universal claim that food choices have no moral consequences.

5.2 Conflict and uncertain classification

Suppose an event has assessed profile $g = (0.6, -0.6)$. With $w = (0.5, 0.5)$, its scalar score is zero, although neither coordinate is neutral. With $w' = (0.75, 0.25)$, the score is 0.3. The difference records a substantive weighting choice; arithmetic cannot resolve the conflict. A prohibition, if applicable, would need to be represented separately rather than outweighed by unrelated gains.

Suppose instead that $\hat{g} = (0.35, 0.15)$, $w = (0.5, 0.5)$, and a justified bound is $\varepsilon = 0.1$. Then $\hat{s} = 0.25$ and the true scalar score lies in $[0.15, 0.35]$; the sign is robust under Proposition 4.4. For $\hat{g} = (0.08, 0.02)$ with the same bound, the interval is $[-0.05, 0.15]$. A positive point estimate of 0.05 does not warrant a definite positive code. The model preserves uncertainty rather than converting hesitation into a negative judgment.

5.3 Repeated formation and a change of course

For a constant target h , constant rate α , and scalar initial state x_0 , Proposition 4.2 gives

$$x_t = h + (x_0 - h)(1 - \alpha)^t. \quad (24)$$

Figure 1 compares sustained positive and negative targets with a change of target after 20 updates. The model approaches the target asymptotically; it does not add an unlimited number of moral points. A course correction changes subsequent formation without rewriting the earlier trajectory.

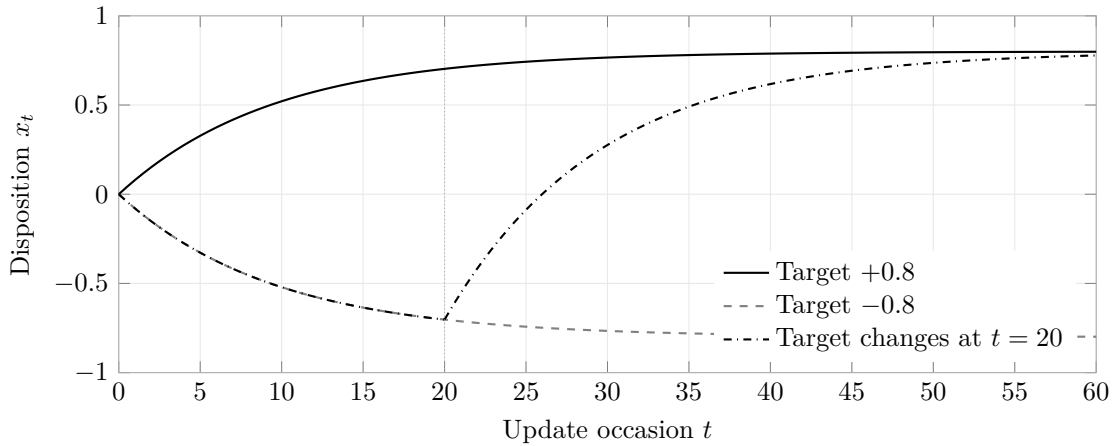


Figure 1: Constructed trajectories with $x_0 = 0$ and $\alpha = 0.1$. The changing-target trajectory uses $h = -0.8$ for updates $t = 0, \dots, 19$ and $h = +0.8$ thereafter. Lines interpolate exact values at integer occasions. These are prescribed-target calculations, not population data.

As a separate calculation, start from $x_T = -0.8$ and impose target $h = 1$, with $\alpha = 0.1$. Equation (22) gives Table 1; the first positive state occurs at the sixth update. Six updates are a consequence of these arbitrary parameters, not a claim about a six-day conversion process.

5.4 A feedback loop can favor either direction

Set $b = 0$ in the mean-response model. At $\beta = 1$, there is a unique fixed point $y = 0$. At $\beta = 3$, the fixed points are approximately $-0.8586, 0, 0.8586$, with the outer two attracting. Figure 2 plots the fixed-point equation. The transition from one to three intersections illustrates the

Table 1: A prescribed reversal, $x_{T+n} = 1 - 1.8(0.9)^n$. Displayed values are rounded to four decimal places.

n	0	1	5	6	10
x_{T+n}	-0.8000	-0.6200	-0.0629	0.0434	0.3724

role of feedback, while the learning rate α controls the speed of deterministic adjustment without changing these fixed-point locations.

