What Affect Permits

A Threshold Theory of Expression

Flyxion
Independent Researcher

August 2026

Abstract

Affective states are commonly interpreted through their motivational consequences: happiness encourages affiliation, anger encourages confrontation, fear encourages withdrawal, and so forth. Yet changes in affect often produce behavioral combinations that are difficult to explain as movement along a single motivational axis. A person in an elevated positive state may become simultaneously more affectionate and more critical, more generous and more tactless, more playful and more willing to disagree. This essay develops a threshold theory of expression in which affect can change behavior through two causally distinct pathways. It may alter the activation of a disposition, but it may also alter the threshold that an already-active disposition must cross before becoming externally expressed.

For behavior b_i, let $a_i(t)$ denote its current activation and let $\theta_i(t, j)$ denote its context- and audience-dependent expression threshold. Define the latent expression margin

$$z_i(t, j) = a_i(t) - \theta_i(t, j),$$

with expression probability

$$\mathbb{P}(E_i = 1 \mid t, j) = \sigma(z_i(t, j)) \,,$$

where σ is any monotonically increasing response function. Activation is further decomposed as

$$a_i(t) = D_i + \eta_i(S_t) + \varepsilon_i(t),$$

where D_i represents a relatively persistent disposition, $\eta_i(S_t)$ a state-dependent motivational contribution, and $\varepsilon_i(t)$ residual fluctuation. The threshold is represented as a function

$$\theta_i(t, j) = \Theta_i \left(C_t, j, \mathcal{H}_{ij,t}, E_t, U_t, N_t, R_t, P_t, \ldots\right),$$

incorporating anticipated social cost, interaction history, executive control, uncertainty, norm enforcement, reversibility, power asymmetry, and related constraints.

This decomposition makes several claims available that a purely motivational account cannot make. First, a common permissive shift can simultaneously increase expression of behaviors with opposing social valences without requiring a common motivational cause. Second, interventions as different as positive affect, anonymity, exhaustion, intoxication, intimacy, institutional permission, and perceived safety may converge behaviorally while acting on different components of the expression system. Third, where threshold relaxation exposes heterogeneous latent dispositions, it predicts trait-conditioned divergence rather than uniform behavioral movement. Fourth, visible disagreement can increase as relationships become safer because the cost of dissent has fallen, making expressed conflict an unreliable monotonic measure of underlying hostility.

The model also yields an identifiability problem with consequences for ordinary claims about authenticity. Observable expression depends on the difference between activation and

inhibition, not on disposition alone. The same behavior can therefore be generated by multiple combinations of trait, state, threshold, and context. Disinhibited behavior is evidence about a person, but it is not a privileged measurement of an otherwise concealed "true self."

Finally, the framework is connected to the companion account developed in *The North Setpoint*. That essay concerns the conditions under which a discrepancy becomes salient enough to register as grievance; the present essay concerns the downstream question of whether an available grievance becomes expressed. Complaint rates consequently combine at least two distinct probabilities,

$$\mathbb{P}(\text{complaint}) = \mathbb{P}(\text{grievance})\mathbb{P}(\text{expression} \mid \text{grievance}),$$

allowing improving conditions and increasing visible complaint to coexist without contradiction. The broader thesis is that permission is itself causal. To explain an observed action, it is not enough to ask what motivated it or what it reveals about the actor. One must also ask what had to become permissible before the disposition could become visible.

1 Kinder and More Critical at the Same Time

Elevated positive affect can produce a pattern that conventional accounts of valence make look contradictory. Someone becomes more affectionate, more candid, more playful, and more generous, while in the same stretch of conversation becoming more willing to criticize, to disagree, to confess an irritation, or to violate a minor social convention that would ordinarily have been observed. The person appears simultaneously nicer and less tactful.

There is no necessary contradiction in this pattern. The contradiction arises only if affect is assumed to provide a behavioral instruction whose semantic content should be reflected uniformly in everything that follows from it. On that interpretation, positive affect should produce positively valenced behavior, negative affect negatively valenced behavior, and the appearance of both at once demands either multiple competing causes or an exception to the rule.

A different possibility is that the common cause occurs one level earlier.

Suppose affection and criticism are already independently available behavioral possibilities. Positive affect need not create either one. Instead, it may alter the conditions under which both are allowed into expression. What changes is not necessarily the content waiting behind the gate, but the gate itself.

Let b_i denote a potentially expressible behavior and E_i the event that it is actually expressed. Let a_i represent the current activation of that behavior and θ_i the threshold it must exceed. In the simplest deterministic case,

$$E_i = \mathbf{1}[a_i > \theta_i].$$

For two behaviors—say,

$$b_1 = \text{affection}, \qquad b_2 = \text{criticism},$$

their simultaneous expression is governed by

$$\mathbb{P}(E_1 = 1, E_2 = 1) = \mathbb{P}(a_1 > \theta_1,\ a_2 > \theta_2).$$

Nothing in this expression requires a_1 and a_2 to share a motivational source. Indeed, consider the deliberately restrictive special case

$$a_1 \perp a_2.$$

Then

$$\mathbb{P}(E_1 = 1, E_2 = 1) = \mathbb{P}(a_1 > \theta_1)\mathbb{P}(a_2 > \theta_2).$$

Now suppose some affective state H lowers both relevant thresholds:

$$\frac{d\theta_1}{dH} < 0, \qquad \frac{d\theta_2}{dH} < 0.$$

Provided there is probability mass near both thresholds, it follows that

$$\frac{d}{dH}\mathbb{P}(E_1 = 1) > 0$$

and

$$\frac{d}{dH}\mathbb{P}(E_2 = 1) > 0.$$

Affection and criticism therefore become more likely to appear together even under the assumption that their activations are statistically independent. Their joint increase does not require

$$\mathrm{Cov}(a_1, a_2) \neq 0.$$

This special case is useful precisely because independence is stronger than the argument requires. Real dispositions will often be correlated. Affection and candor may share causes; criticism and irritation may share others. The point is not that behavioral activations are actually independent, but that a common motivational cause is unnecessary to explain their simultaneous release. A shared change in permissibility is sufficient.

The deterministic threshold is also an idealization. Human expression is noisy. A behavior just below threshold may occasionally appear, while one just above threshold may still fail to do so. It is therefore useful to define an expression margin

$$z_i = a_i - \theta_i$$

and let expression probability be a monotonically increasing function of that margin:

$$p_i = \mathbb{P}(E_i = 1) = \sigma(z_i) = \sigma(a_i - \theta_i),$$

where σ may be logistic or any other suitable monotone response function.

The important quantity is then not activation alone but the distance between activation and inhibition. Increasing a_i and decreasing θ_i both increase z_i:

$$\Delta z_i = \Delta a_i - \Delta \theta_i.$$

Observed expression cannot immediately tell us which term moved.

This ambiguity is the first clue that a motivational theory is incomplete. Suppose a normally reserved person becomes unusually affectionate during an elevated state. One explanation is

$$\Delta a_{\text{affection}} > 0 :$$

the person became more strongly motivated to express affection. Another is

$$\Delta \theta_{\text{affection}} < 0 :$$

the affection was already present at approximately the same activation, but became cheaper, safer, or easier to express. Both mechanisms may of course operate simultaneously.

The distinction becomes more revealing when the behavior is criticism. If the same person becomes more affectionate and more critical at once, a purely valence-directed motivational account must explain why a supposedly positive state is generating tendencies with apparently opposing social signs. The threshold account requires no such semantic reversal. It permits

$$\Delta a_{\text{affection}} > 0, \qquad \Delta a_{\text{criticism}} \approx 0,$$

while

$$\Delta\theta_{\text{affection}} < 0, \qquad \Delta\theta_{\text{criticism}} < 0.$$

Affection may therefore increase partly because the person feels more affectionate, while criticism increases because criticism that was already present has become easier to voice. Similar observable changes need not have identical causal decompositions.

This matters because many ordinary descriptions of affect quietly identify expression with motivation. If a behavior becomes more frequent, we infer that the person must want it more. If it disappears, we infer that the underlying disposition has weakened. But expression is already a selected output. It is what remains after motivation has interacted with anticipated consequences, social norms, inhibition, uncertainty, audience, self-monitoring, and prior experience.

Affective state can therefore change the visible person without changing every disposition that becomes newly visible.

This suggests the distinction on which the rest of the essay depends:

> Affect may change behavior not only by changing what a person wants to do, but by changing what an already-present disposition is permitted to express.

The difference is subtle at the level of observation because both pathways can produce the same outward act. At the level of explanation, however, they are different causal claims. One changes the force pushing toward expression. The other changes the resistance that force must overcome.

The apparent contradiction with which we began is therefore better understood as a problem of causal underdetermination. When a person becomes kinder and more critical at the same time, the observable behaviors reveal that their respective expression boundaries have been crossed, but not why. Greater activation, reduced inhibition, or some combination of the two can produce the same outward result. Expression alone cannot identify the mechanism that made expression possible.

The distinction therefore requires two independently variable quantities where a purely motivational account effectively provides only one: the strength of the disposition pressing toward expression, and the resistance standing between that disposition and its appearance in behavior.

2 What Affect Wants vs. What Affect Permits

The distinction suggested by the opening example can now be stated more precisely. A conventional motivational account treats affect primarily as a cause of changes in behavioral activation. If S_t denotes the person's state at time t, and a_i the activation of disposition i, the basic picture is

$$S_t \longrightarrow a_i.$$

A change in state changes the strength of a behavioral tendency. Increased affectionate motivation produces more affectionate behavior; increased irritation produces more criticism; increased fear produces more avoidance. There is nothing wrong with this account as far as it goes. The difficulty is that it places almost all explanatory weight on the force pushing toward an action and comparatively little on the resistance standing between an activated disposition and its expression.

Introduce therefore a second quantity, the expression threshold θ_i. In the simplest deterministic formulation,

$$b_i \text{ expressed } \iff a_i(S_t) > \theta_i(S_t, C_t),$$

where C_t denotes the relevant context. The observable behavior depends not on activation alone but on the relation between activation and threshold.

It is useful to collect that relation into a single latent expression margin,

$$z_i(t) = a_i(t) - \theta_i(t).$$

When z_i is strongly negative, expression is unlikely. When it is strongly positive, expression is likely. The deterministic boundary

$$z_i > 0$$

is conceptually convenient, but human behavior is not sufficiently precise to justify treating the boundary as perfectly sharp. A probabilistic formulation is therefore more appropriate:

$$p_i(t) = \mathbb{P}(E_i = 1 \mid t) = \sigma(z_i(t)) = \sigma(a_i(t) - \theta_i(t)),$$

where $E_i = 1$ denotes expression and σ is a monotonically increasing response function. A logistic function,

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

is one possible choice, though nothing in the argument depends specifically on logistic form. What matters is simply

$$\sigma'(z) > 0.$$

This formulation immediately reveals why activation and permission must be distinguished. An increase in expression probability can result from

$$\Delta a_i > 0,$$

from

$$\Delta \theta_i < 0,$$

or from both occurring simultaneously. In terms of the expression margin,

$$\Delta z_i = \Delta a_i - \Delta \theta_i.$$

The same observed increase in z_i is therefore compatible with multiple causal histories.

Suppose now that H denotes some dimension of affective state. Differentiating expression probability with respect to H gives

$$\frac{dp_i}{dH} = \sigma'(a_i - \theta_i)\left(\frac{da_i}{dH} - \frac{d\theta_i}{dH}\right).$$

The two terms inside the parentheses correspond to two different causal pathways. The first,

$$\frac{da_i}{dH},$$

is the motivational pathway: affect changes the activation of the disposition. The second,

$$-\frac{d\theta_i}{dH},$$

is the permissive pathway: affect changes how much activation is required before the disposition becomes expressible.

Writing the decomposition explicitly,

$$\frac{dp_i}{dH} = \underbrace{\sigma'(z_i)\frac{da_i}{dH}}_{\text{motivational contribution}} + \underbrace{\sigma'(z_i)\left(-\frac{d\theta_i}{dH}\right)}_{\text{permissive contribution}},$$

makes clear that these are not two descriptions of the same mechanism. They are separately variable causal contributions to the same observable probability.

This distinction also gives the theory a limiting case that is especially important. There may be behaviors for which affect changes expression while having little effect on activation itself:

$$\frac{da_i}{dH} \approx 0, \qquad \frac{d\theta_i}{dH} < 0.$$

Then

$$\frac{dp_i}{dH} > 0$$

despite approximately no increase in the underlying behavioral activation. Something the person already wanted to say, do, reveal, criticize, admit, or offer becomes more likely to appear because the barrier to its expression has changed.

Conversely, affect can increase motivation while leaving permission largely unchanged:

$$\frac{da_i}{dH} > 0, \qquad \frac{d\theta_i}{dH} \approx 0.$$

And in ordinary cases the two pathways may reinforce one another:

$$\frac{da_i}{dH} > 0, \qquad \frac{d\theta_i}{dH} < 0.$$

The usefulness of the distinction does not depend on claiming that one pathway is universally dominant. It depends only on their being causally separable.

There is also an important asymmetry between them. Motivation contains information about behavioral direction. A change in a_i concerns a particular disposition i. A sufficiently general change in permissibility, by contrast, need not specify what comes through the gate. If several dispositions are already active near their respective thresholds, a common reduction in inhibition can expose several of them at once:

$$\theta_i' = \theta_i - \delta, \qquad \delta > 0,$$

for a collection of behaviors $i \in I$. The newly expressible set is then

$$\mathcal{N} = \{i \in I : \theta_i - \delta < a_i \leq \theta_i\}.$$

Nothing requires the members of $\mathcal{N}$ to share a social valence. Affection, criticism, humor, disclosure, dissent, generosity, and tactlessness can all occupy the narrow band uncovered by the same permissive shift.

This is why the distinction between what affect wants and what affect permits is more than terminology. A motivational explanation attributes the direction of behavioral change to the state itself. A permissive explanation allows the direction to come from the pre-existing distribution of dispositions while the state changes their probability of becoming observable.

The central thesis can therefore be written compactly as

state modulates permission; disposition supplies direction.

The statement is deliberately not absolute. Affect can supply direction as well. Positive affect may genuinely increase affiliative motivation; anger may genuinely increase confrontational motivation; fear may genuinely increase avoidance. The claim is instead that behavioral expression contains a second degree of freedom that disappears when motivation and permission are treated as the same thing.

Indeed, even the compact quantity $z_i = a_i - \theta_i$ should not tempt us to collapse them again. Although z_i is sufficient for predicting expression in the minimal model, it is insufficient for explaining it. The transformations

$$a_i \mapsto a_i + \delta$$

and

$$\theta_i \mapsto \theta_i - \delta$$

produce the same change,

$$z_i \mapsto z_i + \delta,$$

while making different claims about what happened inside the person. Prediction can therefore succeed while causal interpretation fails.

This distinction will become increasingly important as the argument proceeds. The behavior visible from outside is not the full behavioral field. It is the portion of that field surviving a selection process. Before asking what a change in expression tells us about motivation, character, or affect, we must first ask what remained present but invisible because it never crossed the boundary into expression.

That hidden population is the subject of the next section.

3 The Hidden Population of Unexpressed Behavior

Once activation and permission are separated, observed behavior can no longer be treated as a transparent sample of disposition. What appears externally is a selected subset of a much larger behavioral field.

Let

$$\mathcal{B}_t = \{b_1, b_2, \ldots, b_n\}$$

denote the set of behavioral dispositions available to a person at time t. Availability here does not mean that every member of $\mathcal{B}_t$ is equally active, equally stable, or equally characteristic. It means only that each is a behavior the system is presently capable of producing under some reachable combination of activation and threshold.

The externally visible subset is

$$\mathcal{E}_t = \{b_i \in \mathcal{B}_t : a_i(t) > \theta_i(t)\}\,.$$

In the probabilistic formulation of the previous section,

$$\mathbb{P}(E_i = 1) = \sigma(a_i - \theta_i)\,,$$

so the boundary between visible and invisible behavior is softened without changing the underlying distinction. Some behaviors are highly likely to appear, some highly unlikely, and a potentially important population lies near the boundary.

This means that silence is not evidence of absence. A person who rarely criticizes may possess little critical disposition, or may possess a great deal of it under a persistently high expression threshold. A person who rarely expresses affection may be relatively unaffectionate, or deeply affectionate and highly reserved. The same ambiguity applies to dissent, confession, eccentricity, aggression, humor, vulnerability, generosity, curiosity, and nearly every other behavior for which expression carries a cost.

The inference problem can be stated formally. Define the binary observation

$$Y_i(t) = \mathbf{1}[a_i(t) > \theta_i(t)].$$

From $Y_i = 0$, one can infer only

$$a_i \leq \theta_i.$$

One cannot infer

$$a_i \approx 0.$$

Likewise, from $Y_i = 1$, one learns that activation exceeded threshold, not that activation was intrinsically large:

$$a_i > \theta_i$$

is compatible both with a large a_i and a large θ_i, and with a small a_i and an even smaller θ_i.

The same observed behavior can therefore arise from very different internal configurations. For example,

$$(a_i, \theta_i) = (10, 9)$$

and

$$(a_i, \theta_i) = (2, 1)$$

produce the same deterministic observation,

$$Y_i = 1,$$

despite representing quite different relations between disposition and constraint. Conversely,

$$(a_i, \theta_i) = (8, 9)$$

and

$$(a_i, \theta_i) = (1, 7)$$

both produce

$$Y_i = 0,$$

although the first contains a strongly activated behavior narrowly suppressed and the second a weakly activated behavior nowhere near expression.

The observable state is consequently many-to-one with respect to the latent state. If

$$\Phi(a_i, \theta_i) = \mathbf{1}[a_i > \theta_i],$$

then in general

$$\Phi^{-1}(Y_i)$$

contains infinitely many possible pairs (a_i, θ_i). Behavioral observation is therefore an inverse problem with no unique solution unless additional information about activation or threshold is independently available.

At the level of the entire behavioral field, define the visible fraction

$$V_t = \frac{|\mathcal{E}_t|}{|\mathcal{B}_t|} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[a_i(t) > \theta_i(t)].$$

If activations and thresholds are treated as random variables with joint density $f_{a,\theta}(a, \theta)$, then for a sufficiently large behavioral population,

$$V_t \longrightarrow \mathbb{P}(a > \theta),$$

with

$$\mathbb{P}(a > \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{a} f_{a,\theta}(a, \theta)\, d\theta\, da.$$

Under the simplifying assumption that activation and threshold are independent,

$$f_{a,\theta}(a, \theta) = f_a(a) f_\theta(\theta),$$

this becomes

$$\mathbb{P}(a > \theta) = \int_{-\infty}^{\infty} f_a(a) F_\theta(a)\, da,$$

where F_θ is the cumulative distribution function of thresholds.

The scalar V_t therefore compresses two distributions into a single observable quantity. A population with high average activation and high thresholds can have the same visible fraction as one with low activation and low thresholds. Even perfect measurement of V_t cannot recover which latent organization generated it.

More generally, suppose

$$a \sim F_a, \qquad \theta \sim F_\theta.$$

Then the mapping

$$(F_a, F_\theta) \longmapsto \mathbb{P}(a > \theta)$$

is not injective. There exist many pairs

$$(F_a^{(1)}, F_\theta^{(1)}), \quad (F_a^{(2)}, F_\theta^{(2)}), \quad \dots$$

such that

$$\mathbb{P}_{(1)}(a > \theta) = \mathbb{P}_{(2)}(a > \theta) = \cdots.$$

This is not merely a statistical inconvenience. It places a principled limit on what can be inferred about disposition from expression alone.

The difficulty becomes even clearer when the expression margin

$$z_i = a_i - \theta_i$$

is used. Since

$$\mathbb{P}(E_i = 1) = \sigma(z_i),$$

behavioral observation can at best provide information about z_i. But the decomposition of z_i remains underdetermined. For any constant δ,

$$a_i' = a_i + \delta, \qquad \theta_i' = \theta_i + \delta$$

gives

$$a_i' - \theta_i' = a_i - \theta_i = z_i.$$

Thus an entire family of latent states is observationally equivalent with respect to expression probability.

This equivalence matters especially when behavior changes. Suppose expression probability rises between t_0 and t_1:

$$p_i(t_1) > p_i(t_0).$$

Then, because σ is monotone,

$$z_i(t_1) > z_i(t_0).$$

But

$$\Delta z_i = \Delta a_i - \Delta \theta_i,$$

so the observation remains compatible with

$$\Delta a_i > 0, \qquad \Delta \theta_i = 0,$$

or

$$\Delta a_i = 0, \qquad \Delta \theta_i < 0,$$

or

$$\Delta a_i > 0, \qquad \Delta \theta_i < 0,$$

as well as mixed cases in which activation and threshold move in opposing directions but one change dominates the other.

A useful way to visualize the latent field mathematically is to define the distance of every behavior from expression:

$$d_i = \theta_i - a_i.$$

Then

$$d_i < 0$$

marks behavior already beyond threshold,

$$d_i \approx 0$$

marks behavior poised near expression, and

$$d_i \gg 0$$

marks behavior deeply suppressed or weakly activated. The distribution

$$D_t(d)$$

of these distances across the behavioral field contains substantially more information than the visible set $\mathcal{E}_t$.

A permissive shift by $\delta > 0$,

$$\theta_i' = \theta_i - \delta,$$

transforms the distances as

$$d_i' = d_i - \delta.$$

The newly visible population is therefore

$$\mathcal{N}_\delta = \{b_i : 0 < d_i \leq \delta\}.$$

This set is theoretically important. It consists not of behaviors created by the intervention, but of behaviors whose previous invisibility depended on their location immediately behind the expression boundary.

The size of the release is

$$|\mathcal{N}_\delta|$$

and, in a continuous approximation with density $f_D(d)$,

$$\mathbb{P}(b_i \in \mathcal{N}_\delta) = \int_0^\delta f_D(d)\, dd.$$

For a small threshold shift,

$$\mathbb{P}(b_i \in \mathcal{N}_\delta) \approx f_D(0)\delta.$$

The behavioral effect of a modest permissive change therefore depends strongly on how much latent behavioral mass was already concentrated near the boundary. Two people undergoing an identical shift in affect can consequently display very different outward changes. If one has many dispositions clustered just below threshold and the other has few, the first may appear dramatically transformed while the second changes hardly at all.

This gives the threshold theory a different picture of behavioral stability. A person may look highly stable because the latent field itself is stable, but they may also look stable because a relatively stable threshold continually censors a turbulent field beneath it. Conversely, a sudden proliferation of new behavior need not imply that the person's underlying dispositions have changed suddenly. The threshold may simply have moved across a population that was already there.

The distinction is particularly important for longitudinal judgments of character. Suppose a person behaves with unusual openness in one environment and unusual reserve in another. It is tempting to infer two different underlying motivational states. Yet the difference could instead be produced by

$$a_i(t_1) \approx a_i(t_2), \qquad \theta_i(t_1, j_1) \neq \theta_i(t_2, j_2).$$

The behavioral field remains approximately constant while the accessible portion of it changes with context.

Observed behavior is thus not the behavioral field itself. It is a projection of that field through a moving boundary:

$$\mathcal{B}_t \xrightarrow{\theta_t} \mathcal{E}_t.$$

Once this is recognized, the threshold cannot remain an undifferentiated quantity. If different environments, relationships, physiological states, and social structures expose different portions of the same latent field, then the next question is not merely whether θ_i rises or falls, but what θ_i is made of. The apparent simplicity of a single boundary conceals several distinct forms of resistance whose separation will matter throughout the rest of the argument.

4 The Threshold Is Not One Thing

Writing expression as

$$\mathbb{P}(E_i = 1) = \sigma(a_i - \theta_i)$$

is useful only so long as θ_i is not allowed to become a name for everything the theory has not explained. If every influence that suppresses expression is simply placed inside a single scalar threshold, the model becomes difficult to falsify: whenever an expected behavior fails to appear, one can declare that its threshold must have been higher. The threshold therefore has to be decomposed into mechanisms that can, at least in principle, vary independently.

The first distinction is between the threshold a person would enforce if their regulatory machinery were fully effective and the degree to which that threshold is actually enforced. Let

$$\theta_i^*$$

denote the *normative expression threshold* for behavior i: the resistance implied by the person's present appraisal of the social, interpersonal, and practical consequences of expression. A minimal decomposition is

$$\theta_i^* = \theta_i^0 + w_C C_i + w_U U_i + w_N N_i + w_P P_i - w_R R_i,$$

where C_i represents anticipated social cost, U_i uncertainty about consequences, N_i norm enforcement, P_i relevant power asymmetry, and R_i perceived reversibility or recoverability of the act. The coefficients w_k determine how strongly each component contributes for the person, behavior, and context under consideration.

This expression should not be read as claiming that human judgment is literally linear. The linear form is useful because it exposes the sign structure. Increasing anticipated cost, uncertainty, normative prohibition, or power asymmetry tends to make expression harder; increasing reversibility tends to make it easier. More elaborate models can replace the linear combination with

$$\theta_i^* = \Theta_i^*(C_i, U_i, N_i, P_i, R_i, \ldots)$$

without altering the distinction that matters here.

Executive control belongs somewhere different.

It would be conceptually misleading to include executive capacity as merely another additive cost. A person may judge an utterance to be dangerous, inappropriate, or socially expensive and nevertheless fail to suppress it. Conversely, another person may make the same judgment and successfully act on it. The difference is not necessarily in the judged threshold θ_i^*. It can lie in the effectiveness with which that threshold is implemented.

Introduce therefore an enforcement gain

$$\kappa = \kappa(E_{\text{control}}), \qquad 0 < \kappa \leq 1,$$

and define the effective threshold as

$$\boxed{\theta_i = \kappa(E_{\text{control}})\theta_i^*.}$$

The limiting cases have straightforward interpretations. When

$$\kappa = 1,$$

the person's regulatory machinery fully implements the threshold generated by their current appraisal. When

$$0 < \kappa < 1,$$

the effective threshold is attenuated. And as

$$\kappa \to 0,$$

one obtains

$$\theta_i \to 0$$

even if

$$\theta_i^* \gg 0.$$

The person may continue to judge a behavior costly while becoming progressively less capable of preventing its expression.

This distinction separates two forms of disinhibition that otherwise look nearly identical from outside. Suppose first that anticipated social cost falls:

$$C_i \downarrow .$$

Then

$$\theta_i^* \downarrow$$

and therefore, holding enforcement constant,

$$\theta_i \downarrow .$$

The person is less inhibited because the act itself is now judged less costly.

Suppose instead that executive enforcement deteriorates:

$$\kappa \downarrow$$

while

$$\theta_i^* \approx \text{constant.}$$

Again,

$$\theta_i \downarrow,$$

but for a completely different reason. The judgment has not necessarily changed. Its behavioral enforcement has.

This gives the expression margin a more informative form:

$$z_i = a_i - \kappa\theta_i^*,$$

and hence

$$\mathbb{P}(E_i = 1) = \sigma(a_i - \kappa\theta_i^*) .$$

A change in expression can now arise through at least three distinguishable routes:

$$\Delta a_i \neq 0, \qquad \Delta\theta_i^* \neq 0, \qquad \Delta\kappa \neq 0.$$

The first is primarily motivational. The second changes the person's assessment of whether expression is costly or permissible. The third changes their capacity to enforce that assessment.

Differentiating the expression margin makes the distinction explicit:

$$dz_i = da_i - \kappa\, d\theta_i^* - \theta_i^*\, d\kappa.$$

Consequently,

$$dp_i = \sigma'(z_i)\,(da_i - \kappa\, d\theta_i^* - \theta_i^*\, d\kappa) .$$

The three terms correspond respectively to changes in activation, changes in the judged threshold, and changes in enforcement capacity. An identical increase in observed expression probability can therefore be generated by different movements through this state space.

This matters particularly for the interpretation of intoxication and exhaustion. To the extent that either acts by degrading executive self-regulation, its characteristic pathway is not necessarily

$$\theta_i^* \downarrow .$$

It may instead be

$$\kappa \downarrow .$$

Someone can therefore say something while intoxicated that they still judge, at the moment of saying it, to be socially costly. Their normative assessment need not have vanished merely because its capacity to regulate action has weakened. The distinction between judgment and enforcement is exactly what an undifferentiated θ_i would conceal.

Anonymity provides a contrasting case. If anonymity primarily reduces anticipated reputational cost, then its effect is more naturally represented as

$$C_i \downarrow \implies \theta_i^* \downarrow$$

with

$$\kappa \approx \text{constant}.$$

The person may retain perfectly intact regulatory capacity while rationally assigning less expected cost to expression because fewer consequences can be attached to their identity.

The same distinction allows perceived safety and intimacy to be represented without treating them as forms of executive impairment. A sufficiently safe relationship may alter several terms at once:

$$C_i \downarrow, \qquad U_i \downarrow, \qquad R_i \uparrow,$$

giving

$$\Delta\theta_i^* < 0$$

even while

$$\kappa$$

remains unchanged or increases. A person can therefore become more candid because they are more regulated and more secure, not because regulation has weakened. Outward "disinhibition" does not identify the mechanism producing it.

The decomposition can be generalized by collecting the threshold-generating variables into a vector

$$\mathbf{x} = (C, U, N, P, R, \ldots)^\top.$$

Then

$$\theta_i^* = \Theta_i^*(\mathbf{x}),$$

and for a small perturbation,

$$\Delta\theta_i^* \approx \nabla_{\mathbf{x}}\Theta_i^* \cdot \Delta\mathbf{x}.$$

Across many behaviors, define the Jacobian

$$J_{ik} = \frac{\partial\theta_i^*}{\partial x_k}.$$

An intervention S produces some state-dependent perturbation

$$\Delta\mathbf{x}^{(S)},$$

giving approximately

$$\Delta\boldsymbol{\theta}^{*(S)} \approx J\Delta\mathbf{x}^{(S)}.$$

The effective threshold change additionally depends on enforcement:

$$\Delta\theta_i^{(S)} \approx \kappa\Delta\theta_i^{*(S)} + \theta_i^*\Delta\kappa^{(S)}.$$

Two interventions can consequently produce comparable aggregate changes,

$$\left\|\Delta\boldsymbol{\theta}^{(A)}\right\| \approx \left\|\Delta\boldsymbol{\theta}^{(B)}\right\|,$$

while operating through nearly orthogonal causal directions:

$$\Delta\mathbf{x}^{(A)} \cdot \Delta\mathbf{x}^{(B)} \approx 0.$$

One might reduce reputational cost while another reduces uncertainty; a third might leave the entire appraisal vector approximately unchanged while reducing κ. Their behavioral outputs may resemble one another even though their internal causal structures do not.

There is one further decomposition that will become important later. Not every component of an expression threshold is specific to a particular behavior. Write

$$\theta_i^* = \theta_{\text{common}}^* + \delta_i,$$

where θ_{common}^* represents constraints shared across a broad region of the behavioral field and δ_i represents behavior-specific resistance.

A general change in perceived safety might act approximately as

$$\Delta\theta_{\text{common}}^* < 0, \qquad \Delta\delta_i \approx 0$$

for many i. By contrast, an explicit invitation such as "criticism is welcome in this meeting" may act primarily on a restricted class $\mathcal{C}$:

$$\Delta\delta_i < 0 \qquad \text{for } i \in \mathcal{C},$$

while

$$\Delta\delta_j \approx 0 \qquad \text{for } j \notin \mathcal{C}.$$

The first is a *global-gate intervention*; the second is a *content-targeted intervention*. Their distinction is empirically useful. A global-gate intervention should expose a comparatively heterogeneous band of latent behaviors because it does not specify which behavioral content is to be released. A content-targeted intervention should produce a more concentrated change in the behaviors whose specific thresholds it alters.

The threshold is therefore not one thing in either a psychological or a mathematical sense. Even the compact quantity

$$\theta_i = \kappa\theta_i^*$$

contains a distinction between judgment and enforcement, while θ_i^* itself contains separable estimates of cost, uncertainty, norms, power, reversibility, and behavior-specific permission. What looks from outside like a single phenomenon—someone becoming less inhibited—can be reached through several different trajectories through this space.

That multiplicity is not an inconvenience to be simplified away. It is what makes apparently similar forms of disinhibition theoretically distinguishable. Positive affect, intoxication, anonymity, exhaustion, intimacy, safety, and social permission need not constitute one psychological class merely because each can, under some circumstances, make previously unexpressed behavior visible.

5 Several Roads Through the Same Gate

The decomposition of the expression threshold makes it possible to use the word *disinhibition* without pretending that it names a single mechanism. Positive affect, intoxication, anonymity, exhaustion, perceived safety, intimacy, institutional permission, and elevated social status can all make previously suppressed behavior more likely to appear. Their similarity may end there. What they share is an outcome in the expression system, not necessarily a cause.

In the notation developed above, expression depends on

$$p_i = \sigma(a_i - \kappa\theta_i^*)\,,$$

with

$$\theta_i^* = \Theta_i^*(\mathbf{x}), \qquad \mathbf{x} = (C, U, N, P, R, \ldots)^\top.$$

A state or intervention S_k can therefore influence expression through at least three broad channels:

$$S_k \longrightarrow a_i,$$

$$S_k \longrightarrow \mathbf{x} \longrightarrow \theta_i^*,$$

or

$$S_k \longrightarrow \kappa.$$

These pathways can coexist, but they should not be collapsed merely because they terminate in the same observable event.

For each intervention S_k, define an *intervention profile*

$$\Pi(S_k) = \left(\frac{\partial a_i}{\partial S_k}, \frac{\partial \mathbf{x}}{\partial S_k}, \frac{\partial \kappa}{\partial S_k} \right).$$

The profile records which parts of the generative system the intervention moves. Two interventions may both increase expression probability while having very different profiles:

$$\frac{\partial p_i}{\partial S_A} > 0, \qquad \frac{\partial p_i}{\partial S_B} > 0,$$

without implying

$$\Pi(S_A) \approx \Pi(S_B).$$

Indeed, their profiles may have little overlap.

Anonymity offers a relatively clean illustration. Suppose its dominant effect is to reduce the expected reputational consequences of expression. Then, to a first approximation,

$$\frac{\partial C_i}{\partial S_{\text{anon}}} < 0,$$

while

$$\frac{\partial \kappa}{\partial S_{\text{anon}}} \approx 0.$$

Consequently,

$$\frac{\partial \theta_i^*}{\partial S_{\text{anon}}} < 0,$$

and expression becomes more probable because the person evaluates the act as less socially costly.

Intoxication can occupy a different region of the model. To the extent that its relevant effect is impaired executive regulation rather than altered social appraisal,

$$\frac{\partial \kappa}{\partial S_{\text{intox}}} < 0,$$

while it remains possible that

$$\frac{\partial \theta_i^*}{\partial S_{\text{intox}}} \approx 0.$$

The person may still regard an utterance as inappropriate, damaging, or dangerous. What has changed is the effectiveness with which that judgment constrains behavior.

Exhaustion can produce a superficially similar result through another degradation of regulatory capacity:

$$\Delta\kappa_{\text{fatigue}} < 0.$$

Yet even intoxication and exhaustion should not be assumed to be mechanistically identical merely because both can reduce κ. One may alter attention, working memory, impulsivity, salience, or activation in ways the other does not. The present model distinguishes causal coordinates; it does not claim that every phenomenon assigned to the same coordinate is physiologically equivalent.

Perceived safety presents almost the opposite case. A safer environment can lower the normative threshold while leaving enforcement intact:

$$C_i \downarrow, \qquad U_i \downarrow, \qquad R_i \uparrow,$$

so that

$$\theta_i^* \downarrow$$

while

$$\kappa \approx 1.$$

The resulting person may appear less inhibited precisely because their regulatory system is functioning normally in an environment that no longer requires as much inhibition.

Intimacy can deepen this effect by making the relevant threshold relationship-specific. If j denotes an interaction partner, then

$$\theta_i^* = \Theta_i^*(\mathbf{x}, j).$$

The same behavior can therefore satisfy

$$a_i < \theta_i^*(j_1)$$

but

$$a_i > \theta_i^*(j_2).$$

Nothing about the underlying disposition need change as the person moves from one audience to another. What changes is the expected consequence of making that disposition visible.

Institutional permission supplies a still different case because it can be explicitly content-targeted. Suppose an organization announces that dissent during a particular review process will not be penalized. The intervention need not lower the person's general social threshold. Instead, it can alter the behavior-specific component

$$\theta_i^* = \theta_{\text{common}}^* + \delta_i.$$

For dissent behaviors $i \in \mathcal{D}$,

$$\Delta\delta_i < 0,$$

while for unrelated behaviors $j \notin \mathcal{D}$,

$$\Delta\delta_j \approx 0.$$

The gate has not opened generally. A particular door within it has.

This distinction suggests a useful classification of permissive interventions. A *global-gate intervention* acts substantially on a common component:

$$\Delta\theta^*_{\text{common}} \neq 0$$

or on the common enforcement gain,

$$\Delta\kappa \neq 0.$$

A *content-targeted intervention* acts primarily on selected residual thresholds:

$$\Delta\delta_i \neq 0 \qquad \text{for } i \in \mathcal{C},$$

with little corresponding movement outside the targeted class.

The distinction matters because the two intervention types make different predictions. Let

$$\mathcal{N}(S) = \{i : E_i(S_0) = 0,\ E_i(S_1) = 1\}$$

denote the behaviors newly admitted to expression after intervention S. For a sufficiently broad global-gate intervention, the content distribution of

$$\mathcal{N}(S)$$

should depend substantially on the latent behavioral field already clustered near threshold. If that field is heterogeneous, the released behaviors should also be heterogeneous.

A content-targeted intervention instead selects a restricted behavioral class in advance. If criticism alone is explicitly licensed, then increased criticism provides little evidence for a general reduction in inhibition. The intervention itself already specifies the direction in which expression should change.

This difference can be represented by assigning each behavior a content vector $\mathbf{q}_i$, encoding dimensions such as affiliation, criticism, self-disclosure, norm violation, generosity, hostility, or other relevant properties. The dispersion of newly expressed content following an intervention can then be written

$$\Sigma_{\mathcal{N}}(S) = \text{Cov}\left(\mathbf{q}_i \mid i \in \mathcal{N}(S)\right).$$

All else equal, the threshold theory predicts greater content dispersion after a broad global-gate intervention than after a narrowly content-targeted one:

$$\text{tr}\,\Sigma_{\mathcal{N}}(S_{\text{global}}) > \text{tr}\,\Sigma_{\mathcal{N}}(S_{\text{targeted}}),$$

provided that the latent near-threshold population is itself heterogeneous. The qualification is essential. Opening a general gate cannot reveal heterogeneity that is not present behind it.

The same framework helps distinguish social status from simple safety. Elevated status may reduce the expected cost of some behaviors because the person is less vulnerable to retaliation:

$$P_i \downarrow \implies \theta_i^* \downarrow .$$

But it may simultaneously increase the expected cost of other behaviors if visibility, responsibility, or reputational exposure rises. Thus

$$\frac{\partial \theta_i^*}{\partial S_{\text{status}}}$$

need not have the same sign for every i. An intervention can be globally important without being globally permissive.

Positive affect is the theoretically interesting case because its intervention profile should not be assumed in advance. It might alter motivation,

$$\frac{\partial a_i}{\partial H} \neq 0,$$

alter appraisal,

$$\frac{\partial C_i}{\partial H} < 0, \qquad \frac{\partial U_i}{\partial H} < 0,$$

alter regulatory enforcement,

$$\frac{\partial \kappa}{\partial H} \neq 0,$$

or produce some combination of these effects. The claim that affect changes permission is therefore not a claim that its mechanism has already been identified. It is a claim that changes in activation alone are insufficient as the default explanation.