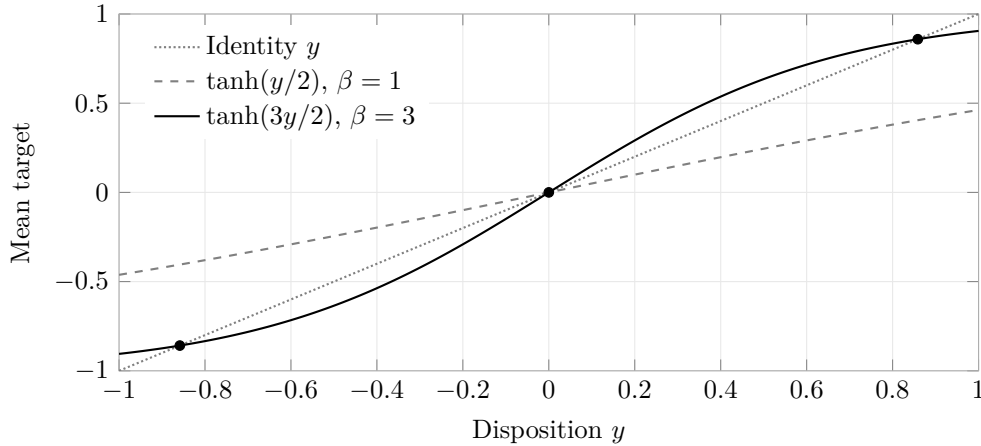


Figure 2: Fixed points are intersections of the mean target with the identity line. For $\beta = 3$, the central fixed point is unstable and the outer fixed points are locally attracting in the deterministic recurrence. The diagram does not display a stochastic stationary distribution or empirical moral categories.

Even at the positive fixed point of the $\beta = 3$ example, the stochastic choice probability is $\sigma(3y^*) = (1 + y^*)/2 \approx 0.9293$, not one. Reinforcement can make a response more likely without making it certain. Moreover, a more stable pattern is not necessarily better: the negative attracting point is equally stable under this symmetric specification.

6 Philosophical interpretation

6.1 Aristotle: formation through practice

Aristotle's account of ethical virtue connects repeated activity to acquired dispositions. Book II of the *Nicomachean Ethics* also distinguishes merely performing an outwardly appropriate act from acting knowingly, choosing it for its own sake, and acting from an established disposition [1, II.1–4]. This supports the conceptual separation between event, intention, and state. It does not license the identification of virtue with an exponential moving average. Aristotle's emphasis on practical judgment and the mean also cautions against treating virtue as maximization of a crude behavioral quantity. The present coordinates represent appropriately calibrated virtues; they do not replace practical wisdom with a scale. His warning about the limited precision available in ethics is especially relevant to the stronger numerical assumptions made here.

6.2 Augustine: direction without ontological dualism

In the *Enchiridion*, Augustine describes evil as a privation or corruption of good rather than an independently existing substance [2, chaps. 11–14]. A signed coordinate is compatible with that asymmetry if negative values denote departures from due goodness, not quantities of a second metaphysical substance. Likewise, the symbols 0 and 1 can encode opposing directions without positing two equal ultimate principles. An affine transformation between $[-1, 1]$ and $[0, 1]$ changes a coordinate convention; it neither proves nor refutes privation theory. The mathematical symmetry of the toy feedback example is an analytical simplification, not a theological claim about the status of good and evil.

6.3 Jesus: fruits, light, and the limits of the analogy

In Matthew 7:15–20, Jesus’ image of trees and their fruits concerns the discernment of false prophets [9]. It supplies an analogy between enduring orientation and outward conduct, but it is not a statistical rule for diagnosing every person’s character from a few observations. The light and darkness imagery in John 3:19–21 connects conduct to openness or resistance to truth [8]. It resonates with directional language while retaining a theological meaning beyond this model. Neither passage establishes a probability distribution, a learning rate, or a moral scoring scheme.

6.4 Stoicism: agency and what is not controlled

Epictetus distinguishes what depends on us from what does not, with particular attention to judgment and the use of impressions [4, secs. 1, 5, 20]. This motivates separating deliberation and intention from uncontrolled outcomes. It also resists a simple equation of fortunate consequences with virtuous agency. The model does not adopt the whole Stoic account of virtue or claim that the different philosophical traditions share one ethical theory. Its event structure merely leaves room for their disagreements to be stated.

6.5 Psychology and reinforcement learning

Lally and colleagues modeled increasing automaticity under repetition in recurring contexts; their study is evidence about particular habits, not the moral quality of those habits [5]. Wood and R  nger review relationships between habitual responses, contexts, and deliberate goals [10]. These works motivate investigating formation and persistence, but do not establish the present state variables or universal rates.

Reinforcement learning offers formal tools for state, action, feedback, and policy [7]. Computational accounts of competing habitual and goal-directed control provide a related comparison [3]. The recurrence here is a bounded filter, not by itself a complete reinforcement-learning algorithm: it specifies neither a Bellman objective nor a validated reward-learning mechanism. Most importantly, experienced reward and objective moral goodness are different quantities. An agent can learn efficiently to pursue a bad objective. The distinction between u_t , $\hat{s}_t$, and g_t preserves that possibility.