This yields a more useful meaning of disinhibition. Rather than naming a mechanism, it names a family of trajectories satisfying

$$\Delta z_i > 0$$

through different combinations of

$$\Delta a_i, \qquad \Delta \theta_i^*, \qquad \Delta \kappa.$$

For two states S_A and S_B, it is entirely possible that

$$\Delta z_i^{(A)} \approx \Delta z_i^{(B)}$$

while

$$\Pi(S_A) \not\approx \Pi(S_B).$$

The outward resemblance is therefore a case of *phenomenological convergence without mechanistic identity*.

That distinction prevents a common explanatory shortcut. Saying that intoxication, intimacy, anonymity, fatigue, safety, status, and positive affect all "lower inhibitions" is not yet an explanation. It is an observation that several causal paths terminate near the same boundary condition. The useful question is which coordinate moved, for which behaviors, by how much, and whether the intervention opened the behavioral field generally or selectively.

Once that distinction is preserved, a further consequence becomes visible. A genuinely broad change in permission carries comparatively little information about what kind of behavior will emerge. The direction of the change must then come increasingly from the distribution of dispositions already waiting near the boundary. In that sense, a gate can be causally powerful while remaining comparatively indifferent to the content that passes through it.

6 Disinhibition Has No Valence

The claim that a permissive shift can release both affectionate and critical behavior does not require the stronger claim that inhibition itself is wholly indifferent to content. Human expression thresholds plainly are not. Social norms may strongly discourage hostility while encouraging affection, or discourage emotional disclosure while permitting criticism. Different behaviors can begin from radically different positions relative to the expression boundary.

The relevant distinction is between the *content-specific structure* of the gate and a perturbation that moves a sufficiently broad part of that gate without specifying the content that will pass through it.

Recall the decomposition

$$\theta_i^* = \theta_{\text{common}}^* + \delta_i,$$

where θ_{common}^* represents a component shared across some region of the behavioral field and δ_i represents the behavior-specific departure from it. The effective threshold is

$$\theta_i = \kappa \left(\theta_{\text{common}}^* + \delta_i \right).$$

The residual δ_i can contain substantial information about content. For one person,

$$\delta_{\text{affection}} < 0$$

may reflect a learned environment in which affection is easy to express, while

$$\delta_{\text{criticism}} > 0$$

reflects a history in which criticism is costly. Another person may have approximately the reverse organization. Nothing in the threshold theory requires these residuals to be small.

Suppose, however, that an intervention changes primarily the common component:

$$\theta_{\text{common}}^{*'} = \theta_{\text{common}}^* - \Delta, \qquad \Delta > 0,$$

while approximately preserving the content-specific structure,

$$\delta_i' \approx \delta_i.$$

Then

$$\theta_i^{*'} = \theta_i^* - \Delta$$

across the affected behavioral class. The intervention has changed permissibility without specifying whether the newly admitted behaviors should be affectionate, hostile, humorous, confessional, generous, critical, or otherwise.

A reduction in enforcement gain can be broader still. If

$$\kappa' < \kappa,$$

then

$$\theta_i' = \kappa' \left(\theta_{\text{common}}^* + \delta_i\right)$$

falls for every positive threshold affected by the common regulatory system. Again, the intervention need not contain information about behavioral valence. It changes the effectiveness of suppression across whatever dispositions the system had previously been suppressing.

This is the sense in which disinhibition has no necessary valence.

It does *not* mean

$$\text{Cov}(\theta_i, \text{val}(b_i)) = 0.$$

That would be an implausibly strong claim. Thresholds can be highly correlated with behavioral content because cultures, relationships, institutions, and individual histories selectively reward and punish particular kinds of expression.

Nor does the theory require

$$\text{Cov}(\Delta\theta_i, \text{val}(b_i)) = 0$$

for every intervention. A content-specific permission can deliberately target one valence or behavioral class. An instruction such as

> Please be as critical as possible.

is explicitly directional. Its purpose is to lower δ_i for critical behavior while leaving other thresholds comparatively intact.

The narrower claim concerns interventions whose dominant effect is global rather than content-targeted. For such an intervention S_G,

$$\Delta\theta_{\text{common}}^* < 0$$

with

$$\Delta\delta_i \approx 0.$$

The change itself then carries relatively little information about the valence of the behavior that will cross threshold. Formally, if v_i denotes a valence or content descriptor associated with behavior

i, the relevant prediction is not that thresholds and valence are uncorrelated, but that the *common perturbation* is approximately independent of v_i:

$$I\left(\Delta\theta^*_{\text{common}}; v_i\right) \approx 0,$$

where $I(\cdot;\cdot)$ denotes mutual information.

The intervention can therefore be powerful while being informationally sparse with respect to behavioral direction.

This distinction clarifies the opening example. Let

$$b_A = \text{affection}, \qquad b_C = \text{criticism},$$

with expression margins

$$z_A = a_A - \kappa\left(\theta^*_{\text{common}} + \delta_A\right)$$

and

$$z_C = a_C - \kappa\left(\theta^*_{\text{common}} + \delta_C\right).$$

Suppose initially

$$z_A < 0, \qquad z_C < 0,$$

but both lie near zero. A common permissive shift

$$\theta^*_{\text{common}} \mapsto \theta^*_{\text{common}} - \Delta$$

produces

$$z'_A = z_A + \kappa\Delta$$

and

$$z'_C = z_C + \kappa\Delta.$$

If

$$-\kappa\Delta < z_A < 0$$

and

$$-\kappa\Delta < z_C < 0,$$

then both cross into positive territory:

$$z'_A > 0, \qquad z'_C > 0.$$

Nothing in this transition requires

$$a_A \uparrow$$

or

$$a_C \uparrow .$$

Nor does it require affection and criticism to be motivationally linked. The same movement of a common boundary is sufficient to expose both.

This generalizes immediately. Let

$$\mathcal{B} = \{b_1, \ldots, b_n\}$$

contain behaviors with content descriptors

$$v_i \in \mathcal{V},$$

where $\mathcal{V}$ may include positive and negative social valence but need not be restricted to a one-dimensional moral scale. A global threshold shift by Δ releases

$$\mathcal{N}_\Delta = \{b_i : -\kappa\Delta < z_i \leq 0\} .$$

The content distribution of the newly expressed set is then inherited from the latent content distribution in that near-threshold band:

$$\mathbb{P}\left(v_i = v \mid b_i \in \mathcal{N}_\Delta\right).$$

The intervention determines the width of the band. The latent behavioral field determines what occupies it.

This yields a useful separation between *directional* and *permissive* information. A motivational intervention can contain directional information:

$$S_M \longrightarrow \Delta a_i$$

for selected i. A global permissive intervention instead changes the selection boundary:

$$S_G \longrightarrow \Delta\theta^*_{\text{common}}.$$

In an idealized limiting case,

$$I(S_M; v_i \mid E_i = 1) > 0,$$

because knowing the intervention helps predict what kind of behavior was activated, whereas

$$I(S_G; v_i \mid E_i = 1)$$

may be much smaller because knowing that a general gate opened tells us little about which latent disposition happened to be waiting immediately behind it.

This distinction prevents a subtle circularity. If every newly expressed behavior is retrospectively declared to be "what the disinhibited state caused," then the theory becomes incapable of distinguishing activation from release. The threshold account instead asks whether the intervention predicts the *content* of the behavior or primarily predicts an increased probability that pre-existing content becomes visible.

The difference can be tested by comparing global and targeted interventions. Let S_G denote a global-gate manipulation and S_T a content-targeted one. Then, under otherwise comparable conditions, one expects

$$H\left(V \mid E = 1, S_G\right) > H\left(V \mid E = 1, S_T\right),$$

where $H(V \mid \cdot)$ is the entropy of expressed behavioral content. A targeted intervention should concentrate expression in the class it licenses. A global intervention should reveal more of whatever heterogeneous content was already distributed near threshold.

This is not an unconditional law. If the latent near-threshold field is itself homogeneous, opening a general gate will produce homogeneous behavior. The prediction therefore depends on the interaction between two quantities:

breadth of threshold perturbation

and

heterogeneity of latent dispositions.

Neither alone determines the observed result.

This point is important because the phrase "positive affect produces positive behavior" conflates the sign of the state with the sign of its behavioral output. If positive affect has a global permissive component, then some of its behavioral consequences should instead be inherited from whatever dispositions the person ordinarily suppresses. A generally warm but reserved person may become conspicuously affectionate. A generally critical but tactful person may become conspicuously blunt. Someone carrying both dispositions may become both at once.

The same logic applies beyond affect. Anonymity can release extraordinary kindness and extraordinary cruelty. Intimacy can permit tenderness and argument. Safety can permit vulnerability and dissent. A decline in inhibition does not tell us, by itself, what will emerge, because inhibition is partly a selection mechanism rather than a source of behavioral content.

That observation makes heterogeneity more than an incidental complication. If a global gate reveals what was already different between people, then lowering that gate should not merely change the mean level of expression. Under the right conditions, it should change the shape and dispersion of the observed behavioral distribution itself.

Disposition Is Already Organized

There is, however, a danger in describing the latent behavioral field as though it were merely a collection of independent impulses waiting behind a gate. Alfred Adler's individual psychology provides a useful correction. In an Adlerian account, behavior is intelligible not simply by tracing it backward to an internal cause, but by asking what place it occupies within an individual's characteristic orientation toward other people, difficulty, belonging, competence, and anticipated

consequence. Behavior is purposive in the broad sense that it participates in an organized "style of life."

This matters for the present theory because the activation variable a_i should not be interpreted as the strength of an isolated impulse. More generally, it is the activation of behavior b_i within an organized field of possible action. We may therefore write

$$a_i(t) = A_i(D, \Lambda_t, C_t),$$

where D denotes relatively persistent dispositions, C_t the immediate context, and Λ_t the person's current organization of goals, expectations, and socially situated purposes. Expression remains governed by

$$\mathbb{P}(E_i = 1) = \sigma[A_i(D, \Lambda_t, C_t) - \kappa\,(\theta^*_{\text{common}} + \delta_i)]\,.$$

The addition is small mathematically but important conceptually. Lowering a threshold does not open a box containing an inventory of psychologically independent behaviors. It reveals more of an already structured system.

Adler's emphasis on *social interest* is especially relevant here. Expression occurs inside a social field whose anticipated consequences are themselves part of behavioral organization. The person who withholds criticism from a colleague, affection from a stranger, vulnerability from an adversary, or disagreement from an authority is not necessarily experiencing the same underlying activation with four arbitrary barriers placed on top. The relationship partly determines both what becomes activated and what becomes permissible.

Thus the model should distinguish

$$C_t \longrightarrow a_i$$

from

$$C_t \longrightarrow \theta^*_i.$$

A social situation can change what a person is disposed to do and, independently, how costly doing it appears. The two effects may reinforce one another or move in opposite directions.

This also gives a more careful interpretation of apparently contradictory behavior. Suppose affection and criticism both become more visible under a permissive state. An atomistic interpretation might say that two unrelated impulses happened to cross threshold simultaneously. An Adlerian reading allows a more organized possibility: both behaviors may belong to the same social orientation. Increased security within a relationship may make a person simultaneously more willing to affirm the other and more willing to correct them. Warmth and disagreement need not merely coexist behind the same gate; they can participate in the same interpersonal project.

Let behaviors be grouped by some latent organization Λ. Then their activations need not be independent:

$$\text{Cov}(a_i, a_j \mid \Lambda) \neq 0.$$

The earlier independence assumption used for simple examples should therefore be understood as a mathematical convenience, not a psychological commitment. A more general joint model permits

$$\mathbf{a} = \big(a_1, \ldots, a_n\big) \sim F_{\mathbf{a}}(\,\cdot \mid D, \Lambda, C_t).$$

A global permissive shift can still expose heterogeneous behaviors, but the newly visible behaviors may arrive in structured clusters rather than as independent samples from a behavioral inventory.

This is also where Adler helps resist the authenticity fallacy developed later in the essay. If behavior is organized relationally and purposively, there is no obvious reason to regard the least inhibited behavior as the most authentic. Self-restraint can itself belong to the person's style of life. So can tact, cooperation, concealment, courage, avoidance, generosity, rivalry, and deliberate silence. Removing a constraint reveals what the system does under reduced constraint; it does not peel away an artificial layer until an unmediated person underneath is finally exposed.

The threshold theory therefore need not choose between constraint and purpose. A useful synthesis is

disposition and purpose shape what approaches the gate; permission shapes what passes through it.

The distinction becomes important when the gate moves globally. A broad reduction in inhibition carries relatively little information about which behavioral content will appear, but this does not mean that the resulting content is random. Its structure can come from the person's existing organization of dispositions, purposes, relationships, and learned social strategies. What appears as increased behavioral variance across people may therefore coexist with considerable coherence within each person.

7 The Variance Prediction

If broad disinhibition changes primarily the boundary of expression rather than the direction of motivation, it should leave a distinctive statistical trace. The effect should not merely be an increase in the mean frequency of some behavior. It should expose differences that were already present but previously censored.

This is the point at which the threshold theory becomes experimentally distinguishable from a sufficiently simple motivational account.

Suppose a population contains individuals $p = 1, \ldots, m$, each with a latent behavioral field

$$\mathcal{B}^{(p)} = \left\{ b_1^{(p)}, \ldots, b_n^{(p)} \right\}.$$

For each behavior,

$$z_i^{(p)} = a_i^{(p)} - \kappa^{(p)} \left(\theta_{\text{common}}^{*(p)} + \delta_i^{(p)} \right)$$

is its expression margin, with

$$\mathbb{P}\Big(E_i^{(p)} = 1 \Big) = \sigma\Big(z_i^{(p)} \Big).$$

Consider first a global-gate intervention S_G whose principal effect is

$$\Delta\theta^*_{\text{common}} < 0$$

while leaving the content-specific residuals approximately intact:

$$\Delta\delta_i \approx 0.$$

Alternatively, the intervention may operate through enforcement,

$$\Delta\kappa < 0,$$

again without specifying which particular behavioral content should become more active.

In either case, behaviors previously clustered immediately below the expression boundary become newly observable. If the global threshold changes by an effective amount $\Delta > 0$, the newly admitted set is approximately

$$\mathcal{N}^{(p)}_{\Delta} = \left\{ b^{(p)}_i : -\Delta < z^{(p)}_i \leq 0 \right\}.$$

The intervention selects behaviors by their *distance from expression*, not by their valence.

If different people contain different distributions of behavior within this near-threshold region, lowering the common gate should therefore expose those differences. The result is a prediction about dispersion:

broad permission should reveal latent heterogeneity.

This requires more care than the simple claim

$$\text{Var}(B \mid \theta \downarrow) > \text{Var}(B \mid \theta \uparrow),$$

because variance can increase or decrease for purely mechanical reasons depending on how behavior is coded, where the threshold begins, and what the latent distribution looks like. The stronger claim is conditional. Given a heterogeneous near-threshold population and an intervention that changes permission without strongly specifying behavioral direction, more of that heterogeneity should become observable.

Let q_i denote a scalar behavioral dimension, or let $\mathbf{q}_i$ denote a multidimensional content representation. Before the intervention, the observed distribution is

$$P_{\text{obs}}(\mathbf{q}) = P\left(\mathbf{q}_i = \mathbf{q} \mid z_i > 0\right).$$

After a global permissive shift,

$$P'_{\text{obs}}(\mathbf{q}) = P\left(\mathbf{q}_i = \mathbf{q} \mid z_i > -\Delta\right).$$

The change in observable diversity therefore depends on the content distribution within

$$-\Delta < z_i \leq 0.$$

If that band contains behavioral content more heterogeneous than the previously visible set, then

$$H(P'_{\text{obs}}) > H(P_{\text{obs}}),$$

where H denotes entropy, and an analogous increase may appear in a variance or covariance measure when the behavioral representation permits one.

The important point is that the heterogeneity is not produced by the intervention. It is uncovered by it.

This produces a different prediction from a strongly directional motivational manipulation. Suppose an intervention selectively increases activation of affiliative behaviors:

$$\Delta a_i > 0 \qquad \text{for } i \in \mathcal{A},$$

with little change elsewhere. Then the intervention itself contains information about the expected direction of behavioral change. Across individuals one should observe some convergence toward the targeted class, subject to differences in responsiveness.

A global-gate intervention instead says comparatively little about direction:

$$\Delta\theta^*_{\text{common}} < 0$$

for a broad behavioral field. Its behavioral consequences should therefore depend more strongly on what each person already carries near threshold.

For two people p and q, it is entirely possible that

$$\Delta\theta^{*(p)}_{\text{common}} \approx \Delta\theta^{*(q)}_{\text{common}}$$

while

$$\Delta\mathbf{B}^{(p)}_{\text{obs}} \not\approx \Delta\mathbf{B}^{(q)}_{\text{obs}}.$$

The common intervention produces divergent expression because the latent fields differ.

This yields a useful population-level prediction. Let

$$\mathbf{Y}^{(p)}$$

represent an individual's observed behavioral profile. If a manipulation is primarily permissive and latent dispositions vary substantially between individuals, then after controlling for the intervention's mean effect one expects increased between-person dispersion:

$$\operatorname{tr}\operatorname{Cov}_p\left(\mathbf{Y}^{(p)} \mid S_G\right) > \operatorname{tr}\operatorname{Cov}_p\left(\mathbf{Y}^{(p)} \mid S_0\right),$$

over the behavioral dimensions whose thresholds were broadly affected.

This is the first form of the variance prediction:

global disinhibition can increase between-person behavioral variance.

But the Adlerian qualification introduced above prevents a mistaken second step. Greater variance *between people* does not imply randomness *within people*.

If an individual's behavioral dispositions are organized by a relatively persistent style of action, social orientation, or purposive structure $\Lambda^{(p)}$, then

$$\mathbf{a}^{(p)} \sim F_{\mathbf{a}}\left(\cdot \mid D^{(p)}, \Lambda^{(p)}, C_t\right).$$

Behaviors belonging to the same organization may therefore have correlated activations:

$$\operatorname{Cov}\left(a_i^{(p)}, a_j^{(p)} \mid \Lambda^{(p)}\right) \neq 0.$$

A permissive shift can consequently reveal a coherent cluster of behaviors within one individual while revealing a very different coherent cluster within another.

One person may become simultaneously more affectionate, confessional, and playful. Another may become more argumentative, competitive, and irreverent. A third may become both more affectionate and more critical because those behaviors belong, for that person, to the same interpersonal orientation: greater security permits both affirmation and disagreement.

The population therefore admits the combination

$$\text{between-person heterogeneity} \quad + \quad \text{within-person organization}.$$

Indeed, this combination is more characteristic of the threshold theory than an indiscriminate increase in behavioral noise would be.

Let

$$\Sigma_{\text{between}} = \operatorname{Cov}_p\left(\mathbb{E}[\mathbf{Y}^{(p)} \mid p]\right)$$

measure between-person dispersion, while

$$\Sigma_{\text{within}}^{(p)} = \operatorname{Cov}\left(\mathbf{Y}^{(p)} \mid p\right)$$

describes variation within an individual. A global permissive intervention need not imply

$$\Sigma_{\text{within}}^{(p)} \uparrow$$

in every sense. What it predicts more directly is that the mapping from latent individual differences to visible individual differences becomes less censored.

This can be stated as an attenuation problem. Let $D_i^{(p)}$ denote an independently estimated latent disposition toward behavior i. Under strong inhibition, observed behavior may correlate only weakly with that disposition:

$$\operatorname{Corr}\left(D_i^{(p)}, Y_i^{(p)} \mid \theta_{\text{high}}\right) \approx 0,$$

because many individuals with substantial D_i remain below threshold.

After a sufficiently broad permissive shift, the threshold model predicts

$$\operatorname{Corr}\left(D_i^{(p)}, Y_i^{(p)} \mid \theta_{\text{lower}}\right) > \operatorname{Corr}\left(D_i^{(p)}, Y_i^{(p)} \mid \theta_{\text{high}}\right),$$

provided the intervention does not itself substantially rewrite $D_i^{(p)}$ or selectively activate the same behavior.

This trait-expression correlation is a sharper test than variance alone. A mere increase in behavioral variability could arise from noise. The threshold account predicts something more structured: newly visible behavior should become more predictable from information about dispositions that existed *before* the gate moved.

That qualification creates an important methodological constraint. $D_i^{(p)}$ cannot be estimated from the same disinhibited episode whose behavior it is then used to predict. Doing so would make the test circular. If a person becomes unusually critical under intervention and that episode is then used as evidence that the person possessed a latent critical disposition, the theory has predicted nothing.

The estimate must therefore be independent.

One possibility is informant evidence collected before the intervention. Another is behavior measured in a separate context in which the relevant behavior already carries low expression cost. Longitudinal evidence can also be used if it identifies stable patterns without incorporating the episode under investigation. Formally, if

$$\widehat{D}_i^{\text{latent}} = \Psi(X_{\text{independent}}),$$

then the required independence condition is approximately

$$X_{\text{independent}} \perp Y_i^{\text{test}} \mid D_i,$$

except through the latent disposition being estimated.

The empirical prediction is then

$$\widehat{D}_i^{\text{latent}} \longrightarrow Y_i^{\text{test}}$$

more strongly under a broad permissive intervention than under a strongly censoring condition.

This produces a useful comparison among three hypotheses.

Under a predominantly directional motivational account,

$$S \longrightarrow a_i$$

for a selected behavioral class, so the intervention itself should explain a large fraction of the direction of change.

Under a predominantly noisy-disruption account,

$$S \longrightarrow \varepsilon_i \uparrow,$$

so behavioral variance may increase without becoming more predictable from prior dispositions.

Under the threshold account,

$$S \longrightarrow \theta^*_{\text{common}} \downarrow$$

or

$$S \longrightarrow \kappa \downarrow,$$

so variance can increase while the newly visible behavior simultaneously becomes *more* related to independently measured latent structure.

Schematically,

mechanism	dispersion	relation to prior disposition
directional activation	possibly lower or targeted	not necessarily stronger
random disruption	higher	weaker
broad permission	potentially higher	stronger

The third combination is the distinctive one.

A further prediction follows from the distinction between global and content-targeted interventions developed in the previous section. If an intervention explicitly licenses criticism, then observing more criticism does not discriminate between activation and threshold accounts very effectively. The manipulation has already selected a content class.

By contrast, if an intervention alters a broad gate,

$$\Delta\theta^*_{\text{common}} < 0$$

while

$$\Delta\delta_i \approx 0,$$

then behavioral direction should be less predictable from the intervention itself and more predictable from independent information about the person.

One can express the contrast information-theoretically. Let S identify the intervention, D the independently measured latent organization, and Y the newly expressed behavior. A strongly content-targeted intervention may satisfy

$$I(Y;S) \gg I(Y;D \mid S),$$

whereas a sufficiently global permissive intervention should shift explanatory weight toward

$$I(Y;D \mid S).$$

The strongest version of the threshold prediction is therefore not simply

$$\text{Var}(Y) \uparrow .$$

It is the joint pattern

$$\boxed{\text{Var}_{\text{between}}(Y) \uparrow, \qquad I(Y;D_{\text{independent}} \mid S) \uparrow,}$$

under an intervention independently shown to act broadly on permission.

This also explains why apparently contradictory expressions are theoretically valuable rather than troublesome observations. If elevated positive affect were found to increase affection in one

person, criticism in another, playfulness in a third, and both warmth and bluntness in a fourth, such heterogeneity would initially look like noise under a simple valence account. Under a threshold account, it is exactly the pattern worth examining. The critical question is whether those directions can be predicted from what was already known about each person before the affective shift occurred.

The gate hypothesis therefore becomes strongest precisely where a vague theory of "disinhibition" would become weakest. It does not explain heterogeneous behavior by saying that inhibition disappeared and anything could happen. It predicts that when a broad gate moves, the variation that appears should bear the signature of the organized behavioral field that had previously been partly hidden.

What appears newly spontaneous may consequently have a history. The threshold at a particular moment is only the current boundary of expression; repeated approval, punishment, embarrassment, safety, intimacy, and successful disclosure can move that boundary over much longer periods. The behavioral field and its gate are both products of time, but they need not be products of time in the same way.

8 State, Trait, and Suppression History

The variance prediction depends on a distinction that ordinary descriptions of personality often blur. A behavior can be characteristic of a person without being continuously activated, activated without being expressed, and expressed without representing a stable trait. These are different levels of the generative process.

For behavior i, write activation as

$$a_i(t) = D_i + \eta_i(S_t) + \xi_i(C_t) + \varepsilon_i(t),$$

where D_i is a relatively persistent dispositional component, $\eta_i(S_t)$ is a state-dependent perturbation, $\xi_i(C_t)$ captures contextual activation, and $\varepsilon_i(t)$ collects shorter-lived variation not otherwise represented.

The expression probability remains

$$\mathbb{P}(E_i(t) = 1) = \sigma[a_i(t) - \kappa(t)\theta_i^*(t)]\,.$$

Substituting the activation decomposition gives

$$\mathbb{P}(E_i(t) = 1) = \sigma[D_i + \eta_i(S_t) + \xi_i(C_t) + \varepsilon_i(t) - \kappa(t)\theta_i^*(t)]\,.$$

Observed behavior therefore depends simultaneously on something relatively persistent, something state-dependent, something contextual, and something regulatory. Calling the resulting expression a "trait" merely because it occurred repeatedly can confuse these sources.

The threshold has a history of its own.

Suppose a person expresses behavior b_i and receives some consequence $r_i(t)$. A simple learning rule is

$$\theta_i^*(t+1) = \theta_i^*(t) - \gamma r_i(t),$$

where $\gamma > 0$ is an updating rate and the sign convention assigns positive r_i to outcomes that make future expression seem safer or more worthwhile. Thus

$$r_i(t) > 0 \quad \Longrightarrow \quad \theta_i^*(t+1) < \theta_i^*(t),$$

while

$$r_i(t) < 0 \quad \Longrightarrow \quad \theta_i^*(t+1) > \theta_i^*(t).$$

This is intentionally schematic. Social learning is not plausibly a one-parameter reinforcement process. The point is that suppression history can be represented as an updating process rather than treated as a mysterious static property of personality.

A more general formulation is

$$\theta_i^*(t+1) = \theta_i^*(t) + \Gamma_i\left(o_i(t), j, C_t, \theta_i^*(t)\right),$$

where $o_i(t)$ is the experienced outcome of expression, j identifies the interaction partner or audience, and Γ_i specifies how that experience changes future appraisal.

This allows learning to be asymmetric. A single humiliating consequence may raise a threshold much more than several mildly successful expressions lower it. One can represent this with separate rates,

$$\theta_i^*(t+1) = \theta_i^*(t) - \gamma_+ r_i^+(t) + \gamma_- r_i^-(t),$$

where

$$r_i^+(t) = \max(r_i(t), 0), \qquad r_i^-(t) = \max(-r_i(t), 0),$$

and there is no requirement that

$$\gamma_+ = \gamma_-.$$

Indeed, for socially dangerous behavior one might expect

$$\gamma_- > \gamma_+.$$

A threshold can therefore acquire hysteresis. The path by which it reached its present value matters.

Two people facing the same present environment can consequently have

$$C_t^{(p)} \approx C_t^{(q)}$$

while nevertheless satisfying

$$\theta_i^{*(p)}(t) \neq \theta_i^{*(q)}(t)$$

because their histories differ.

This makes the threshold explicitly path-dependent:

$$\theta_i^*(t) = \Theta_i^*\left(C_t, \mathcal{H}_i(t)\right),$$

where

$$\mathcal{H}_i(t) = \{o_i(s), C_s, j_s\}_{s<t}$$

denotes the relevant history of prior expressions and consequences.

The behavioral field is therefore filtered not only through present social conditions but through accumulated estimates of what similar actions have cost before.

This gives a more precise meaning to *suppression history*. Consider two people with approximately equal dispositions,

$$D_i^{(p)} \approx D_i^{(q)},$$

and approximately equal present activation,

$$a_i^{(p)}(t) \approx a_i^{(q)}(t).$$

If their histories have produced

$$\theta_i^{*(p)} \ll \theta_i^{*(q)},$$

then

$$\mathbb{P}(E_i^{(p)} = 1) \gg \mathbb{P}(E_i^{(q)} = 1)$$

can follow without any substantial difference in the disposition itself.

The inverse case is equally important. Similar rates of expression need not imply similar dispositions. If

$$D_i^{(p)} > D_i^{(q)}$$

but

$$\theta_i^{*(p)} > \theta_i^{*(q)},$$

the two differences can partly cancel:

$$D_i^{(p)} - \kappa^{(p)}\theta_i^{*(p)} \approx D_i^{(q)} - \kappa^{(q)}\theta_i^{*(q)}.$$

The observed behavioral profiles may then look nearly identical even though their latent organizations differ substantially.

This is another reason the trait-correlation test developed in the previous section requires independent evidence. Repeated non-expression cannot safely be used as evidence that D_i is small when the same observations are also consistent with a large and historically reinforced threshold.

The familiar claim that some disinhibiting condition "reveals the real person" becomes especially unstable under this decomposition. Consider the common intuition that intoxication exposes dispositions normally concealed by self-control. If intoxication primarily reduces enforcement,

$$\kappa_{\text{intox}} < \kappa_{\text{sober}},$$

then it may indeed make some previously suppressed dispositions visible. But the same state may also alter activation:

$$\eta_i(S_{\text{intox}}) \neq 0,$$

attention,

$$\xi_i(C_t) \mapsto \xi_i'(C_t),$$

and short-term variability,

$$\text{Var}(\varepsilon_i \mid S_{\text{intox}}) \uparrow .$$

The observed behavior is therefore generated by

$$D_i + \eta_i(S_{\text{intox}}) + \xi_i(C_t) + \varepsilon_i - \kappa_{\text{intox}}\theta_i^*,$$

not by D_i alone.

There is no reason to expect all of these terms to move in the same direction.

A person might become more affectionate partly because an affiliative disposition was already near threshold, more argumentative because enforcement has weakened, less cautious because consequences receive less weight, and less capable of expressing some complicated thought because the state has impaired the cognitive machinery required to formulate it. The intervention can expose, amplify, distort, and suppress different behaviors simultaneously.

The language of a single hidden self is too coarse for this structure.

An Adlerian formulation makes the problem deeper still. The persistent component D_i should not necessarily be imagined as a collection of independent trait magnitudes. Behavior can be organized around a more general orientation Λ, representing characteristic expectations, purposes, social strategies, and patterns of interpretation. We can therefore replace

$$D_i$$

with

$$D_i(\Lambda),$$

or more generally write

$$\mathbf{D} = \mathcal{D}(\Lambda).$$

Then

$$\mathbf{a}(t) = \mathcal{A}\left(\Lambda, S_t, C_t\right) + \boldsymbol{\varepsilon}_t.$$

The significance of this move is that history can modify more than thresholds. Repeated social experience can alter both what a person expects expression to cost and the organization from which particular actions become attractive in the first place.

It is therefore useful to distinguish two learning processes:

$$\mathcal{H}_t \longrightarrow \theta_i^*(t)$$

and

$$\mathcal{H}_t \longrightarrow \Lambda_t.$$

The first is learning about permission. The second is learning that changes the organization of behavior itself.

A child repeatedly punished for disagreement might learn

$$\theta^*_{\text{dissent}} \uparrow,$$

while retaining a strong disposition toward dissent. But sufficiently extended experience might also alter how disagreement is conceived, anticipated, or pursued, changing

$$\Lambda_t$$

and therefore changing activation itself. The two outcomes can look similar from outside while representing very different developmental histories.

This suggests a two-timescale model. Let threshold learning occur relatively quickly,

$$\theta_i^*(t+1) = \theta_i^*(t) + \Gamma_i(o_t),$$

while dispositional organization changes more slowly,

$$\Lambda_{t+1} = \Lambda_t + \epsilon\Omega(\mathcal{H}_t), \qquad 0 < \epsilon \ll 1.$$

A person may therefore learn first to suppress a behavior and only later, perhaps after years, become someone for whom the behavior is no longer strongly activated at all.

The reverse is possible as well. Repeated safe expression can lower a threshold before it changes the person's broader social orientation:

$$\theta_i^* \downarrow$$

may precede

$$D_i(\Lambda) \uparrow .$$

What begins as permission can eventually participate in personality development.

This temporal structure complicates any sharp distinction between "who the person is" and "what the environment permits." The distinction remains analytically useful at a given time, but over longer horizons the variables can act on one another:

$$\mathcal{H} \to \theta^* \to E \to \mathcal{H}' \to \Lambda \to a.$$

Expression changes experience; experience changes future thresholds; repeated experience can alter the organization of activation; that altered organization then changes what approaches the threshold in later situations.

The resulting system contains feedback:

$$a_i(t) \to E_i(t) \to o_i(t) \to \theta_i^*(t+1) \to E_i(t+1),$$

with a slower loop through dispositional organization,

$$o_i(t) \to \Lambda_{t+k} \to a_i(t+k).$$

A threshold is therefore neither a fixed wall nor merely an instantaneous calculation. It is a learned boundary embedded in a history of expression.

This provides a more precise interpretation of social reserve. Reserve can arise because activation is low, because normative thresholds are high, because enforcement is strong, because past outcomes have raised particular thresholds, because a person's organized purposes make certain expressions unlikely, or because several of these conditions coincide.

Likewise, a sudden period of openness need not have a single psychological meaning. It can represent a change in activation,

$$\Delta a_i > 0,$$

a change in appraisal,

$$\Delta \theta_i^* < 0,$$

a temporary reduction in enforcement,

$$\Delta \kappa < 0,$$

or the delayed consequence of repeated successful interactions that have gradually changed the threshold over time.

This last possibility matters because expression itself can become evidence for future permission. If an action repeatedly crosses threshold and produces no serious cost, the system receives information that its previous estimate may have been too conservative.

For a sequence of non-negative outcomes,

$$r_i(t), r_i(t+1), \ldots, r_i(t+k) \geq 0,$$

one may obtain

$$\theta_i^*(t+k+1) < \theta_i^*(t),$$

until some lower asymptote is reached.

A behavior that once required unusual courage can thereby become ordinary. Nothing about the behavior itself need have changed. What has changed is the history against which its cost is estimated.

This is particularly visible in relationships. Repeated disclosure without punishment, disagreement without rupture, affection without rejection, and eccentricity without ridicule can progressively change what is expressible between particular people. The resulting phenomenon is commonly

called trust, but in the present framework trust can be given a narrower operational meaning: it is partly a learned alteration in the expected cost structure of expression.

That alteration is not yet the whole of trust. But it is enough to explain why the same behavioral disposition can remain almost permanently hidden in one relationship and become effortless in another.

9 Mood as Information, Mood as Regulation

The threshold model has so far treated affect as one possible source of permissive change without committing to a particular psychological mechanism. That restraint matters because there are at least two importantly different ways in which affect could alter an expression threshold. A positive state can change what a person infers about the surrounding world, or it can change the regulatory machinery through which behavior is selected and suppressed. These routes may converge behaviorally while remaining causally distinct.

The first possibility is close to the mood-as-information tradition associated with Schwarz and Clore. On this account, affective state can function as information during judgment. A positive state may contribute to an appraisal that circumstances are comparatively benign, that vigilance can be relaxed, or that the immediate environment contains fewer threats than it otherwise would appear to contain.

In the present notation, let

$$H_t$$

denote an affective state and let

$$A_t = \mathcal{A}(H_t, C_t)$$

denote an appraisal partly informed by that state. The appraisal can then alter the components from which the normative threshold is constructed:

$$H_t \longrightarrow A_t \longrightarrow (C_i, U_i, R_i, \ldots) \longrightarrow \theta_i^*.$$

For example, if positive affect contributes to a safer-world appraisal, one might observe

$$H_t \uparrow \implies C_i \downarrow, \qquad U_i \downarrow, \qquad R_i \uparrow,$$

and therefore

$$\theta_i^* \downarrow .$$

Nothing in this pathway requires executive control to deteriorate. Indeed,

$$\kappa \approx 1$$

is perfectly compatible with increased expression. The person may be regulating themselves effectively while operating with a different estimate of what requires regulation.

This is an important alternative to the loose language of "losing inhibitions." A person who becomes more candid while happy may not have lost anything. Their regulatory system may simply be applying a lower normative threshold because the situation is being appraised as safer.

The second pathway is different. Affect may alter some component of behavioral regulation without first being interpreted as evidence about external conditions. Schematically,

$$H_t \longrightarrow \text{regulatory state} \longrightarrow \theta_i.$$

This pathway can itself take several forms. Affect might alter enforcement,

$$H_t \longrightarrow \kappa,$$

or modify some threshold-generating process directly,

$$H_t \longrightarrow \theta_i^*,$$

without the change being mediated by a proposition resembling "the world is safe."

It is therefore useful to distinguish an appraisal-mediated contribution $M_A(H_t)$ from a regulatory contribution $M_R(H_t)$. A simplified structural model is

$$\theta_i^*(t) = \theta_{i,0}^* - \alpha_i M_A(H_t) - \beta_i M_R(H_t),$$

or, retaining the enforcement distinction,

$$\theta_i(t) = \kappa(H_t)\left[\theta_{i,0}^* - \alpha_i M_A(H_t) - \beta_i M_R(H_t)\right].$$

The coefficients α_i and β_i need not be equal across behaviors. Affect may change the perceived cost of social disclosure more than the perceived cost of physical risk, for example, while another regulatory effect may act much more broadly.

The two causal structures can be represented as

$$H \to A \to \theta^* \to E$$

and

$$H \to R \to (\theta^*, \kappa) \to E.$$

They produce the same qualitative endpoint when

$$\frac{d\theta_i}{dH} < 0,$$

but the derivative alone does not identify which pathway generated it.

Expanding the derivative makes the ambiguity visible. Since

$$\theta_i = \kappa(H)\theta_i^*(H),$$

we have

$$\frac{d\theta_i}{dH} = \theta_i^* \frac{d\kappa}{dH} + \kappa \frac{d\theta_i^*}{dH}.$$

If

$$\theta_i^* = \Theta_i^*(A(H), R(H), C_t),$$

then

$$\frac{d\theta_i^*}{dH} = \frac{\partial \Theta_i^*}{\partial A}\frac{dA}{dH} + \frac{\partial \Theta_i^*}{\partial R}\frac{dR}{dH}.$$

Thus even before considering motivation, an affective state can alter expression through several mathematically separable routes.

Adding activation produces the complete local decomposition:

$$p_i = \sigma(z_i), \qquad z_i = a_i - \kappa\theta_i^*,$$

so that

$$\frac{dp_i}{dH} = \sigma'(z_i)\left[\frac{da_i}{dH} - \theta_i^*\frac{d\kappa}{dH} - \kappa\frac{d\theta_i^*}{dH}\right].$$

Substituting the appraisal and regulatory pathways gives

$$\frac{dp_i}{dH} = \sigma'(z_i)\left[\frac{da_i}{dH} - \theta_i^*\frac{d\kappa}{dH} - \kappa\left(\frac{\partial \Theta_i^*}{\partial A}\frac{dA}{dH} + \frac{\partial \Theta_i^*}{\partial R}\frac{dR}{dH}\right)\right].$$

This is a more demanding theory than the claim that good mood causes disinhibition. The same observed increase in expression can be decomposed into a motivational effect, an appraisal-mediated permissive effect, a direct regulatory effect on normative thresholds, and an enforcement effect.

The distinction also prevents the threshold theory from simply renaming mood-as-information. If all apparent permission effects disappeared once appraisal were held constant, then the direct regulatory contribution would have little independent work to do. In the limiting case,

$$\beta_i = 0, \qquad \frac{d\kappa}{dH} = 0,$$

and the permissive effect of mood would be entirely mediated by appraisal:

$$H \to A \to \theta^*.$$

The threshold model would still provide a useful description of where that appraisal enters behavioral selection, but it would not have discovered an independent regulatory pathway.