7 Objections and limitations

7.1 Is the binary language either false or trivial?

If the thesis were that every human choice is unambiguously good or evil, neutral events, conflicting considerations, and ignorance would refute its scope. The formal thesis is narrower: where a justified signed assessment exists and is nonzero, its direction can be coded with one bit. That coding is elementary mathematics. The explanatory work lies in the proposed formation and feedback mechanisms, not in the sign function itself. A directional vocabulary can organize reflection without constituting an independently discovered law of nature.

7.2 Does objectivity justify numbers or moral aggregation?

No. Axiom 2.2 goes beyond Axiom 2.1. Interpersonal comparability, common units across virtues, and exchange rates between moral considerations are not established. Some duties may be lexically ordered or noncompensable. The admissibility set $\mathcal{A}_G(c)$ can represent prohibitions, and a partial order can preserve incomparability, but neither solves all ethical conflicts. Results that use $w^\top g$ are conditional on that scalarization being appropriate. A successful fit of behavior does not justify the chosen weights.

7.3 Does the model explain formation, or only rename it?

An unconstrained H could absorb almost any observed change. The general framework is therefore a specification language, not yet a falsifiable psychological theory. The aligned benchmark, constant-rate filter, and two-action model are narrower hypotheses with explicit implications. An empirical application would have to fix the form of H , select measures, constrain parameters, and test predictions against alternatives before seeing the outcomes. Merely reporting that the model can be adjusted to describe a life would not validate it.

7.4 Why should character have short, smooth memory?

Equation (8) represents incremental adjustment and may miss abrupt insights, enduring commitments, trauma, forgetting, developmental discontinuities, and identity-level change. Its constant-rate version discounts old experience geometrically. Other mechanisms could require multiple timescales, nonlocal memory, or state-dependent rates. A sudden policy change can be represented by changing b or q , but that representation does not explain how such a change arises. Repentance as a religious or personal phenomenon is richer than crossing a coordinate's zero point.

7.5 Can conduct reveal an underlying virtuous person?

Observed conduct is jointly produced by disposition, incentives, opportunity, and constraint. Different combinations can generate the same actions. Even behavioral parameters can be non-identifiable: without a scale convention, changes in the size of the features ϕ can be offset by changes in β . Latent intentions add further ambiguity. Proposition 4.9 shows a separate obstacle to recovering objective moral assessment from behavior alone. The model therefore provides no warrant for assigning people total moral scores or immutable types.

7.6 Does the model individualize structural problems?

A poor choice set can reflect deprivation, coercion, harmful social norms, or institutions. Representing these factors in c_t is necessary but does not make them disappear. Two people with the same disposition can face very different available actions and costs. An account of improvement that changes only x_t while ignoring $\mathcal{A}(c_t)$, incentives, and relationships would be incomplete. Consequences for others remain ethically relevant independently of whether an action improves the actor's future character.

7.7 Do the equations settle freedom, responsibility, or salvation?

They do not. A stochastic choice rule is compatible with several accounts of agency; it does not distinguish them. The framework has no variable that measures human dignity, ultimate responsibility, grace, or salvation. A negative state is a revisable representation of selected dispositions, never an assertion that a person lacks worth. Nor does modeled improvement cancel past harm, restitution, or obligations to those affected.

8 Discussion

The framework separates three possible uses. As a conceptual tool, it disciplines claims about direction, habituation, and change. As a mathematical object, it supports exact conditional reasoning. As an empirical model, it remains a proposal requiring operational definitions and independent validation.

A credible empirical program would begin with a narrow domain, such as repeated opportunities to disclose an error, and a justified measurement model. It would distinguish observed conduct from reported intention, perceived moral judgment, and situational pressure. Parameters estimated from one set of occasions should predict later choices in held-out contexts. A no-learning baseline, a context-only model, and alternative memory kernels would test whether the character state adds explanatory value.

Several predictions are conditional and potentially informative. Fixed-target learning predicts diminishing increments through (24). At constant context and incentives, (11) predicts how a disposition change alters log odds. The scalar feedback reduction predicts qualitatively different deterministic behavior across specified parameter regimes. Failure of these predictions would challenge the corresponding model, not moral realism or the existence of character.

For personal reflection, the model directs attention to four distinct questions: what matters morally in this situation; how certain the assessment is; which response a practice tends to strengthen; and how the next situation's opportunities or incentives could change. Answering those questions need not involve assigning numbers. A useful representation may remain qualitative when numerical precision would be spurious.