Conversely, evidence for

$$\frac{\partial p_i}{\partial H} \neq 0$$

while the relevant appraisal variables are held approximately constant would support the stronger claim that affect participates directly in behavioral regulation.

In causal notation, the quantity of interest is something like

$$\frac{\partial}{\partial H}\mathbb{P}(E_i = 1 \mid do(A = a)),$$

rather than merely

$$\mathbb{P}(E_i = 1 \mid H).$$

The empirical challenge is therefore one of dissociation.

One experimental family could manipulate appraisal while changing affect as little as possible. Information indicating that a social environment is safe, forgiving, reversible, or non-punitive could alter

$$A_t$$

and therefore

$$\theta_i^*$$

without requiring a substantial change in hedonic state.

A second family would attempt the converse: alter affective or regulatory state while holding explicit beliefs about environmental safety comparatively stable. The cleaner the dissociation, the more informative the resulting comparison becomes.

The idealized factorial design would independently manipulate affect H and appraisal A:

$$(H_{\text{low}}, A_{\text{unsafe}}), \qquad (H_{\text{low}}, A_{\text{safe}}),$$

$$(H_{\text{high}}, A_{\text{unsafe}}), \qquad (H_{\text{high}}, A_{\text{safe}}).$$

If appraisal alone drives the relevant permission effect, then after conditioning on A,

$$E \perp H \mid A$$

should approximately hold.

If affect has an additional regulatory role, then

$$E \not\perp H \mid A.$$

The interaction term is also informative. Affect might not simply add a second independent pathway; it may alter how strongly appraisal affects thresholds. For example,

$$\theta_i^* = \theta_{i,0}^* - \alpha A - \beta H - \chi AH.$$

Then

$$\chi \neq 0$$

would mean that the behavioral significance of perceived safety itself depends on affective state.

This possibility is psychologically plausible. Knowing abstractly that a setting is safe need not make it *feel* safe enough for expression, while positive affect may make the same evidence behaviorally effective. Conversely, a sufficiently negative state may preserve high thresholds even when the person can explicitly report that no serious external threat exists.

The distinction between appraisal and regulation therefore need not map neatly onto a distinction between cognition and emotion. Appraisal can itself be affectively shaped, and regulation can depend on representations. The purpose of separating the pathways is causal rather than ontological: the model asks whether a state changes expression because it changes what consequences are expected, because it changes how those expectations regulate action, or both.

An Adlerian perspective again adds a useful complication. The relevant appraisal is not necessarily a neutral estimate of objective environmental risk. It occurs within an organized social orientation. Two people receiving the same evidence about the same audience may infer very different consequences because the situation occupies a different place within their characteristic expectations of belonging, rivalry, humiliation, cooperation, or competence.

We can represent this as

$$A_t = \mathcal{A}(H_t, C_t, \Lambda_t),$$

where Λ_t is the organized interpersonal orientation introduced earlier. Consequently,

$$\frac{\partial A}{\partial H}$$

may itself vary systematically with Λ.

Positive affect need not send the same informational signal through every person because the receiving interpretive system is not the same.

This helps explain why the variance prediction and the mood-as-information account are not competitors. A common affective shift can provide broadly similar safety information while producing different behavioral consequences because it encounters different latent fields, different suppression histories, and different interpersonal organizations.

For person p,

$$H \uparrow \longrightarrow A^{(p)} \longrightarrow \Delta\theta^{*(p)} \longrightarrow \mathcal{N}^{(p)},$$

while for person q,

$$H \uparrow \longrightarrow A^{(q)} \longrightarrow \Delta\theta^{*(q)} \longrightarrow \mathcal{N}^{(q)}.$$

Even if

$$\Delta\theta^{*(p)}_{\text{common}} \approx \Delta\theta^{*(q)}_{\text{common}},$$

there is no reason to expect

$$\mathcal{N}^{(p)} \approx \mathcal{N}^{(q)}.$$

The same apparent signal can open onto different behavioral terrain.

This gives the original observation a more precise interpretation. Elevated positive affect may increase warmth because it genuinely increases affiliative activation. It may increase candor because circumstances feel safer. It may increase criticism because anticipated social cost falls. It may alter self-monitoring or enforcement. Several of these processes may occur at once.

The fact that they converge on expression does not make them one mechanism.

Nor does the threshold model require deciding in advance which pathway is dominant. Its contribution is to preserve the distinctions long enough for that question to become empirically meaningful. Mood can inform judgment; it can alter regulation; it can change motivation. The resulting behavior is the product of all three, filtered through a history that determines what a particular person has learned is safe to reveal.

That history is rarely social in the abstract. It is accumulated with particular people. Repeated interactions can teach that criticism survives, that disclosure is not punished, that disagreement does not terminate a relationship, or that affection will not be rejected. Once thresholds are allowed to carry such partner-specific histories, trust becomes describable not merely as a feeling or belief but as a change in the geometry of what can be expressed.

10 The Social World as a Variable Threshold

Expression does not occur before an abstract audience. It occurs toward particular people, within particular relationships, under expectations built from previous encounters. A threshold that ignores the recipient therefore throws away one of the most important sources of variation in whether a behavior becomes visible.

Let j denote the person, group, or institution toward which behavior b_i would be expressed. The normative threshold should then be written

$$\theta_i^*(t,j) = \Theta_i^*\left(C_i(t,j), U_i(t,j), N_i(t,j), P_i(t,j), R_i(t,j), \mathcal{H}_{ij}(t)\right),$$

where $\mathcal{H}_{ij}(t)$ is the history relevant to behavior i in relation to recipient j. The effective threshold remains

$$\theta_i(t,j) = \kappa(t)\theta_i^*(t,j).$$

Expression is consequently recipient-relative:

$$\mathbb{P}(E_i(t,j)=1) = \sigma\left[a_i(t,j) - \kappa(t)\theta_i^*(t,j)\right].$$

The same person, with approximately the same disposition, can therefore satisfy

$$a_i(t,j_1) > \theta_i(t,j_1)$$

and

$$a_i(t,j_2) < \theta_i(t,j_2).$$

A criticism that is easy to express to a close friend may remain unspeakable to an employer. Affection that is effortless toward one family member may be difficult toward another. An eccentric

opinion may be expressed freely among trusted peers, cautiously among strangers, and not at all before an audience capable of imposing serious consequences.

None of these differences requires a corresponding change in the underlying disposition.

The simplest interpretation is that recipients change anticipated cost. If

$$C_i(t, j_1) < C_i(t, j_2),$$

then, all else equal,

$$\theta_i^*(t, j_1) < \theta_i^*(t, j_2).$$

But recipient dependence can enter through several terms simultaneously. With a familiar person, uncertainty may be lower,

$$U_i(t, j_1) < U_i(t, j_2),$$

reversibility may be higher,

$$R_i(t, j_1) > R_i(t, j_2),$$

and the expected severity of norm enforcement may be lower,

$$N_i(t, j_1) < N_i(t, j_2).$$

A relationship can therefore make expression easier without changing any single quantity dramatically. Several modest reductions can accumulate into a substantial movement of the threshold.

This gives a natural formal role to interaction history. Let

$$\mathcal{H}_{ij}(t) = \{E_i(s, j), o_i(s, j)\}_{s<t}$$

denote previous attempts to express behavior i toward j, together with their outcomes. A simple partner-specific update rule is

$$\theta_i^*(t+1, j) = \theta_i^*(t, j) - \gamma_+ r_{ij}^+(t) + \gamma_- r_{ij}^-(t).$$

Repeated low-cost expression then tends to produce

$$\theta_i^*(t+k, j) < \theta_i^*(t, j),$$

while costly outcomes can move the same threshold upward.

The resulting quantity is path-dependent. It matters not merely what the other person is like now, but what has happened when similar behavior has been expressed toward them before.

This suggests a restricted operational definition of trust. Let

$$T_{ij}(t)$$

represent accumulated evidence that expressing behavior i toward j does not produce consequences severe enough to justify continued suppression. One possible bounded update is

$$T_{ij}(t+1) = (1-\lambda)T_{ij}(t) + \lambda q_{ij}(t), \qquad 0 < \lambda \leq 1,$$

where $q_{ij}(t) \in [0,1]$ summarizes the safety of the most recent relevant interaction. Then

$$0 \leq T_{ij}(t) \leq 1.$$

The threshold can depend on this accumulated estimate through

$$\theta_i^*(t,j) = \bar{\theta}_i^*(t,j) - \gamma_i T_{ij}(t), \qquad \gamma_i > 0.$$

As

$$T_{ij} \uparrow,$$

the threshold tends to fall:

$$\frac{\partial \theta_i^*}{\partial T_{ij}} = -\gamma_i < 0.$$

This is not intended as a complete theory of trust. Trust includes expectations about competence, loyalty, honesty, care, predictability, and much else. The quantity T_{ij} isolates only one dimension relevant to the present argument: learned permission to expose portions of the behavioral field.

Even this narrow component need not be scalar. A person may trust another with vulnerability but not money, with disagreement but not secrets, with humor but not professional risk. A more realistic representation is therefore a vector

$$\mathbf{T}_j = (T_{1j}, T_{2j}, \ldots, T_{nj}),$$

or, after grouping related behaviors,

$$\mathbf{T}_j = \left(T_{\text{affection}}, T_{\text{dissent}}, T_{\text{disclosure}}, T_{\text{dependence}}, T_{\text{criticism}}, \ldots\right)_j.$$

Trust in this sense is not a scalar quantity distributed uniformly over a relationship. It has a topology.

The corresponding threshold vector is

$$\boldsymbol{\theta}^*(t,j) = (\theta_1^*(t,j), \ldots, \theta_n^*(t,j)),$$

and two relationships can have very different geometries even if both would ordinarily be described as "high trust."

For example,

$$T_{\text{affection},j} \gg 0, \qquad T_{\text{dissent},j} \ll 1$$

describes a relationship in which warmth is easy but disagreement remains dangerous. Another may satisfy

$$T_{\text{affection},k} \approx T_{\text{dissent},k} \approx 1,$$

making both affirmation and criticism comparatively easy to express.

This gives a more precise interpretation of the observation with which the essay began. Becoming simultaneously warmer and more critical need not represent a contradiction. Within some relationships, those behaviors may become jointly permissible because the relationship has reached a state in which neither affection nor disagreement carries much threat.

Indeed, one of the characteristic consequences of sufficiently robust trust may be that apparently opposite behaviors cease to function as opposites.

If criticism no longer implies rejection, then

$$C_{\text{criticism}} \downarrow .$$

If affection no longer implies dangerous vulnerability, then

$$C_{\text{affection}} \downarrow .$$

Both thresholds can fall:

$$\theta^*_{\text{criticism}} \downarrow, \qquad \theta^*_{\text{affection}} \downarrow .$$

The resulting relationship can become simultaneously more affectionate and more argumentative without either movement cancelling the other.

This is compatible with the Adlerian qualification developed earlier. Relationships do not merely alter a set of external costs imposed on an otherwise fixed individual. They can participate in the organization of behavior itself. If

$$\Lambda_t$$

represents the person's characteristic organization of social purposes and expectations, then recipient j can enter both sides of the expression inequality:

$$a_i(t, j) = A_i(D, \Lambda_t, C_t, j)$$

and

$$\theta^*_i(t, j) = \Theta^*_i(C_t, \mathcal{H}_{ij}, j).$$

The recipient can change what approaches the gate as well as the height of the gate.

This matters because a simplistic threshold theory might otherwise interpret every increase in expression under trust as the revelation of something that was already present in exactly the same form. Sometimes that will be approximately correct. In other cases the relationship itself has helped produce the disposition being expressed.

A long friendship can lower the threshold for a kind of humor that already existed, but it can also generate a repertoire of humor specific to that friendship. Intimacy can make an existing vulnerability expressible, but it can also create forms of vulnerability that would not have existed outside the relationship. Repeated disagreement can make dissent safer, while also making the participants better at producing useful disagreement.

Thus

$$\mathcal{H}_{ij} \to \theta_i^*$$

is only one pathway. There may also be

$$\mathcal{H}_{ij} \to \Lambda_t \to a_i.$$

The social world does not merely censor a finished person. It participates in the person's continuing organization.

Nevertheless, the threshold distinction remains useful because changes in permission can occur much faster than changes in disposition. A person entering a familiar room can become visibly different within seconds. It would be implausible to suppose that their personality was reconstructed at the doorway. More often, a previously learned relationship-specific threshold has been reinstated:

$$j_{\text{trusted}} \quad \Longrightarrow \quad \theta_i^*(t, j_{\text{trusted}}) \downarrow .$$

This makes context switching mathematically significant. Let the same activation satisfy

$$a_i = a_i(t)$$

over a short interval while the audience changes from j_1 to j_2. Then

$$\Delta z_i = -\kappa \left[\theta_i^*(t, j_2) - \theta_i^*(t, j_1)\right].$$

A behavioral change can occur with

$$\Delta a_i \approx 0.$$

The audience alone can move the expression margin across zero.

Power relations make this especially consequential. Suppose P_{ij} represents the capacity of recipient j to impose costs on person p. Then

$$\frac{\partial \theta_i^*}{\partial P_{ij}} > 0$$

for behaviors whose expression risks retaliation. As power asymmetry grows,

$$P_{ij} \uparrow \quad \Longrightarrow \quad \theta_i^* \uparrow .$$

Visible agreement can therefore increase while actual agreement remains unchanged.

Let a_{dissent} denote activation of dissent. Then

$$\mathbb{P}(E_{\text{dissent}} = 1) = \sigma \left[a_{\text{dissent}} - \kappa \theta_{\text{dissent}}^*\right].$$

Increasing the cost of dissent drives

$$\theta_{\text{dissent}}^* \uparrow$$

and therefore

$$\mathbb{P}(E_{\text{dissent}} = 1) \downarrow$$

even if

$$a_{\text{dissent}}$$

is constant or rising.

The observable absence of disagreement consequently does not identify the absence of disagreement as a latent state.

This is the social version of the inverse problem established earlier. If

$$Y_{\text{dissent}} = 0,$$

one cannot infer

$$a_{\text{dissent}} \approx 0.$$

One knows only that

$$a_{\text{dissent}} \leq \kappa \theta^*_{\text{dissent}}.$$

A harmonious group and a frightened group can therefore generate the same surface observation.

Conversely, a relationship in which dissent is extremely cheap can display frequent disagreement even when the underlying magnitude of disagreement is modest. Suppose safety lowers the threshold:

$$C_{\text{social}} \downarrow \quad \Longrightarrow \quad \theta^*_{\text{dissent}} \downarrow .$$

Then increasingly small dissent activations can cross it.

The observable quantity

$$Y_{\text{dissent}}$$

may rise while the latent quantity

$$a_{\text{dissent}}$$

remains constant or even falls.

This produces a correlation reversal that is easy to misread socially. Across some range of conditions, greater safety can be associated with more visible conflict:

$$\frac{d\,\mathbb{P}(E_{\text{dissent}} = 1)}{d\,\text{Safety}} > 0.$$

That inequality does not mean safety creates hostility. It can mean safety makes hostility, disagreement, correction, annoyance, and ordinary difference less expensive to reveal.

At the same time, the model does not license the opposite shortcut that visible conflict is evidence of safety. High conflict can arise from high underlying hostility, low expression thresholds, strong activation, weak enforcement, or some combination of these. The mapping remains many-to-one.

The narrower conclusion is more useful:

expressed conflict is not a monotonic measure of social hostility.

Once expression thresholds are recipient-relative and historically learned, quiet and conflict both become ambiguous observations. A silent relationship may contain little disagreement or a great deal that cannot safely cross the gate. A noisy relationship may be unstable, or it may have become sufficiently secure that small disagreements no longer threaten anything important.

The difference cannot be read from the volume of disagreement alone. It depends on what disagreement costs.

11 Why Safety Can Produce More Conflict

Once expression thresholds are allowed to vary with social cost, a familiar interpretive problem becomes mathematically unavoidable. The amount of conflict that can be observed in a relationship, institution, or society is not simply the amount of conflict that exists within it. It is the amount of latent disagreement that succeeds in crossing the prevailing thresholds of expression.

Let g denote the prevalence or intensity of some underlying disagreement and let

$$e = \mathbb{P}(\text{expression} \mid \text{disagreement})$$

denote the probability that an existing disagreement becomes visible. A first approximation to the observable conflict rate is then

$$c = ge.$$

This factorization is deliberately simple. It separates two quantities that ordinary interpretation often treats as one. The first asks how much disagreement exists. The second asks how much of that disagreement can be expressed.

Within the threshold model,

$$e = \sigma\left(a_{\text{dissent}} - \kappa\theta^*_{\text{dissent}}\right),$$

so that

$$c = g\,\sigma\left(a_{\text{dissent}} - \kappa\theta^*_{\text{dissent}}\right).$$

A rise in observed conflict can therefore be generated by an increase in underlying disagreement,

$$g \uparrow,$$

by an increase in its expressibility,

$$e \uparrow,$$

or by both.

The distinction becomes especially important when safety changes. Let s denote some relevant dimension of interpersonal or institutional safety. If greater safety lowers the expected cost of dissent, then

$$\frac{\partial \theta^*_{\text{dissent}}}{\partial s} < 0.$$

Holding activation and enforcement fixed,

$$\frac{\partial e}{\partial s} = -\sigma'(z)\kappa \frac{\partial \theta^*_{\text{dissent}}}{\partial s} > 0,$$

where

$$z = a_{\text{dissent}} - \kappa \theta^*_{\text{dissent}}.$$

Thus safer conditions can increase the probability that an existing disagreement becomes observable.

If underlying disagreement remains constant,

$$\frac{dg}{ds} = 0,$$

then

$$\frac{dc}{ds} = g\frac{de}{ds} > 0.$$

Visible conflict rises even though latent conflict has not.

More strikingly, visible conflict can rise while latent conflict falls. Since

$$c = ge,$$

differentiation with respect to safety gives

$$\frac{dc}{ds} = e\frac{dg}{ds} + g\frac{de}{ds}.$$

Suppose improved conditions reduce underlying disagreement:

$$\frac{dg}{ds} < 0,$$

while simultaneously making the remaining disagreement easier to express:

$$\frac{de}{ds} > 0.$$

Then

$$\frac{dc}{ds} > 0$$

whenever

$$g\frac{de}{ds} > -e\frac{dg}{ds}.$$

The increase in expressibility can exceed the decline in disagreement itself.

This is the central correlation reversal. A system can become less hostile while appearing more argumentative.

The result becomes particularly transparent in proportional terms. For $g, e, c > 0$,

$$\frac{\dot{c}}{c} = \frac{\dot{g}}{g} + \frac{\dot{e}}{e}.$$

Observable conflict grows whenever

$$\frac{\dot{e}}{e} > -\frac{\dot{g}}{g}.$$

A ten percent decline in latent disagreement accompanied by a twenty percent increase in its probability of expression can therefore produce more visible conflict, not less.

This is not a paradox in the generative system. It becomes paradoxical only when

$$c$$

is treated as though it directly measured

$$g.$$

The error is analogous to interpreting the number of reported events as a direct measure of the number of underlying events while ignoring changes in reporting probability. Once reporting itself varies, the observable cannot be read without a model of the observation process.

The same structure appears at the individual level. Suppose two people have some disagreement activation

$$a_{\text{dissent}} > 0.$$

In a fragile relationship,

$$\theta^*_{\text{dissent}} \gg a_{\text{dissent}},$$

and the disagreement remains invisible. As repeated interactions demonstrate that dissent does not cause abandonment, humiliation, retaliation, or lasting damage,

$$C_{\text{dissent}} \downarrow, \qquad U_{\text{dissent}} \downarrow, \qquad R_{\text{dissent}} \uparrow,$$

and therefore

$$\theta^*_{\text{dissent}} \downarrow .$$

Eventually,

$$a_{\text{dissent}} > \kappa\theta^*_{\text{dissent}},$$

and the disagreement becomes expressible.

From outside, the relationship has acquired conflict.

From inside, it may have acquired permission.

This difference is not semantic. The two interpretations imply different dynamics. If increased visible disagreement reflects

$$g \uparrow,$$

then one expects deterioration in other indicators associated with the underlying conflict. If it reflects primarily

$$e \uparrow,$$

then the increased disagreement may coexist with greater intimacy, lower fear, faster repair, greater willingness to disclose mistakes, and greater confidence that the relationship will survive the exchange.

This suggests that conflict should not be studied only by frequency. Its *consequence structure* also matters.

Let L denote the expected loss associated with an episode of expressed disagreement. Two systems can satisfy

$$c_1 < c_2$$

while simultaneously satisfying

$$L_1 \gg L_2.$$

The first system has fewer visible conflicts but each carries substantial risk; the second has more visible conflicts precisely because individual conflicts have become comparatively cheap.

A crude measure of conflict burden might therefore be

$$B_c = cL,$$

rather than c alone. Even this remains incomplete, but it immediately shows how

$$c \uparrow$$

can coexist with

$$B_c \downarrow$$

when the cost per conflict falls sufficiently rapidly:

$$\frac{\dot{L}}{L} < -\frac{\dot{c}}{c}.$$

A relationship can become noisier and less dangerous at the same time.

This also clarifies why harmony is difficult to infer from silence. Let

$$c \approx 0.$$

Since

$$c = ge,$$

the observation is compatible with

$$g \approx 0,$$

which would represent genuine low disagreement, but also with

$$e \approx 0,$$

which would represent severe suppression.

At the limiting extremes,

$$g \to 0, \qquad e \to 1$$

and

$$g \to 1, \qquad e \to 0$$

can both yield

$$c \to 0.$$

One system is quiet because almost nobody objects. The other is quiet because almost nobody can.

This is a structural non-identifiability problem. Given only

$$c = ge,$$

there are infinitely many pairs (g, e) compatible with the same observation. For any $c \in (0, 1)$,

$$(g, e) = \left(x, \frac{c}{x}\right)$$

produces the same c for every admissible

$$x \in [c, 1].$$

No measurement of visible conflict alone can recover the underlying disagreement rate.

Power asymmetry makes the ambiguity especially important. Suppose the cost of dissent depends positively on power differential P:

$$\frac{\partial \theta^*_{\text{dissent}}}{\partial P} > 0.$$

Then

$$P \uparrow \implies e \downarrow .$$

A highly asymmetric organization can therefore exhibit unusually little visible disagreement without possessing unusually little latent disagreement. Its apparent consensus may partly be an observation effect generated by the cost of contradicting whoever occupies the stronger position.

Conversely, deliberately protecting dissent can increase measured conflict. Suppose an institution reduces retaliation risk,

$$C_{\text{dissent}} \downarrow,$$

explicitly weakens norms against disagreement,

$$N_{\text{dissent}} \downarrow,$$

and increases recoverability after mistaken or unpopular criticism,

$$R_{\text{dissent}} \uparrow .$$

Then

$$\theta^*_{\text{dissent}} \downarrow$$

and

$$e \uparrow .$$

If the reform succeeds, the immediate empirical result may be an increase in reported disagreement.

Judging the reform solely by

$$c$$

could therefore reverse its interpretation. The mechanism intended to make disagreement safer has made disagreement more visible, and that visibility is then mistaken for evidence that the institution has become more conflictual.

The same problem appears in measurements of psychological and social safety. If a survey asks whether people have recently voiced disagreement, reported a problem, challenged a decision, admitted uncertainty, or disclosed a mistake, high rates can indicate that many problems exist. They can also indicate that the threshold for reporting them is low.

Observed frequency is produced jointly by incidence and permeability.

A more general formulation is

$$O_i = L_i E_i,$$

where L_i denotes the latent occurrence of condition i and E_i its probability of becoming observable. Then

$$\dot{O}_i = E_i \dot{L}_i + L_i \dot{E}_i.$$

The structure is not peculiar to disagreement. Whenever the probability of expression changes with the environment, changes in the observable can have the opposite sign from changes in the latent condition.

For social conflict,

$$\boxed{\dot{c} = e\dot{g} + g\dot{e}.}$$

This equation will become important again when the argument is connected to *The North Setpoint*. There the first term will acquire a richer internal structure because what counts as a grievance can itself change as conditions improve. For the moment, however, the simpler result is enough: even if the production of grievances were held completely constant, lowering the threshold of expression would be sufficient to change the amount of complaint that becomes visible.

The Adlerian perspective adds one further qualification. Low-cost disagreement need not merely expose a fixed quantity g. Repeated safe disagreement can change the organization of the relationship itself. If dissent is expressed, survives, and is repaired, then the outcome enters the history

$$\mathcal{H}_{ij}(t),$$

which can lower future thresholds:

$$\theta^*_{\text{dissent}}(t+1, j) < \theta^*_{\text{dissent}}(t, j).$$

But the same history may also alter the participants' social organization Λ. They can become better at distinguishing disagreement from rejection, criticism from hostility, and correction from domination. The relationship learns not merely that conflict is permissible but how conflict fits within continued cooperation.

Thus the feedback can take the form

$$\text{safe dissent} \to \text{successful continuation} \to \theta^*_{\text{dissent}} \downarrow \to \text{more expressible dissent},$$

while a slower process modifies

$$\Lambda_t \to \Lambda_{t+1}.$$

More visible disagreement can then become part of the mechanism by which a relationship becomes more robust.

The converse loop is equally possible:

$$\text{dissent} \to \text{retaliation} \to \theta^*_{\text{dissent}} \uparrow \to \text{less visible dissent}.$$

Such a system can become quieter as it becomes less safe.

This is why neither silence nor conflict has a fixed social meaning. Their meaning depends on the latent field, the expression threshold, the history that produced that threshold, and what happens after expression occurs.

The principle can therefore be stated more carefully than "safe groups argue more." They need not. A safe group with very little underlying disagreement may remain quiet. Nor does abundant conflict demonstrate safety. A hostile group can obviously be both unsafe and argumentative.

What follows from the model is only the non-monotonicity:

expressed conflict is not a monotonic measure of underlying hostility.

In some regimes, increasing safety can reduce conflict. In others it can increase visible conflict by making disagreement cheaper to reveal. Without an independent estimate of the expression process, the sign of the observable alone cannot tell us which regime we are seeing.

That same identification problem becomes still sharper when the latent quantity is not fixed. A person or society can change not only how easily a grievance is spoken, but what discrepancy becomes large enough to register as a grievance in the first place. At that point two moving boundaries intervene between conditions and complaint: one determines what becomes objectionable, and another determines what becomes speakable.

12 The Bridge to *The North Setpoint*

The preceding argument separates latent disagreement from its probability of expression. That separation becomes more important once the latent quantity is allowed to change with the conditions under which it is evaluated. A complaint does not arise merely because some objectively measurable harm exists and then passes through an expression threshold. Before expression becomes relevant, some feature of experience must first register as a discrepancy worth objecting to.

This introduces a second threshold-like process upstream of the one developed in this essay.

The companion essay, *The North Setpoint*, considers the adaptation of the reference frame against which conditions are evaluated. Its central quantity is a locally adapted reference point

$$N_t = \mathbb{E}[H]_{t-\tau:t},$$

where H represents experienced conditions along some relevant hedonic or evaluative dimension and τ specifies the adaptation window. Sustained improvement can produce

$$N_{t+1} > N_t.$$

The important consequence is not simply that people "get used to" improved conditions. It is that the location from which deviations are evaluated has changed.

A condition H_i can therefore acquire significance partly through its distance from the adapted reference:

$$d_i(t) = N_t - H_i(t).$$

The probability that this discrepancy becomes a grievance can be represented abstractly as

$$g_i(t) = \mathbb{P}(G_i = 1) = \Gamma\left(d_i(t), H_i(t), X_i(t)\right),$$

where X_i collects whatever additional variables affect evaluation. The companion essay is concerned primarily with the structure of Γ, the movement of N_t, and the possibility that relative discrimination persists even as absolute suffering declines.

The present essay begins one stage later.

Conditional on a grievance having formed, the question here is whether it becomes behaviorally visible. Let

$$e_i(t) = \mathbb{P}(E_i = 1 \mid G_i = 1).$$

Then the probability of an observable complaint is

$$c_i(t) = \mathbb{P}(C_i = 1) = \mathbb{P}(G_i = 1)\mathbb{P}(E_i = 1 \mid G_i = 1),$$

or simply

$$\boxed{c_i = g_i e_i.}$$

This small equation marks the division of labor between the two essays.

The North Setpoint studies g_i.

What Affect Permits studies e_i.

Neither factor can generally be inferred from their product.

The full pipeline is therefore

experienced conditions $\longrightarrow$ reference adaptation $\longrightarrow$ discrepancy $\longrightarrow$ grievance $\longrightarrow$ expression threshold $\longrightarrow$ c

Writing the stages separately prevents several common explanatory collapses. A bad condition is not identical to a grievance about that condition. A grievance is not identical to its expression. An expressed complaint is not a direct measurement of the condition that ultimately contributed to producing it.

There are at least two filters between condition and observable complaint.

The first is evaluative:

$$H_i, N_t \longrightarrow g_i.$$

The second is expressive:

$$g_i, \theta_i \longrightarrow e_i.$$

The first asks,

> How large does a discrepancy have to become before it registers as something wrong?

The second asks,

Once something registers as wrong, what must happen before the objection can be expressed?

These questions are independent enough that neither answer determines the other.

A person can experience a substantial grievance while maintaining a very high expression threshold:

$$g_i \approx 1, \qquad e_i \approx 0.$$

The resulting complaint rate is low:

$$c_i \approx 0.$$

This is silent dissatisfaction.

Another person or population can have both a substantial grievance rate and a low threshold for expression:

$$g_i \approx 1, \qquad e_i \approx 1,$$

giving

$$c_i \approx 1.$$

This is visible dissatisfaction.

A third case has few grievances but strong inhibition:

$$g_i \approx 0, \qquad e_i \approx 0.$$

Again,

$$c_i \approx 0.$$

The observable resembles the first case despite an almost opposite latent condition.

Finally, grievance can be rare while expression is highly permissible:

$$g_i \approx 0, \qquad e_i \approx 1,$$

which also produces

$$c_i \approx 0.$$

Here the silence is not produced by suppression. The gate is open; comparatively little is trying to pass through it.

The four limiting regimes can be written compactly as

	e low	e high
g low	low grievance, inhibited	low grievance, permissive
g high	silent dissatisfaction	visible dissatisfaction

and three of the four can generate relatively little observable complaint.

This is a stronger identification problem than the one encountered in the previous section. There, visible conflict could not identify latent disagreement because the expression probability was unknown. Here even the production of the latent grievance is itself conditioned by an adaptive reference process.

Suppose only the observable complaint rate c is known. Since

$$c = ge,$$

for any $c \in (0, 1)$ there is a continuum of admissible decompositions

$$(g, e) = \left(x, \frac{c}{x}\right), \qquad x \in [c, 1].$$

Thus even perfect measurement of complaint frequency cannot identify grievance frequency without additional information about expression probability.

Nor can grievance frequency by itself identify absolute conditions, because

$$g = \Gamma(N - H, H, X)$$

and N is adaptive.

The inverse problem therefore has two stages:

$$c \not\Rightarrow g \not\Rightarrow H.$$

Observed complaint does not uniquely recover grievance, and grievance does not uniquely recover absolute suffering.

This does not make complaints uninformative. It specifies what additional information is needed before they can be interpreted.

The temporal version makes the relationship between the two essays especially clear. Differentiating

$$c(t) = g(t)e(t)$$

gives

$$\boxed{\dot{c} = e\dot{g} + g\dot{e}.}$$

The first term,

$$e\dot{g},$$

is the contribution of changing grievance production to the observable complaint rate.

The second,

$$g\dot{e},$$

is the contribution of changing expressibility.

This equation is the natural mathematical hinge between the two theories.
The North Setpoint provides one mechanism capable of producing

$$\dot{g} > 0$$

without requiring

$$\dot{H} < 0.$$

What Affect Permits provides mechanisms capable of producing

$$\dot{e} > 0$$

without requiring any increase in the underlying grievance.
The two effects can occur independently.
Consider first reference adaptation alone. Suppose absolute conditions improve,

$$\dot{H} > 0,$$

but the local reference point rises as well,

$$\dot{N} > 0.$$

Since

$$d = N - H,$$

we have

$$\dot{d} = \dot{N} - \dot{H}.$$

If the reference point rises faster than the particular condition being evaluated,

$$\dot{N} > \dot{H},$$

then

$$\dot{d} > 0$$

despite absolute improvement.
If grievance probability increases with discrepancy,

$$\frac{\partial g}{\partial d} > 0,$$

then this pathway can yield

$$\dot{g} > 0$$

while

$$\dot{H} > 0.$$

Now consider expression independently. Suppose the environment becomes safer for complaint:

$$C_{\text{social}} \downarrow, \qquad U \downarrow, \qquad R \uparrow .$$

Then

$$\theta^*_{\text{complaint}} \downarrow$$

and therefore

$$\dot{e} > 0.$$

Combining the two gives

$$\dot{H} > 0, \qquad \dot{N} > 0, \qquad \dot{\theta}^*_{\text{complaint}} < 0.$$

Under suitable magnitudes,

$$\dot{g} > 0$$

and

$$\dot{e} > 0,$$

so

$$\dot{c} = e\dot{g} + g\dot{e} > 0.$$

Absolute conditions improve while visible complaint increases.

Nothing inconsistent has occurred.

The evaluative reference moved north while the expressive gate opened.

This is precisely why the two mechanisms should not be collapsed into a single theory of dissatisfaction. The movement of N changes what enters the grievance population. The movement of θ changes how much of that population becomes visible.

In differential form, complaint responds to both:

$$dc = \frac{\partial c}{\partial N}\, dN + \frac{\partial c}{\partial H}\, dH + \frac{\partial c}{\partial \theta}\, d\theta + \cdots .$$

Because

$$c = ge,$$

these derivatives factor naturally:

$$\frac{\partial c}{\partial N} = e\frac{\partial g}{\partial N},$$

$$\frac{\partial c}{\partial H} = e\frac{\partial g}{\partial H},$$

and

$$\frac{\partial c}{\partial \theta} = g\frac{\partial e}{\partial \theta}.$$

If

$$\frac{\partial g}{\partial N} > 0$$

and

$$\frac{\partial e}{\partial \theta} < 0,$$

then

$$dN > 0 \qquad \text{and} \qquad d\theta < 0$$

both push complaint upward even though they operate at different stages of the generative process.

This also clarifies an apparent tension between the essays' subjects. The North Setpoint is not claiming that people complain more simply because they are safer. What Affect Permits is not claiming that safer conditions generate more grievances. One theory concerns the reference frame under which discrepancies acquire evaluative significance; the other concerns the permeability of the boundary between private evaluation and public behavior.

The distinction matters particularly when the same social improvement affects both.

Suppose some institutional change reduces a class of severe harms. This may raise the ordinary baseline against which remaining discrepancies are judged:

$$N \uparrow .$$

At the same time, the institution may become more tolerant of criticism:

$$\theta^*_{\text{complaint}} \downarrow .$$

The first transformation can make finer discrepancies increasingly salient. The second can make those discrepancies increasingly speakable.

The observable record can then contain more criticism after the improvement than before it.

Reading that increase as direct evidence of worsening conditions commits an inverse error:

$$c \uparrow \not\Rightarrow \; g \uparrow \not\Rightarrow \; H \downarrow .$$

But the opposite inference is equally invalid:

$$c \uparrow \not\Rightarrow \; H \uparrow .$$

Abundant complaint does not prove that a society is safe, prosperous, or improving. Severe underlying harms can obviously produce abundant grievances, and sufficiently activated grievances can cross even high thresholds.

The theory therefore supplies no shortcut from complaint frequency to social diagnosis.

Its claim is about non-identifiability.

To infer the underlying condition from the observable, one needs independent estimates of at least some of the intermediate variables:

$$H, \qquad N, \qquad g, \qquad \theta, \qquad e.$$

Without them, several radically different social worlds can project onto the same complaint rate.

There is also an important asymmetry between the two moving boundaries. The reference point N determines what becomes evaluatively salient; the expression threshold θ determines what becomes socially observable. A person can therefore know that something bothers them while refusing to say it:

$$G_i = 1, \qquad E_i = 0.$$

But there are also discrepancies that never acquire the status of grievance in the first place:

$$d_i \neq 0, \qquad G_i = 0.$$

These are not suppressed complaints. They are differences that have not become objections.

That distinction prevents the threshold theory from expanding until it explains everything. Not every unexpressed possibility is a censored grievance. The gate studied here acts only after the relevant activation has formed.

Schematically,

$$H_i \to d_i \to G_i \to a_{\text{complaint},i} \to \theta_{\text{complaint},i} \to E_i.$$

The upstream theory asks why

$$d_i \to G_i.$$

The present theory asks why

$$G_i \not\Rightarrow E_i.$$

Keeping those questions separate also prevents a subtle moral confusion. If someone has a low threshold for expressing grievances, that fact says nothing by itself about whether those grievances are justified. Likewise, explaining why a discrepancy became salient does not determine whether expressing it is wise, proportionate, socially useful, or morally appropriate.

The equations describe the production and visibility of complaint. They do not adjudicate its content.

This is especially important because the same machinery can operate on excellent complaints and terrible ones. A lowered expression threshold can make whistleblowing possible; it can make petty hostility easier; it can make an apology possible; it can release prejudice; it can permit a needed correction; it can permit pointless cruelty. The threshold itself does not settle the value of what crosses it.

Likewise, reference adaptation can increase sensitivity to genuine residual injustice or to trivial deviations. The fact that a discrepancy is relative does not make it unreal, and the fact that it is experienced strongly does not make it large in every other coordinate.

The two essays therefore converge on a methodological principle rather than a verdict:

the frequency of complaint is not a monotonic measure of the magnitude of suffering.

The result follows because complaint is generated through at least two moving processes. What becomes grievance-worthy can change, and what becomes speakable can change.

Once those processes are separated, the otherwise puzzling observation that better conditions sometimes coexist with more visible dissatisfaction ceases to require either cynicism or denial. It becomes a question about which term in

$$\dot{c} = e\dot{g} + g\dot{e}$$

is moving, by how much, and why.

The next step is therefore not to decide whether abundant complaint signifies progress or decline. It is to examine the more interesting case in which both terms move together: the reference frame rises while the cost of expression falls. That combination makes it possible for increasingly favorable conditions to generate an increasingly dense record of increasingly small objections.

13 Why More Complaint Can Accompany Better Conditions

The two preceding mechanisms can now be allowed to move simultaneously. A change in conditions can alter what becomes grievance-worthy while also altering the cost of expressing whatever grievances remain. Once both boundaries move, the apparent relation between suffering and complaint can become surprisingly weak.

Recall the basic decomposition

$$c(t) = g(t)e(t),$$

where $g(t)$ is the probability that a relevant condition becomes a grievance and $e(t)$ is the probability that an existing grievance becomes expressed. Differentiation gives

$$\boxed{\dot{c} = e\dot{g} + g\dot{e}.}$$

This identity is simple enough that its implications are easy to overlook. Complaint can increase because grievances become more frequent, because expression becomes easier, or because both processes occur together. None of these possibilities, considered by itself, determines whether absolute conditions have improved or deteriorated.

Let H_t denote the relevant absolute condition, oriented so that

$$\dot{H}_t > 0$$

means improvement. Let N_t denote the locally adapted reference point introduced in *The North Setpoint*. A simple discrepancy variable is

$$d_t = N_t - H_t.$$

If grievance probability is increasing in discrepancy,

$$g_t = \Gamma(d_t), \qquad \Gamma'(d_t) > 0,$$

then

$$\dot{g}_t = \Gamma'(d_t)\left(\dot{N}_t - \dot{H}_t\right).$$

This immediately separates absolute improvement from relative evaluation.