No human-subject data are presented, and no parameters are estimated. The numerical illustrations are reproducible from the displayed equations and the self-contained source. They demonstrate arithmetic and dynamical implications only.

9 Conclusion

The defensible content of a 0/1 theory is a distinction between opposed moral directions combined with a nonbinary account of intensity, character, and uncertainty. Under an assumed objective reference, some events admit a positive or negative orientation; neutrality, conflict, and unresolved assessment remain explicit. Character is modeled as a vector of dispositions that can affect future choice and be modified by subsequent experience.

The recursive model proves that bounded formation can retain a weighted memory of action, that feedback can stabilize either direction, and that a change of trajectory is possible under stated conditions. It also shows why repetition, confidence, and stability cannot be equated with goodness. Its philosophical premise is not proved by its mathematics, and its psychological assumptions require testing. The result is a working framework for asking how choices shape the person who chooses next, with its scope and uncertainties kept visible.

A Notation and reproducibility

Symbol	Meaning
$\mathcal{G}$	Assumed objective moral reference
$e = (c, a, i, o)$	Context, action, intention/knowledge, and outcome
$g_{\mathcal{G}}(e)$	Moral profile on the declared coordinate scales
$s_{\mathcal{G}}, b_{\mathcal{G}}, m_{\mathcal{G}}$	Scalar assessment, directional bit, and magnitude
w	Declared weights for a scalar projection
$x_t \in [-1, 1]^d$	Selected disposition state
$q_t, \hat{s}_t$	Epistemic distribution and estimated assessment
h_t, D_t	Formation target and coordinate update rates
ϕ, u_t	Disposition features and anticipated nonmoral utility
π_t	Conditional action distribution
β, κ, γ	Policy sensitivities to disposition, judgment, utility
S_t, K	Augmented state and its transition kernel
α, b, F	Common update rate, scalar policy bias, mean-response map

Table 2: The normative, epistemic, and psychological quantities have different roles even when a benchmark assigns them the same numbers.

Figure 1 is generated directly by (24). For its changing-target path, compute $x_{20} = -0.8[1 - (0.9)^{20}]$ and then use $x_t = 0.8 + (x_{20} - 0.8)(0.9)^{t-20}$ for $t \geq 20$. Figure 2 plots y , $\tanh(y/2)$, and $\tanh(3y/2)$; bisection of $\tanh(1.5y) - y = 0$ on $[0.5, 1]$ gives the positive nonzero intersection $0.8585596366\dots$. Table 1 uses (22) without simulation. No random seed, external dataset, or fitted parameter is involved.

The source contains its references and vector figures and requires no external bibliography or image files. Standard LaTeX with the listed packages suffices; repeated compilation resolves cross-references.

References

- [1] Aristotle. *Nicomachean Ethics*, Book II, especially chapters 1–6. Translated by W. D. Ross. Internet Classics Archive, MIT. <https://classics.mit.edu/Aristotle/nicomachaen.2.ii.html>.
- [2] Augustine of Hippo. *Enchiridion on Faith, Hope, and Love*, chapters 11–14. Translated by J. F. Shaw, in *Nicene and Post-Nicene Fathers*, First Series, vol. 3, edited by P. Schaff (1887). New Advent edition. <https://www.newadvent.org/fathers/1302.htm>.
- [3] N. D. Daw, Y. Niv, and P. Dayan (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8, 1704–1711. doi:10.1038/nn1560.
- [4] Epictetus. *Enchiridion*, especially sections 1, 5, and 20. Translated by E. Carter. Internet Classics Archive, MIT. <https://classics.mit.edu/Epictetus/epicench.html>.
- [5] P. Lally, C. H. M. van Jaarsveld, H. W. W. Potts, and J. Wardle (2010). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology*, 40(6), 998–1009. doi:10.1002/ejsp.674.
- [6] W. MacAskill, K. Bykvist, and T. Ord (2020). *Moral Uncertainty*. Oxford University Press. doi:10.1093/oso/9780198722274.001.0001.
- [7] R. S. Sutton and A. G. Barto (2018). *Reinforcement Learning: An Introduction*, 2nd edition. MIT Press. <https://mitpress.mit.edu/9780262039246/reinforcement-learning/>.
- [8] *New American Bible, Revised Edition* (2011). John 3:19–21. United States Conference of Catholic Bishops. <https://bible.usccb.org/bible/john/3>.
- [9] *New American Bible, Revised Edition* (2011). Matthew 7:15–20. United States Conference of Catholic Bishops. <https://bible.usccb.org/bible/matthew/7>.
- [10] W. Wood and D. R  nger (2016). Psychology of habit. *Annual Review of Psychology*, 67, 289–314. doi:10.1146/annurev-psych-122414-033417.