If

$$\dot{H}_t > 0$$

but

$$\dot{N}_t > \dot{H}_t,$$

then

$$\dot{d}_t > 0$$

and consequently

$$\dot{g}_t > 0.$$

Conditions have improved, yet the discrepancy from the newly adapted reference point has increased.

This is not the only possible outcome. If

$$\dot{H}_t > \dot{N}_t,$$

then

$$\dot{g}_t < 0.$$

The model therefore does not require grievances to proliferate under improving conditions. It says only that the sign of

$$\dot{g}$$

cannot be inferred from the sign of

$$\dot{H}$$

without knowing how the reference point is moving.

The expression side introduces a second independent derivative. Write

$$e_t = \sigma\left[a_t - \kappa_t \theta_t^*\right].$$

Then

$$\dot{e}_t = \sigma'(z_t)\left[\dot{a}_t - \theta_t^* \dot{\kappa}_t - \kappa_t \dot{\theta}_t^*\right],$$

where

$$z_t = a_t - \kappa_t \theta_t^*.$$

Suppose improving social conditions reduce the anticipated cost of expressing a complaint. Perhaps retaliation becomes less likely, criticism becomes more institutionally protected, relationships become more secure, or recovery from an unsuccessful objection becomes easier. These changes can produce

$$\dot{\theta}_t^* < 0,$$

and therefore, holding the other terms approximately constant,

$$\dot{e}_t > 0.$$

The combined system can consequently satisfy

$$\dot{H}_t > 0, \qquad \dot{g}_t > 0, \qquad \dot{e}_t > 0,$$

which necessarily gives

$$\dot{c}_t > 0.$$

The important case, however, is even stronger. Complaint can increase while grievance itself decreases.

Suppose

$$\dot{H}_t > 0, \qquad \dot{g}_t < 0, \qquad \dot{e}_t > 0.$$

Then

$$\dot{c}_t > 0$$

whenever

$$g\dot{e} > -e\dot{g}.$$

Equivalently,

$$\frac{\dot{e}}{e} > -\frac{\dot{g}}{g}.$$

The rate at which grievances become easier to express need only exceed the rate at which grievances themselves are disappearing.

This produces perhaps the cleanest version of the apparent paradox:

$$\boxed{\dot{H} > 0, \qquad \dot{g} < 0, \qquad \dot{c} > 0}$$

is entirely possible.

Absolute conditions improve. Fewer grievances exist. More complaints are heard.

The missing variable is permeability.

Consider a simple numerical example. Suppose that in an earlier environment, half of a population experiences some class of grievance,

$$g_0 = 0.50,$$

but only one tenth of those grievances are expressed:

$$e_0 = 0.10.$$

The observable complaint rate is

$$c_0 = g_0 e_0 = 0.05.$$

Now suppose conditions improve substantially and grievance prevalence falls to

$$g_1 = 0.30.$$

At the same time, expression becomes much safer, raising

$$e_1 = 0.40.$$

Then

$$c_1 = g_1 e_1 = 0.12.$$

The underlying grievance rate has fallen by forty percent, while the observable complaint rate has more than doubled.

Nothing about the increase

$$0.05 \to 0.12$$

reveals the improvement

$$0.50 \to 0.30$$

unless the change in expression probability is independently known.

A historical observer with access only to public complaints could therefore rank the two environments backward.

The same inversion can occur without any change in the total number of grievances if their severity distribution changes. Let $s \geq 0$ denote the absolute severity of a grievance and let

$$f_t(s)$$

denote its distribution at time t. The expected absolute grievance burden is

$$B_t = \int_0^\infty s f_t(s)\, ds.$$

Suppose social improvement shifts probability mass toward lower severities, so

$$B_{t+1} < B_t.$$

Expression probability, however, may itself depend on severity:

$$e_t(s) = \mathbb{P}(E = 1 \mid s).$$

The observable complaint rate is then

$$c_t = \int_0^\infty e_t(s) f_t(s)\, ds.$$

If expression thresholds fall sufficiently,

$$e_{t+1}(s) > e_t(s)$$

over a broad range of s, it is possible to have

$$B_{t+1} < B_t$$

while

$$c_{t+1} > c_t.$$

The society is producing a larger number of visible complaints about a smaller aggregate burden.

This is particularly likely when the threshold reduction exposes the lower tail of the severity distribution. Under high expression costs, perhaps only very severe grievances are worth voicing:

$$E = 1 \quad \text{primarily when} \quad s > s_{\text{high}}.$$

Under safer conditions,

$$s_{\text{low}} < s_{\text{high}},$$

and grievances satisfying

$$s_{\text{low}} < s \leq s_{\text{high}}$$

enter the observable record.

The resulting increase in complaint is partly a measurement-resolution effect. The social instrument has become sensitive to smaller signals.

This suggests an analogy with improvements in scientific instrumentation. Suppose a detector originally registers only events exceeding magnitude m_0. After technical improvement it registers events down to

$$m_1 < m_0.$$

The number of recorded events can increase dramatically even while the total energy represented by those events decreases. It would be a mistake to infer that the underlying system had necessarily become more energetic simply because the detector became more sensitive.

Expression thresholds can behave similarly.

A social environment with

$$\theta_{\text{complaint}} \downarrow$$

has increased the resolution with which latent dissatisfaction becomes publicly measurable.

This does not make complaints mere measurement artifacts. The newly observed events were real in the relevant sense before the detector became capable of registering them. The point is instead that the measurement process changed.

The same qualification applies to historical comparison. Suppose one period contains relatively severe conditions but high expression thresholds,

$$H_{\text{old}} \ll H_{\text{new}}, \qquad \theta_{\text{old}} \gg \theta_{\text{new}}.$$

It does not follow that

$$c_{\text{old}} > c_{\text{new}}.$$

Indeed the reverse may occur:

$$c_{\text{old}} < c_{\text{new}}.$$

Comparing complaint frequencies across those periods without estimating the change in threshold treats two different measurement regimes as though they were one.

The same problem occurs cross-culturally, across organizations, and across relationships. Different environments impose different costs on criticism, confession, accusation, dissent, vulnerability, and requests for repair. Their observable complaint rates therefore cannot be compared as though the probability of expression were invariant.

Formally, suppose populations A and B satisfy

$$c_A = g_A e_A$$

and

$$c_B = g_B e_B.$$

Then

$$c_A > c_B$$

does not imply

$$g_A > g_B.$$

Indeed,

$$g_A < g_B$$

is compatible with the observation whenever

$$\frac{e_A}{e_B} > \frac{g_B}{g_A}.$$

The inference fails because the observation process differs between the two populations.

Reference adaptation makes the comparison harder still. If

$$g_A = \Gamma(N_A - H_A)$$

and

$$g_B = \Gamma(N_B - H_B),$$

then even recovering

$$g_A > g_B$$

would not establish

$$H_A < H_B.$$

Different reference points can produce different grievance rates under similar absolute conditions, or similar grievance rates under very different absolute conditions.

The complete mapping is therefore

$$(H, N, \theta, a, \kappa, \ldots) \longrightarrow c.$$

It is many-to-one.

This is why complaint frequency is an especially poor standalone measure of social deterioration. It sits at the end of a causal chain containing several variables that can move independently and sometimes in opposite directions.

The argument can be expressed locally through the total differential

$$dc = \frac{\partial c}{\partial H}\, dH + \frac{\partial c}{\partial N}\, dN + \frac{\partial c}{\partial \theta^*}\, d\theta^* + \frac{\partial c}{\partial \kappa}\, d\kappa + \frac{\partial c}{\partial a}\, da + \cdots .$$

A single observed quantity,

$$dc,$$

cannot determine the signs of all the terms that produced it.

That fact becomes politically and morally important because visible complaint is often treated as retrospective evidence about the conditions from which it emerged. A period rich in public criticism can appear more troubled than a period from which comparatively little criticism survives. Yet a low-complaint environment may have achieved silence by keeping

$$\theta^*$$

high rather than by keeping

$$g$$

low.

Conversely, a high-complaint environment may be revealing genuine deterioration. Nothing in the model prevents

$$\dot{H} < 0, \qquad \dot{g} > 0, \qquad \dot{c} > 0.$$

The framework does not turn complaint into evidence of progress. It removes the inference in both directions.

One must know something about the generative process.

This is also where the threshold theory differs from a simple claim that people become more demanding as life improves. That claim concerns primarily the movement of the evaluative reference point:

$$N \uparrow .$$

The present essay adds a second mechanism:

$$\theta_{\text{expression}} \downarrow .$$

A population can become more sensitive to smaller discrepancies while also becoming more capable of talking about them.

Those changes can reinforce one another.

As severe dangers recede, smaller discrepancies may acquire greater relative salience. As social retaliation becomes less severe, those discrepancies may become increasingly inexpensive to express. The resulting public discourse can become denser, finer-grained, and more critical even while the absolute conditions generating it improve.

The temptation is then to compare the content of present complaints with the content of earlier suffering and conclude that something has gone wrong with the complainants. But such a comparison confuses the scale on which an event is evaluated with the mechanism by which it becomes visible.

A complaint can be about something historically minor and still be generated by a perfectly ordinary local discrepancy process.

Likewise, the ability to voice it can be evidence that the cost of expression has fallen.

Neither fact determines whether the complaint is justified.

The model therefore supports no inference of the form

$$\text{minor complaint} \Rightarrow \text{moral triviality},$$

just as it supports no inference of the form

$$\text{frequent complaint} \Rightarrow \text{severe conditions}.$$

Magnitude, salience, expression, and justification are different variables.

This distinction becomes particularly important at the far end of the process imagined by Pearce. If suffering were dramatically reduced while information-sensitive discrimination remained intact, the evaluative system could continue distinguishing

$$H_1 < H_2$$

even when both values lay far above anything presently experienced as good. If such distinctions retained behavioral significance, then some analogue of grievance could persist without anything resembling contemporary suffering.

Whether it would still deserve the word *grievance* is partly a semantic question. The structural point does not depend on the label. A system capable of preferring one state over another must preserve some representation of the difference

$$H_2 - H_1.$$

If that difference can motivate correction, then negative comparison has survived even if negative valence has not.

At the same time, a sufficiently permissive expressive environment could make those tiny comparisons highly visible.

The limiting case would therefore be strange only from the standpoint of a culture accustomed to treating complaint as evidence of suffering. One could imagine agents occupying overwhelmingly positive states while continuously detecting, discussing, and correcting extremely small deviations from still better states.

Their discourse might sound extraordinarily exacting.

Their suffering could nevertheless be approximately zero.

The point is not that this outcome must occur. It is that nothing in the reduction of suffering alone rules it out. Information-sensitive evaluation and permissive expression are separate dimensions from hedonic depth.

The social statistic we observe sits downstream of all three.

This gives the paired theory its strongest general form:

better conditions can coexist with fewer grievances and more complaints.

They can also coexist with more grievances and more complaints, fewer grievances and fewer complaints, or more grievances and fewer complaints. The signs depend on the movement of the reference frame and the expression gate.

Complaint therefore cannot serve as a one-dimensional thermometer for the quality of a social world.

It is an output of a system.

And once expression is understood as an output of several jointly hidden variables, a further temptation becomes difficult to sustain: the belief that what finally crosses the gate must reveal what the person "really" was underneath it. The same inverse problem that prevents complaint from uniquely revealing social conditions also prevents disinhibited behavior from uniquely revealing a supposedly authentic self.

14 Against the Authenticity Fallacy

A seductive misreading of everything developed so far holds that disinhibited behavior reveals what someone "really" thinks or feels—that tact, restraint, and self-regulation are a mask, and that lowering θ far enough eventually exposes the face underneath it. The threshold model gives this reading no support, and the argument against it is stronger than it first appears.

The weak version of the objection is that disinhibited behavior is *state-contaminated*: whatever emerges under intoxication, exhaustion, or elation reflects the transient state as much as any stable disposition, and therefore should not be treated as a clean window onto the person. This is true, but it understates the problem. It suggests that a better measurement—a cleverer intervention, a purer form of disinhibition—might eventually get past the contamination and reach D_i directly.

Within the present model, it cannot.

The problem is not measurement quality. It is identifiability.

Recall the latent quantity governing expression:

$$z_i(t, j) = D_i + \eta_i(S_t) + \varepsilon_i(t) - \theta_i(t, j),$$

with

$$\theta_i(t, j) = \kappa_t \, \theta_i^* \left(S_t, C_t, j, \mathcal{H}_{ij}\right).$$

Here D_i denotes the comparatively stable dispositional component, $\eta_i(S_t)$ the state-dependent perturbation to activation, $\varepsilon_i(t)$ transient variation not otherwise represented, and $\theta_i(t, j)$ the effective expression threshold in the present context and relationship.

Observed expression depends on their combination:

$$\Pr(E_i = 1) = \sigma[z_i(t, j)].$$

In the deterministic limiting form,

$$E_i = 1 \iff z_i(t, j) > 0.$$

The observer therefore encounters D_i only through a quantity in which it is already combined with state, noise, context, enforcement, and threshold. Behavior does not present these terms separately.

Even if $\varepsilon_i(t)$ is temporarily ignored, the difficulty remains. Consider

$$z_i = D_i + \eta_i - \theta_i.$$

For any δ, the transformation

$$(D_i, \eta_i, \theta_i) \longmapsto (D_i + \delta, \eta_i - \delta, \theta_i)$$

leaves the expression margin unchanged:

$$(D_i + \delta) + (\eta_i - \delta) - \theta_i = D_i + \eta_i - \theta_i = z_i.$$

Consequently,

$$\sigma\left[(D_i + \delta) + (\eta_i - \delta) - \theta_i\right] = \sigma(z_i).$$

The two latent decompositions are observationally equivalent with respect to expression.

The same ambiguity can be generated between activation and threshold. For any δ,

$$(D_i, \eta_i, \theta_i) \longmapsto (D_i + \delta, \eta_i, \theta_i + \delta)$$

also leaves z_i invariant:

$$(D_i + \delta) + \eta_i - (\theta_i + \delta) = z_i.$$

More generally, transformations satisfying

$$\Delta D_i + \Delta\eta_i - \Delta\theta_i = 0$$

belong to an equivalence class of latent configurations producing the same expression margin. Define

$$(D_i, \eta_i, \theta_i) \sim (D_i', \eta_i', \theta_i')$$

whenever

$$D_i + \eta_i - \theta_i = D_i' + \eta_i' - \theta_i'.$$

Expression identifies, at most, the equivalence class

$$[(D_i, \eta_i, \theta_i)]_\sim,$$

not its unique decomposition.

This is the structural point. A larger stable disposition compensated by a smaller state contribution can be behaviorally indistinguishable from a smaller stable disposition accompanied by a larger state contribution. Likewise, greater activation accompanied by a greater threshold can be indistinguishable from weaker activation accompanied by weaker inhibition.

The observable remains

$$z_i.$$

Its decomposition does not.

This is a different and harder claim than saying that "the drunk utterance is unreliable evidence." It says that the drunk utterance does not identify D_i. Nor can additional observations of expression, considered only through this generative model, resolve the ambiguity. The problem is not that the observer has collected too little behavioral evidence. It is that the same behavioral evidence is compatible with infinitely many decompositions of the latent quantity that generated it.

Recovering D_i therefore requires additional identifying assumptions or measurements not contained in expression itself.

Repeated observations can certainly constrain the possibilities. If a behavior appears reliably across many states, recipients, and contexts, an observer may have increasingly good grounds for estimating some stable component. One might write

$$z_i(t, j) = D_i + \eta_i(S_t) - \kappa_t \theta_i^*(S_t, C_t, j, \mathcal{H}_{ij}) + \varepsilon_i(t)$$

and attempt to estimate D_i from a sufficiently rich longitudinal design. But doing so requires assumptions about the distributions or functional forms of the other terms. The behavioral record does not supply the decomposition by itself.

This matters for the familiar claim that some intervention "reveals the real person." Suppose an intervention Q lowers the effective threshold:

$$Q \longrightarrow \theta_i \downarrow .$$

Then a previously unexpressed behavior may satisfy

$$z_i^{\text{before}} < 0$$

but

$$z_i^{\text{after}} > 0.$$

The behavior becomes visible.

What has been learned?

At minimum, one has learned that the post-intervention system occupied a state for which

$$D_i + \eta_i(S_t) + \varepsilon_i(t) > \theta_i(t, j).$$

That is real information. It establishes something about the configuration that actually generated the behavior.

What it does not establish is

$$D_i = \text{the expressed behavior.}$$

Nor does it establish that the behavior was previously present as a complete, fully formed intention imprisoned behind inhibition. State can alter activation as well as permission. Context can alter both. Relationships can change the behavioral repertoire itself.

The limit

$$\theta_i \to 0$$

does not solve the problem.

If

$$\theta_i \to 0,$$

then

$$z_i(t) \to D_i + \eta_i(S_t) + \varepsilon_i(t),$$

not

$$D_i.$$

Even perfect removal of the gate leaves the state-dependent activation term intact.

Thus

$$\lim_{\theta_i \to 0} z_i \neq D_i$$

in general.

The imagined experiment in which all inhibition is stripped away therefore does not expose a pure dispositional core. It exposes a person in the highly specific state produced by stripping away inhibition.

That state is itself part of the causal system.

There is a deeper conceptual problem with the authenticity interpretation as well. It treats inhibition as though it were external to the person whose behavior is being explained. Disposition belongs to the self; restraint is treated as an obstruction placed around it.

But the present model contains no such asymmetry.

Both activation and inhibition belong to the generative architecture:

$$E_i = F\left(D_i, \eta_i, \varepsilon_i, \theta_i^*, \kappa, C_t, j, \mathcal{H}_{ij}\right).$$

A person's anticipation of consequences, learned restraint, capacity for self-control, concern for another person's feelings, uncertainty, sense of social obligation, and accumulated relationship history are not foreign objects placed between an inner person and the world. They are among the processes by which that person produces behavior.

Suppose someone experiences an impulse to make a cruel remark but refrains because they anticipate its effect on another person. The authenticity fallacy assigns ontological privilege to the impulse:

$$\text{cruel impulse} = \text{real self},$$

while treating the inhibition as concealment:

$$\text{restraint} = \text{mask}.$$

Nothing in the causal structure warrants this ranking.

The impulse contributed to the outcome.

So did the restraint.

Had the restraint remained effective, silence would have been the person's behavior just as genuinely as the remark becomes their behavior when the threshold is crossed.

The inhibition was equally them.

This does not mean that disinhibited behavior teaches us nothing about latent dispositions. It can reveal possibilities that ordinary conditions rarely make observable. A person who repeatedly becomes affectionate under lowered social inhibition has supplied evidence that affectionate activation exists somewhere in the relevant behavioral system. A person who repeatedly becomes hostile has supplied evidence of something different.

But evidence of a latent possibility is not identification of an authentic essence.

The distinction is especially important because inhibition itself can be stable. If D_i is granted trait-like status because it persists across time, then persistent threshold structure deserves comparable attention. Suppose

$$\theta_i^*(t, j) \approx \bar{\theta}_i^*(j)$$

over long intervals. A person's characteristic unwillingness to humiliate others, characteristic reserve around strangers, characteristic candor among friends, or characteristic caution around irreversible consequences may be as stable as the dispositions these thresholds regulate.

A theory that calls the first set of regularities "the person" and the second set "the mask" has inserted its conclusion into its vocabulary.

The threshold model does not need that distinction.

It needs only the generative relation

$$z_i = a_i - \theta_i.$$

Both sides are causal.

Both sides vary.

Both sides can carry history.

And neither acquires metaphysical priority merely because one becomes easier to notice after the other has weakened.

The authenticity fallacy can therefore be stated precisely. It mistakes

$$E_i = 1$$

under a lowered threshold for an observation of

$$D_i.$$

But what has actually been observed is

$$E_i \sim \sigma \left[D_i + \eta_i(S_t) + \varepsilon_i(t) - \kappa_t \theta_i^*(S_t, C_t, j, \mathcal{H}_{ij}) \right].$$

The inverse mapping is not unique.

There is no privileged threshold setting at which behavior suddenly ceases to be generated by a system and becomes a transparent report from an underlying self.

There are only different states of the system.

15 Conclusion: Permission Is Causal

The essay began with an observation that looked, on the surface, like a contradiction. Elevated positive affect can make someone simultaneously warmer and more critical, more affectionate and less tactful, more generous and more willing to disagree—sometimes in the same conversation, sometimes in the same sentence.

A purely directional account of affect has difficulty with this pattern. If positive affect pushes behavior toward some positive class of outcomes, the simultaneous release of criticism, dissent, confession, eccentricity, or tactlessness appears to require a second explanation moving in another direction.

The account developed here does not require affect to push in two directions at once.

It requires only that affect sometimes act on a variable upstream of direction itself.

Across the preceding sections, one generative structure has done most of the work. Activation was represented as

$$a_i(t) = D_i + \eta_i(S_t) + \varepsilon_i(t),$$

where D_i is a comparatively stable dispositional component, $\eta_i(S_t)$ is a state-dependent contribution, and $\varepsilon_i(t)$ collects transient variation not otherwise represented.

The effective expression threshold was represented as

$$\theta_i(t, j) = \kappa_t \, \theta_i^* \left(S_t, C_t, j, \mathcal{H}_{ij} \right),$$

where θ_i^* represents the threshold implied by the person's evaluation of the behavior and its anticipated consequences, while

$$0 < \kappa_t \leq 1$$

represents the capacity to enforce that threshold.

The expression margin is

$$z_i(t, j) = a_i(t) - \theta_i(t, j),$$

and observable expression is generated probabilistically by

$$\Pr(E_i = 1) = \sigma[z_i(t, j)] \,.$$

Substitution gives the unified form

$$\boxed{\Pr(E_i = 1) = \sigma[D_i + \eta_i(S_t) + \varepsilon_i(t) - \kappa_t \, \theta_i^* \, (S_t, C_t, j, \mathcal{H}_{ij})] \,.}$$

The fifteen sections of this essay can be read as investigations of the terms inside this single expression.

The dispositional component D_i describes part of what is available to be expressed, but Section XIV showed why it cannot simply be read backward from behavior. Expression identifies a composite latent margin, not a uniquely recoverable trait.

The state-dependent term

$$\eta_i(S_t)$$

contains the ordinary motivational pathway. Affect can genuinely change what someone wants, attends to, approaches, avoids, or feels inclined to do. Nothing in the threshold theory requires denying that familiar mechanism.

Its claim is instead that motivation is not the only moving part.

The normative threshold

$$\theta_i^*$$

represents the behavioral cost structure as the person presently encounters it: anticipated social consequences, uncertainty, norms, power relations, reversibility, recipient identity, and accumulated history. A behavior can become easier to express because this cost structure changes even if its activation does not.

The enforcement factor

$$\kappa_t$$

adds a further distinction. A person can continue to judge a behavior costly while becoming temporarily less capable of enforcing that judgment. This is why exhaustion, intoxication, anonymity, perceived safety, and elevated affect need not be treated as interchangeable examples of one vague phenomenon called disinhibition.

They can arrive at similar observable behavior through different causal routes.

One intervention may produce

$$\theta_i^* \downarrow$$

while leaving

$$\kappa_t \approx 1.$$

Another may leave

$$\theta_i^*$$

nearly unchanged while producing

$$\kappa_t \downarrow .$$

A third may alter

$$\eta_i(S_t),$$

and a fourth may alter several terms simultaneously.

The observable fact

$$E_i = 1$$

does not tell us which route was taken.

This was the reason for decomposing the threshold rather than allowing θ to become a catch-all explanation. Permission is causal only if the mechanisms governing permission can themselves be distinguished.

The same requirement produced the distinction between global and content-specific threshold movement. A broadly acting intervention may shift

$$\theta_{\text{common}},$$

making many kinds of behavior simultaneously easier to express, while an explicit permission concerning one category of behavior may alter only a content-specific residual

$$\delta_i.$$

Thus

$$\theta_i = \theta_{\text{common}} + \delta_i$$

allows a common gate to move without requiring all behaviors to become identically permissible.

This matters for the essay's original observation. Warmth and criticism need not share a motivational cause. They need only occupy the same newly admitted region of the behavioral field.

If

$$z_{\text{affection}} < 0$$

and

$$z_{\text{criticism}} < 0$$

under one state, but a common permissive shift produces

$$z_{\text{affection}} > 0$$

and

$$z_{\text{criticism}} > 0,$$

both behaviors can become visible together.

Nothing requires

$$D_{\text{affection}}$$

and

$$D_{\text{criticism}}$$

to point in the same social direction.

The gate is not the content.

That observation generated the essay's variance prediction. If an intervention primarily relaxes a broadly shared gate, it should expose heterogeneous latent dispositions rather than move everyone uniformly toward one behavioral target. The newly visible population should therefore become, under appropriate conditions, more behaviorally diverse.

The direction of that newly visible behavior should also correlate with independently measured prior dispositions or suppression histories rather than with some universal valence supplied by the intervention itself.

These predictions distinguish a permission account from a theory that merely renames motivation.

The social argument extended the same structure across relationships. The threshold is not simply

$$\theta_i(t).$$

It is

$$\theta_i(t, j),$$

because behavior is expressed toward someone, or before some audience, whose history and capacity to impose consequences matter.

Trust therefore acquired a restricted but useful formal meaning. Repeated interactions can change the region of the behavioral field that can safely become visible:

$$\mathcal{H}_{ij} \longrightarrow \theta_i^*(t, j).$$

This made it possible to understand why secure relationships can become both warmer and more argumentative. Affection and disagreement cease to be opposites when neither threatens continuation.

The same result complicated the interpretation of social conflict. If g denotes latent disagreement or grievance and e its probability of expression, then observable complaint satisfies

$$c = ge.$$

Consequently,

$$\dot{c} = e\dot{g} + g\dot{e}.$$

This small equation connected the threshold theory to *The North Setpoint*. The companion essay concerns processes capable of changing g: reference adaptation, discrepancy, and the changing scale on which conditions become grievance-worthy. The present essay concerns e: the permeability of the boundary between an available grievance and an expressed complaint.

The distinction makes several otherwise puzzling combinations possible:

$$\dot{H} > 0, \qquad \dot{g} < 0, \qquad \dot{c} > 0,$$

or

$$\dot{H} > 0, \qquad \dot{g} > 0, \qquad \dot{c} > 0,$$

depending on how reference adaptation and expression thresholds move.

Complaint frequency is therefore not a monotonic measure of suffering.

Nor is silence a monotonic measure of harmony.

Nor is disinhibited behavior a transparent measure of personality.

These are instances of the same more general problem: the observable sits at the end of a generative process containing latent variables whose movements cannot be reconstructed from the endpoint alone.

This is where the essay's engagement with Adler becomes useful. Behavior does not have to be imagined as a collection of isolated impulses waiting behind independent gates. Dispositions, thresholds, and interpretations can themselves be organized within a person's characteristic way of approaching the social world. The relationship between person and environment is therefore not merely one of censorship. Social history can change what is permitted, what is expected, and sometimes what is generated in the first place.

The threshold model remains deliberately narrower than a complete psychology. It does not claim that all behavior is latent behavior waiting for permission. It does not claim that affect works primarily through inhibition. It does not claim that every increase in visible disagreement indicates safety, or that every reduction in visible disagreement indicates suppression.

Its claim is methodological and causal.

Observed behavior depends on more than activation.

A disposition can be present without being expressed.

Expression can change without disposition changing.

Different mechanisms can produce the same apparent disinhibition.

The same permissive change can release behaviors of opposite valence.

And the behavior that emerges after inhibition weakens does not thereby acquire special status as the authentic person beneath restraint.

The inhibition was part of the causal system too.

The distinction can be summarized by separating three questions.

The first asks what makes a behavior available:

What is activated?

The second asks whether that available behavior can cross into expression:

What is permitted?

The third asks what an observer can infer after expression has occurred:

What does the observable identify?

Conflating these questions produces many of the errors examined throughout the essay. Motivation is mistaken for permission. Permission is mistaken for character. Silence is mistaken for absence. Expression is mistaken for authenticity. Complaint is mistaken for suffering. Conflict is mistaken for hostility.

Separating them does not make behavior less meaningful.

It makes the causal structure richer.

The original puzzle can therefore be stated one final time without paradox. Someone can become kinder and more critical in the same conversation because kindness and criticism need not have been produced by the affective state that made them visible. They may already have occupied neighboring regions of a heterogeneous behavioral field. Affect altered activation, appraisal, enforcement, permission, or some combination of them, and more of that field crossed into expression.

The question "What made this person want to say that?" remains important.

So does the question "What does saying it tell us about them?"

But between them lies another question. It concerns the causal boundary across which latent behavior becomes observable, the history that shaped that boundary, and the conditions under which it moves.

That is the question this essay has tried to isolate:

What had to become permissible before this disposition could become visible?

References

[1] Schwarz, N., and Clore, G. L. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45(3):513–523, 1983. doi:10.1037/0022-3514.45.3.513.

[2] Schwarz, N., and Clore, G. L. Mood as information: 20 years later. *Psychological Inquiry*, 14(3–4):296–303, 2003.

[3] Isen, A. M., and Shalker, T. E. The effect of feeling state on evaluation of positive, neutral, and negative stimuli: When you "accentuate the positive," do you "eliminate the negative"? *Social Psychology Quarterly*, 45(1):58–63, 1982.

[4] Isen, A. M., Daubman, K. A., and Nowicki, G. P. Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology*, 52(6):1122–1131, 1987.

[5] Fredrickson, B. L. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist*, 56(3):218–226, 2001.

[6] Clore, G. L., Gasper, K., and Garvin, E. Affect as information. In Forgas, J. P. (ed.), *Handbook of Affect and Social Cognition*, pp. 121–144. Lawrence Erlbaum Associates, Mahwah, NJ, 2001.

[7] Gray, J. A. Précis of *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System*. *Behavioral and Brain Sciences*, 5(3):469–484, 1982. doi:10.1017/S0140525X00013066.

[8] Gray, J. A. *The Psychology of Fear and Stress*. Second edition. Cambridge University Press, Cambridge, 1987.

[9] Gray, J. A., and McNaughton, N. *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System*. Second edition. Oxford University Press, Oxford, 2000.

[10] Carver, C. S., and White, T. L. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, 67(2):319–333, 1994.

[11] Logan, G. D. On the ability to inhibit simple thoughts and actions: A users' guide to the stop signal paradigm. In Dagenbach, D., and Carr, T. H. (eds.), *Inhibitory Processes in Attention, Memory, and Language*. Academic Press, San Diego, 1994.

[12] Aron, A. R. The neural basis of inhibition in cognitive control. *The Neuroscientist*, 13(3):214–228, 2007.

[13] Wiecki, T. V., and Frank, M. J. A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, 120(2):329–355, 2013.

[14] Baumeister, R. F., Heatherton, T. F., and Tice, D. M. *Losing Control: How and Why People Fail at Self-Regulation*. Academic Press, San Diego, 1994.

[15] Muraven, M., Tice, D. M., and Baumeister, R. F. Self-control as limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology*, 74(3):774–789, 1998.

[16] Baumeister, R. F., Vohs, K. D., and Tice, D. M. The strength model of self-control. *Current Directions in Psychological Science*, 16(6):351–355, 2007. doi:10.1111/j.1467-8721.2007.00534.x.

[17] Baumeister, R. F., Tice, D. M., and Vohs, K. D. The strength model of self-regulation: Conclusions from the second decade of willpower research. *Perspectives on Psychological Science*, 13(2):141–145, 2018.

[18] Inzlicht, M., and Schmeichel, B. J. What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science*, 7(5):450–463, 2012.

[19] Shenhav, A., Botvinick, M. M., and Cohen, J. D. The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013.

[20] Steele, C. M., and Josephs, R. A. Alcohol myopia: Its prized and dangerous effects. *American Psychologist*, 45(8):921–933, 1990. doi:10.1037/0003-066X.45.8.921.

[21] Josephs, R. A., and Steele, C. M. The two faces of alcohol myopia: Attentional mediation of psychological stress. *Journal of Abnormal Psychology*, 99(2):115–126, 1990. doi:10.1037/0021-843X.99.2.115.

[22] Giancola, P. R. Executive functioning: A conceptual framework for alcohol-related aggression. *Experimental and Clinical Psychopharmacology*, 8(4):576–597, 2000.

[23] Fillmore, M. T. Drug abuse as a problem of impaired control: Current approaches and findings. *Behavioral and Cognitive Neuroscience Reviews*, 2(3):179–197, 2003.

[24] Suler, J. The online disinhibition effect. *CyberPsychology & Behavior*, 7(3):321–326, 2004. doi:10.1089/1094931041291295.

[25] Joinson, A. N. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2):177–192, 2001.

[26] Zimbardo, P. G. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In Arnold, W. J., and Levine, D. (eds.), *Nebraska Symposium on Motivation*, Vol. 17, pp. 237–307. University of Nebraska Press, Lincoln, 1969.

[27] Diener, E. Deindividuation: The absence of self-awareness and self-regulation in group members. In Paulus, P. B. (ed.), *Psychology of Group Influence*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

[28] Spears, R., and Lea, M. Panacea or panopticon? The hidden power in computer-mediated communication. *Communication Research*, 21(4):427–459, 1994.

[29] Goffman, E. *The Presentation of Self in Everyday Life*. Doubleday, Garden City, NY, 1959.

[30] Goffman, E. *Interaction Ritual: Essays on Face-to-Face Behavior*. Pantheon Books, New York, 1967.

[31] Noelle-Neumann, E. The spiral of silence: A theory of public opinion. *Journal of Communication*, 24(2):43–51, 1974.

[32] Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs*, 70(9):1–70, 1956.

[33] Cialdini, R. B., and Goldstein, N. J. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621, 2004.

[34] Foucault, M. *Discipline and Punish: The Birth of the Prison*. Translated by Alan Sheridan. Pantheon Books, New York, 1977.

[35] Rotter, J. B. A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35(4):651–665, 1967.

[36] Rempel, J. K., Holmes, J. G., and Zanna, M. P. Trust in close relationships. *Journal of Personality and Social Psychology*, 49(1):95–112, 1985.

[37] Rusbult, C. E. Commitment and satisfaction in romantic associations: A test of the investment model. *Journal of Experimental Social Psychology*, 16(2):172–186, 1980.

[38] Clark, M. S., and Mills, J. Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology*, 37(1):12–24, 1979.

[39] Reis, H. T., Clark, M. S., and Holmes, J. G. Perceived partner responsiveness as an organizing construct in the study of intimacy and closeness. In Mashek, D. J., and Aron, A. P. (eds.), *Handbook of Closeness and Intimacy*, pp. 201–225. Lawrence Erlbaum Associates, Mahwah, NJ, 2004.

[40] Mischel, W. *Personality and Assessment*. Wiley, New York, 1968.

[41] Mischel, W., and Shoda, Y. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2):246–268, 1995.

[42] Fleeson, W. Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, 80(6):1011–1027, 2001.

[43] Fleeson, W. Moving personality beyond the person-situation debate: The challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, 13(2):83–87, 2004.

[44] Wood, A. M., Linley, P. A., Maltby, J., Baliousis, M., and Joseph, S. The authentic personality: A theoretical and empirical conceptualization and the development of the Authenticity Scale. *Journal of Counseling Psychology*, 55(3):385–399, 2008.

[45] Kernis, M. H., and Goldman, B. M. A multicomponent conceptualization of authenticity: Theory and research. In Zanna, M. P. (ed.), *Advances in Experimental Social Psychology*, Vol. 38, pp. 283–357. Elsevier Academic Press, San Diego, 2006.

[46] Adler, A. *Understanding Human Nature*. Greenberg, New York, 1927.

[47] Adler, A. *The Science of Living*. Greenberg, New York, 1929.

[48] Adler, A. *The Education of Children*. Greenberg, New York, 1930.

[49] Adler, A. *What Life Should Mean to You*. Little, Brown, and Company, Boston, 1931.

[50] Adler, A. *Social Interest: A Challenge to Mankind*. Capricorn Books, New York, 1964. Originally published 1938.

[51] Ansbacher, H. L., and Ansbacher, R. R. (eds.). *The Individual Psychology of Alfred Adler: A Systematic Presentation in Selections from His Writings*. Basic Books, New York, 1956.

[52] Gross, J. J. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299, 1998.

[53] Gross, J. J. Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39(3):281–291, 2002.

[54] Gross, J. J., and John, O. P. Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85(2):348–362, 2003.

[55] Gross, J. J. Emotion regulation: Current status and future prospects. *Psychological Inquiry*, 26(1):1–26, 2015.

[56] Keltner, D., and Haidt, J. Social functions of emotions at four levels of analysis. *Cognition & Emotion*, 13(5):505–521, 1999.

[57] Van Kleef, G. A. The emerging view of emotion as social information. *Social and Personality Psychology Compass*, 4(5):331–343, 2010.

[58] Van Kleef, G. A. How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science*, 18(3):184–188, 2009.

[59] Elliot, A. J. Approach and avoidance motivation and achievement goals. *Educational Psychologist*, 34(3):169–189, 1999.

[60] Elliot, A. J., and Thrash, T. M. Approach-avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology*, 82(5):804–818, 2002.

[61] Carver, C. S., and Scheier, M. F. *On the Self-Regulation of Behavior*. Cambridge University Press, Cambridge, 1998.

[62] Green, D. M., and Swets, J. A. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

[63] MacCallum, R. C. Working with imperfect models. *Multivariate Behavioral Research*, 38(1):113–139, 2003.

[64] Bollen, K. A. *Structural Equations with Latent Variables*. Wiley, New York, 1989.

[65] Pearl, J. *Causality: Models, Reasoning, and Inference*. Second edition. Cambridge University Press, Cambridge, 2009.

[66] Pearl, J., and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 2018.

[67] Brickman, P., and Campbell, D. T. Hedonic relativism and planning the good society. In Appley, M. H. (ed.), *Adaptation-Level Theory: A Symposium*, pp. 287–302. Academic Press, New York, 1971.

[68] Brickman, P., Coates, D., and Janoff-Bulman, R. Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology*, 36(8):917–927, 1978.

[69] Helson, H. *Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior*. Harper & Row, New York, 1964.

[70] Kahneman, D., and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[71] Frederick, S., and Loewenstein, G. Hedonic adaptation. In Kahneman, D., Diener, E., and Schwarz, N. (eds.), *Well-Being: The Foundations of Hedonic Psychology*, pp. 302–329. Russell Sage Foundation, New York, 1999.

[72] Diener, E., Lucas, R. E., and Scollon, C. N. Beyond the hedonic treadmill: Revising the adaptation theory of well-being. *American Psychologist*, 61(4):305–314, 2006.

[73] Pearce, D. *The Hedonistic Imperative*. Online monograph, 1995–. Available from the author's writings at Hedweb.

[74] Pearce, D. Information-sensitive gradients of bliss. *The Hedonistic Imperative / Hedweb*, 2016.

[75] Pearce, D. Life in the Far North: An information-theoretic perspective on heaven. Essay on gradients of information-sensitive well-being. Hedweb.

[76] James, W. *The Principles of Psychology*. Henry Holt and Company, New York, 1890.

[77] Lewin, K. *Principles of Topological Psychology*. McGraw-Hill, New York, 1936.

[78] Heider, F. *The Psychology of Interpersonal Relations*. Wiley, New York, 1958.

[79] Festinger, L. A theory of social comparison processes. *Human Relations*, 7(2):117–140, 1954.

[80] Bandura, A. *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[81] Bandura, A. Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50(2):248–287, 1991.

[82] Gibson, J. J. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

[83] Clark, A. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, Cambridge, MA, 1997.